Earliest Predictor of Dropout in MOOCs: A Longitudinal Study of FutureLearn Courses

Alexandra Ioana Cristea *alexandra.i.cristea@durham.ac.uk*
**Computer Science, Durham University, Durham, UK*

Ahmed Alamri* *ahmed.alamri@durham.ac.uk*

Mizue Kayama *kayama@cs.shinshu-u.ac.jp*
Shinshu University, Shinshu, Japan

Craig Stewart *craig.stewart@coventry.ac.uk*
Coventry University, Coventry, UK

Mohammad Alshehri* *mohammad.alshehri@durham.ac.uk*

Lei Shi *lei.shi@liverpool.ac.uk*
University of Liverpool, Liverpool, UK

Abstract

Whilst a high dropout rate is a well-known problem in MOOCs, few studies take a data-driven approach to understand the reasons of such a phenomenon, and to thus be in the position to recommend and design possible adaptive solutions to alleviate it. In this study, we are particularly interested in finding a *novel early detection mechanism* of potential dropout, and thus be able *to intervene at an as early time as possible*. Additionally, unlike previous studies, we explore a light-weight approach, based on as little data as possible – since different MOOCs store different data on their users – and thus strive to create a truly generalisable method. Therefore, we focus here specifically on the generally available *registration date* and its relation to the course start date, via a *comprehensive, larger than average, longitudinal study of several runs of all MOOC courses at the University of Warwick between 2014-1017*, on the less explored European FutureLearn platform. We identify specific periods where different interventions are necessary, and propose, based on statistically significant results, specific *pseudo-rules for adaptive feedback.*
Keywords: Learning Analytics, FutureLearn, longitudinal study, MOOC.

1. Introduction

Massively Open Online Courseware (MOOC) has become a key mainstream approach [2] to democratise knowledge. Building on efforts from the Open Education Resources movement [26] originating from the 1990's, and significantly boosted in 2001 by MIT's Open Courseware (ocw.mit.edu) announcement, on making available quality teaching materials to all academics, MOOCs have since proven to be a popular, if not entirely effective [8], choice of education.

Whilst MOOC courses can scale their delivery to many tens of thousands of students (or more [25]), only a small percentage of those students actually complete the course. Completion rage figures vary typically between 3-15% [6, 13] for non-fee charging courses and rises over 70% for those courses that charge [15]. Such a situation undermines the goal of making educational resources available to enable mass-access and learning. Thus, there has been a great deal of interest and research into why these students drop out and how to keep them engaged with the course till completion [3,12,15]. However, most solutions involve a large number of parameters and are what we call 'heavy-weight'. Whilst such approaches may provide higher

accuracy, they are less applicable in real-life, as they require real-time processing of large quantities of data, as well as may provide results too late in the course-cycle to be of real effect. Instead, thus, we focus on an Occam's razor[1] approach, attempting to use here the very first data available, at the earliest stage in the process, in a '*light-weight*' *approach*: the *registration date*.

FutureLearn[2] is a free European online learning initiative, similar to Coursera[3] in the United States. FutureLearn started in 2012 as a partnership between several UK universities, the BBC and the British library, expanding later to include courses from international institutions, NGOs and businesses. Considering the breadth of the courses on offer, very little has so far been done to analyse the success of FutureLearn, specifically. Indeed, a quick literature search on Google Scholar[4] (since 2014) on 'FutureLearn analytics' renders only 865 results, as opposed to 3870 on 'Coursera analytics'. Thus, a further contribution of this paper is to perform a study on a platform that is popular and growing, however, being less explored. Yet, the results presented here are of a *generic nature*, as the registration date is available to any MOOC currently used.

Moreover, this paper presents the results of the analysis of a larger than average data set of FutureLearn MOOC users over several runs, focusing specifically on non-completion, to determine *if there are factors that can be identified before the students even start the course*, that can *guide teachers to target and support these students*, so that they do not disengage from their learning. To analyse students' non-completion, various variables can be considered, such as student profile data (e.g., age, gender, country), behavioural patterns related to the consumption and generation of data when interacting with the course (e.g., reading, watching, writing, taking quizzes). This paper instead, however, uses only one relatively simple variable, which, to the authors' best knowledge, has not been studied in prior research in relation to completion predictability: the *registration date*. The advantage is that this data is available, in most of cases, even before the course starts; thus, if completion can be predicted from it, very early intervention is possible. Hence, this paper targets the following research questions:

RQ1: Can the date of registration (in terms of distance from the course start) of students predict their completion (or non-completion)?

RQ2: How can the dropout-rates be alleviated based on the registration date?

2. Related Research and Setup

2.1. Learning Analytics and Educational Data Mining

Learning Analytics (LA) and Educational Data Mining (EDM) are two fields which are the basis of our research presented in this paper. LA is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and environments in which it occurs" [18]. EDM is a process of applying computerised methods, such as machine learning and data mining, to the enormous volume of educational data [17]. LA is thus arguably mainly aimed at human consumption, whereas EDM is mainly aimed at computer processing. However, the boundaries are not very strict. In terms of applicable techniques for educational data, most are appropriate for both EDM and LA, and encompass statistical methods, data mining, machine learning, network analysis and visualisation. Three techniques often used by both are as follows [17]. *Clustering methods* are used to categorise groups of learners based on similar features. *Prediction techniques* are used to estimate a target variable, based on existing data of other variables. *Relationship mining techniques* are used to identify the relationship between variables, such as learner behaviour

[1] www.britannica.com/topic/Occams-razor

[2] www.futurelearn.com

[3] www.coursera.org

and learner difficulties. In terms of popularity, the research trend is gradually moving towards LA rather than EDM, although both areas are still growing [17] (see **Fig. 1**).

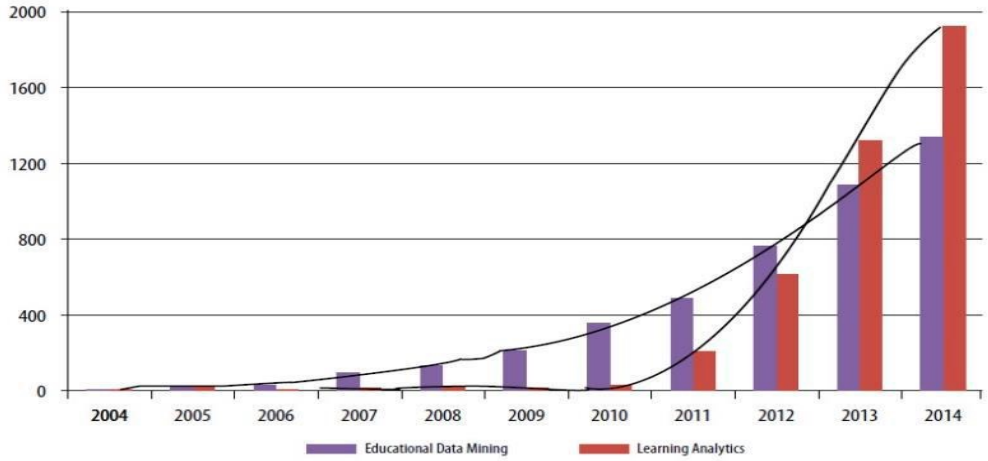

Fig. 1. Evolution of terms on EDM and LA occurring in research papers, according to [17].

In the current paper, the authors apply statistical methods with the goal of informing both the *teacher* (as *course designer*) and potentially the *student*, about the retention opportunities for that student. Specifically, we aim at forming the basis of creating rules for specific feedback, to be returned to the student, in order to ensure retention.

2.2. MOOCs Analytics and Mining

Our work as presented in this paper is closest related to the area of retention-versus-dropout in MOOCs. The issue of massive open online course systems (MOOCs) having high dropout rate has been observed, as said, by many researchers [9, 13, 16]. Various solutions have been proposed, such analysing a students' activity in online forums [27], or analysing the students' click-stream [14, 22], classification methods of longitudinal engagement trajectories [6] and monitoring video views [11]. However, most of these approaches are only able to predict retention or dropout once a student has started learning and, importantly, interacting with a MOOC. For example, in [14], correlation was observed between activities in the latter part of the course and dropping out. Also [27] observed the relationship between learner sentiments expressed on forums and the chance of dropping out.

To the best of our knowledge, **none of these studies attempts to predict dropout only from the very first interaction with the system – the registration**.

Moreover, the FutureLearn platform has not been studied as frequently as other MOOC platforms (e.g., Coursera and edX) [24]. Recent work on FutureLearn data exploration includes social aspects [5], dashboard development [24], pedagogies on MOOCs [19], new courses on FutureLearn [23], and reviews of empirical MOOC literature [22, 28].

Another common point for prior researches is that they have analysed only a few courses in a MOOC (e.g. [11] claims to be the largest study, with only 4 courses, with only one run each; one MOOC with one run in [4]; three courses used in [1]). They have often only analysed courses on the same, or similar, subjects [5].

Thus, unlike their research, we **have performed comparative longitudinal studies of several runs of a large number of different courses on varied subjects**.

2.3. Setup: Terms and Methodology

Firstly, a few definitions are required, as follows. Here we are studying *synchronous MOOC courses*, i.e., courses which have an official starting date (considered in the rest of the paper as date 0) and which are expected to run over a specific number of weeks, after which they end. *Non-completing students* are students who have not completed the course by the data collection

point (October 2017, more than 3 months after the last course ended). *Enrolled students* are students that completed enrolment. Note however that these can be also students who have never logged into the course, but just have enrolled for it. *Completing students* are students who have completed the course before the data collection point. These are students who have done all of the activities below, where applicable (as not all courses allow for all of them): all pages read; all videos watched; all tasks done; all quizzes done.

To address the research questions (section 1), we analyse 6 courses on different subjects delivered by the University of Warwick (from literature to computer science to social sciences: Literature and Mental Health, Shakespeare and His World, Big Data, Supply Chains, Babies in mind, The Mind is Flat), each with several runs, for a total of 23 runs, for a total of 240,568 students, employing a variety of statistical methods. These courses were freely available for anyone during the analysed period 2014-2017 (since FutureLearn started and Warwick joined), and allowed for enrolment at any time. A notification was sent automatically upon enrolment, as well as just before the start of the course. To assess if the data available is normally distributed, we use the Pearson chi-squared test (establishing 'goodness of fit'). Depending on this, we then use a T-test for normally distributed data, or the Wilcoxon signed-ranked test otherwise. The Bonferroni correction is used for compensation of multiple comparisons.

3. Results

Results show that out of 240,568 registered students, only 7,437 (~3%) complete (see Table 1); thus, this highlights an extreme MOOC non-completion issue, at the lower end of the boundary of 3-15% [6, 13]. We further analyse the normality of the registration data, results showing that registration is not normally distributed ($p < 2.2e\text{-}16$). Thus, the T-test cannot be applied. Thus, we have applied the non-parametric Wilcoxon test instead – firstly, to all data across all courses and runs (column 'Total' in table below). We notice that students register, on average, 1 month (30.47 days) in advance of their FutureLearn courses start. We can also see that non-completers tend to register, possibly non-intuitively, around 3.5 days earlier, on average, than completers (see more discussion on this in section 5) and that this difference is statistically significant. We also can estimate that non-completers are the ones influencing the overall average (due to their larger number) and the large variance (with a maximum of 256 days in advance, up to 809 days after the course starts).

Table 1. Initial analysis of impact of registration date (Reg.) versus course starting date (here, 0), onto completion.

		Total	**Reg. > 90**	**90 <= Reg. <= -30**	**-30 > Reg.**
ALL	**Data size**	240568	16522	214676	9370
	Avg.	**30.47**	**142.14**	**25.37**	**-49.39**
	Var.	2142.46	1160.13	1008.42	863.49
	Max.	256	256	90	-31
	Min.	-809	91	-30	-809
Completers	**Data size**	7437	279	7016	142
	Avg.	**27.06**	**126.85**	**24.54**	**-44.69**
	Var.	1459.73	1206.21	990.29	131.03
	Max.	210	210	90	-31
	Min.	-83	91	-30	-83
Non-completers	**Data size**	233131	16243	207660	9228
	Avg.	**30.58**	**142.4**	**25.39**	**-49.46**
	Var.	2163.86	115.33	1009.01	904.9
	Max.	256	256	90	-31
	Min.	-809	91	-30	-809
Who registers earlier?		Non-Completers	Non-Completers	Non-Completers	Completers
Wilcoxon's p		*p = 0.0010*	*p =1.94e^{-13}*	*p = 0.093*	*p = 0.033*

This large spread becomes more obvious in the box diagram (Fig. 2). The figure shows that *non-completers* are responsible for most outliers, as well as the largest spread. It becomes visible from the figure that registering too early - or too late - will possibly result in non-completion; this is studied further, below, in order to quantify what 'too late' or 'too early' means in this context.

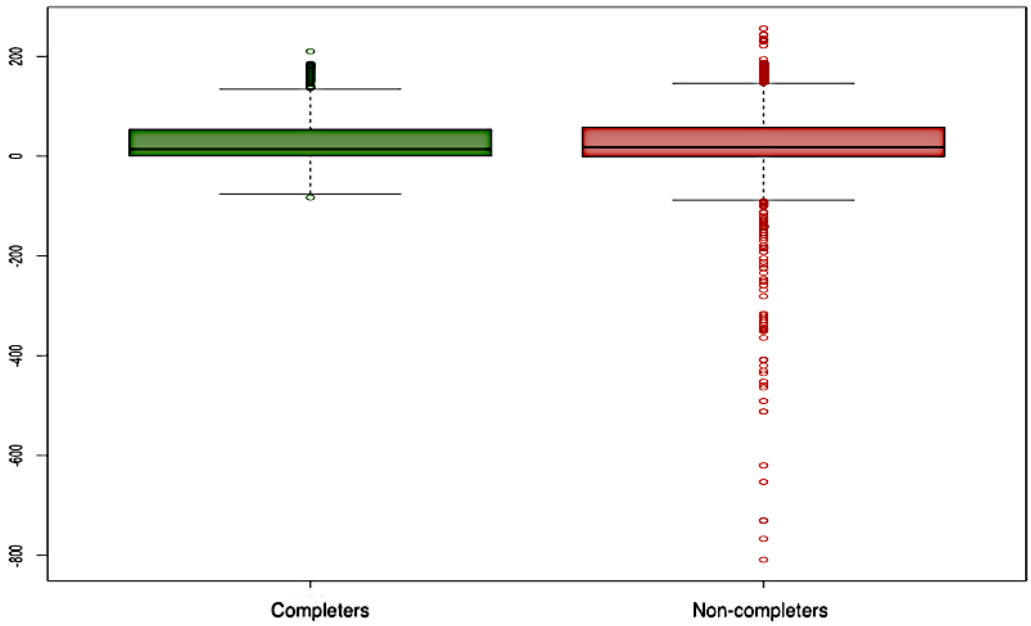

Fig. 2. Box diagram for registration date for *completers* and *non-completers* across all courses and runs, in absolute values.

A further visual analysis of the spread of registration dates is shown in Fig. 3, where the number of *completers* (in small green dots) and *non-completers* (in large red dots) are placed side-by-side, for each registration date. Beside the larger spread of the *non-completers* in terms of date, it can be clearly seen that their numbers are much larger as well (visually confirming that only less than 3% of the students actually complete). As these two spreads are at such very different scales, this data is further analysed separately in Fig. 4 (for *completers*) and Fig. 5 (for *non-completers*). The images show that, surprisingly, the shapes of the two graphs are relatively similar: beside the peak around the actual course starting date, there is a peek somewhere around 90-100 days after the course starting date.

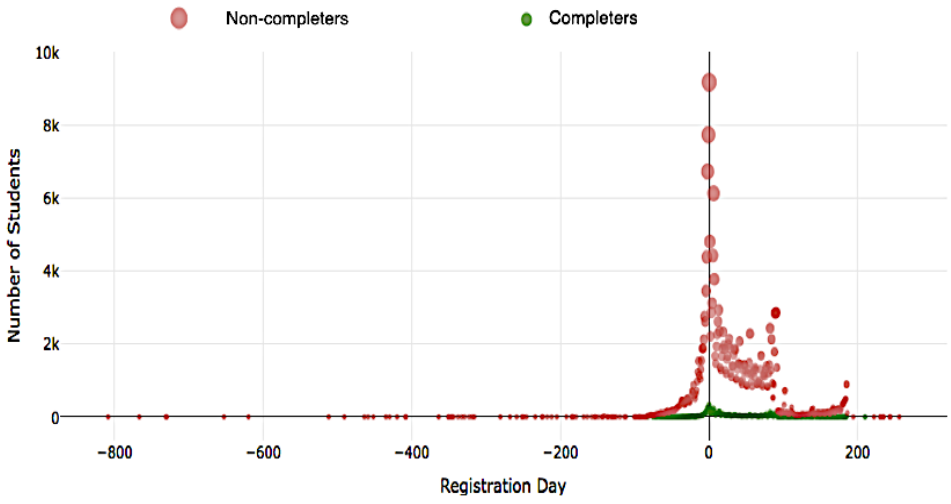

Fig. 3. Completers (green) versus non-completers (red) across all courses and runs, in absolute values.

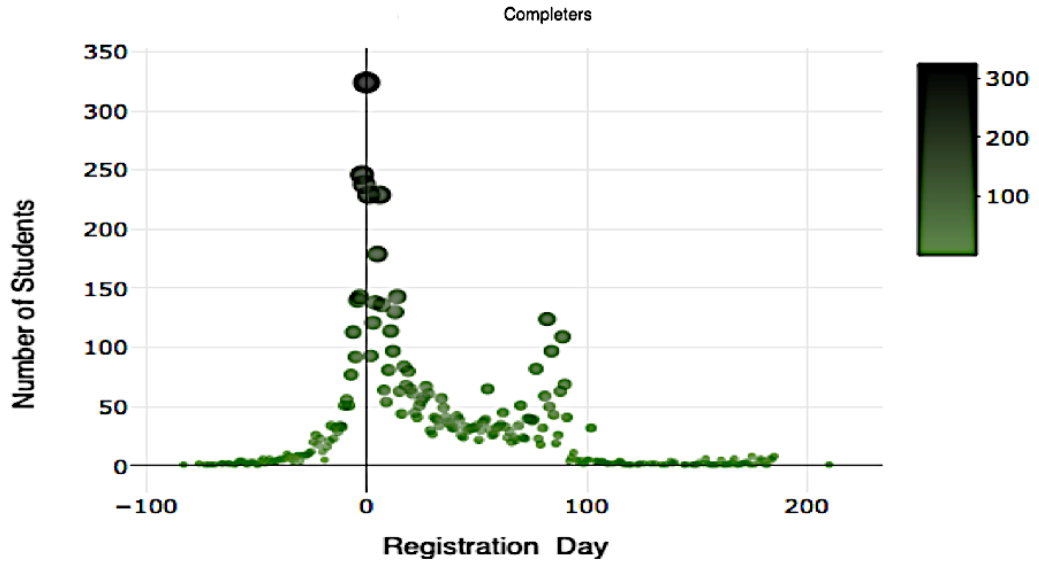

Fig. 4. Completers and their registration dates.

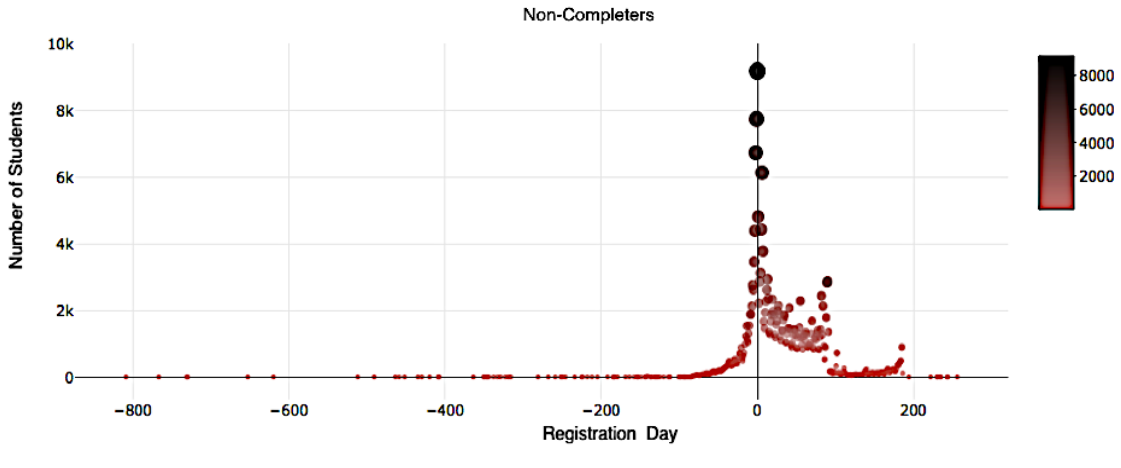

Fig. 5. Non-completers and their registration dates.

Thus, we analyse this data further, taking 90 days later as one transition point, and using its symmetrical counterpoint of 90 days earlier as another. The latter results from Fig. 4, where *completers* tend to disappear around that date. Thus, we specifically look at very early registrations (initially about 3 months – 90 days – in advance), late ones (1 month – 30 days – after course start) as well as the period in-between. Table 1 further shows these initial results for the overall cohort for all registrations. It can be seen that the averages shift considerably, with early registrations averaging at 142.4 days before course start for *non-completers*, who register a significant 15 days earlier than *completers* (even with Bonferroni correction at $p<.0167$); late registrations averaging at 49.46 days after course start, with *non-completers* registering about 5 days later, on average, than *completers* (significant at $p<0.05$ only). The overwhelming majority of *completers*, however, are in the middle region (7016/7437 or 94%). They register, on average, 24.54 days in advance, with *completers* registering about 1 day later than non-completers (but this is not significant). Fig. 6 helps visualise this data, for the three periods. The total numbers are less informative (Fig. 6, left side), as the number of *non-completers* dominates the overall numbers. Thus, we use the percentage view (Fig. 6, right side), which shows that there is a larger percentage of *completers* than *non-completers* who register closer to the course starting time, and a smaller percentage of *completers* who register

very early, or very late. However, the figure also shows that the majority of both *completers* and *non-completers* register in the identified central period.

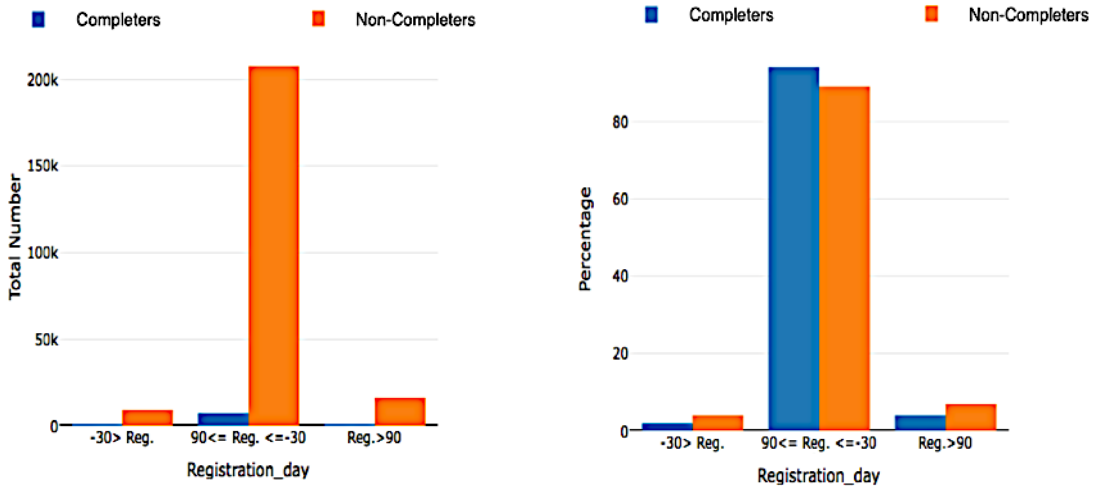

Fig. 6. Completers (in blue) and non-completers (in orange) visualised as total numbers (left) and as a percentage (right) for the initial three periods identified in Table 1.

Thus, it is clear that the central period needs further analysis, as, additionally, the statistical results and especially polarity start changing there (as shown in Table 1). Also, the labels 'early' and 'late' have been applied based on visual information only. Thus, we considered defining periods more rigorously, based on the features of the data, starting with Avg.=30.47, the overall average number of days in advance of the course start that students register on, as well as the overall standard deviation, σ. Interestingly, the 'early' (P1) and 'late' (P5) periods remain relatively similar - albeit better supported by the data - at 99.9 days in advance, and 38.96 days after, respectively, confirming our initial intuition. The results for these periods also remain relatively similar, as can be seen in Table 2.

Table 2. Periods identified based on σ, the standard deviation; 'Reg.' stands for registration date; 'Avg.' stands for average.

Avg.	***30.47***	**P1**	**P2**[99.9, 53.6]	**P3[53.6, 7.33]**	**P4**[7.33, 8.96]	**P5**
σ	***46.29***	Reg.> Avg. + 3/2σ =~99.9	Avg.+3/2σ <= Reg. < Avg. + 1/2σ	Avg. + 1/2σ <=Reg. <=Avg. - 1/2σ	Avg.-1/2σ < Reg. <= Avg. - 3/2σ	Avg. -3/2σ> Reg.=~-38.96
ALL	**Data size**	13941	52744	74489	93264	6130
	Avg.	**151.25**	**74.27**	**27.97**	**-4.59**	**-57.17**
	Var.	842.91	143.78	170.81	99.97	1187.73
	Max	256	99	53	7	-39
	Min.	100	54	8	-38	-809
Completers	**Data size**	198	1669	2392	3091	87
	Avg.	**140.75**	**75.5**	**25.44**	**-2.92**	**-51.24**
	Var.	1030.78	133.96	166.96	72.96	99.58
	Max.	210	99	53	7	-39
	Min.	100	54	8	-38	-83
Non-completers	**Data size**	13743	51075	72097	90173	6043
	Avg.	**151.4**	**74.23**	**28.06**	**-4.65**	**-57.26**
	Var.	838.67	144.06	170.72	100.8	1202.9
	Max.	256	99	53	7	-39
	Min.	100	54	8	-38	-809
Who registers earlier?		Non-Completers	Completers	Non-Completers	Completers	Completers
Wilcoxon's p		$p = 1.29e^{-06}$	$p = 4.75e^{-05}$	$p < 2.2e^{-16}$	$p < 2.2e^{-16}$	p = 0.041

However, now we can analyse in more details the middle period, by splitting it into 3 parts: the centre is half a deviation (*1/2 σ*) from the overall average *Avg.* both ways, and the sides contain the remaining periods, up to *3/2 σ*, each way. Here, we see some very interesting and potentially surprising fine-grained results: P2 *completers* and *non-completers* have only 1, however, significant, day, on average, between them. Interestingly, P3 and P4 show strong significant differences between the *completers* and *non-completers*, of opposite signs (2.62 and -1.73, respectively). Thus, *completers* register earlier in P2, later in P3 and earlier in P4. Fig. 7 shows that, for all three middle periods, the percentage of *completers* is consistently larger than the percentage of *non-completers*. However, it shows that, for both *completers* and *non-completers*, both their absolute numbers (left) and their percentages (right) grow steadily between periods P2, P3 and P4, with their peak in P4.

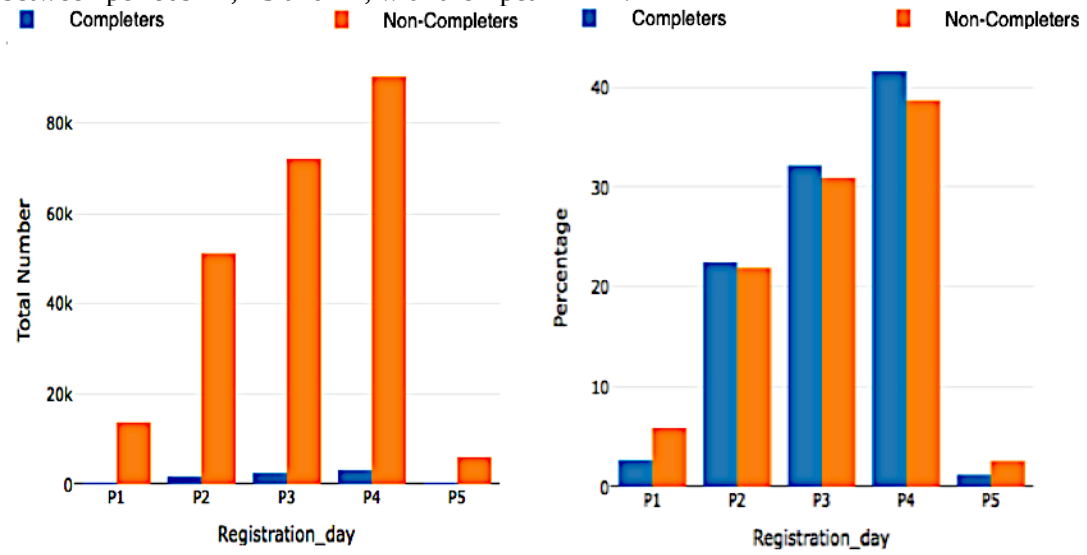

Fig. 7. Completers (in blue) and non-completers (in orange) visualised as total numbers (left) and as a percentage (right) for the five periods identified in Table 2.

4. Discussion and Extracted Rules

The results obtained are worthy of discussion, because some were not as straightforward as we initially expected. As in previous research [3, 8, 12, 13, 14, 15, 20], in our Warwick courses there are a substantial number of students who don't complete. In answering **RQ1**, indeed, registration time is a significant predicting factor. However, there is, for instance, not a simple answer to the question if the students should register early or late. Whilst we initially expected that "the early bird catches the worm" and thus students who register early would have a higher chance of completion, the general answer is in fact that, on average, registering later seems to be more advisable. Specifically, and interestingly, we were able to find explicit periods of time, related to the course starting date, for which this question can be answered in a statistically significant way, i.e.: **P1** (99.9 days before the start of the course); **P2** (99.9 to 53.6 days before); **P3** (53.6 to 7.33 days before); **P4** (7.33 before to 38.96 days after).

Intuitively, if students register too early (here, above 3 months in advance, covering P1), this is not very beneficial, as they may possibly forget that they have done so in the first place, so it is 'better' during this period to be one of the ones who register later, rather than earlier. Thus, being somewhat closer to the actual start of the course, when registering, is more desirable. On the other hand, if students register during a slightly closer time to the start (between 3 and above 2 months before the course has started), then, students completing or not are close, whilst it is slightly better to register earlier. However, even closer to the starting point (about 2 months before course start, up to about one week before), students enrolling slightly later are again more likely to complete. Interestingly, just around course start (1 week before course start up to about 1 month after course start) it suddenly becomes better to enrol slightly

earlier. This surprising result may be explained by noting that, for most of the time, it translates to registration being more likely to lead to a successful outcome if it is closer to the course starting date. The only exception is period P2, which would need further analysis in future research. Finally, for students enrolling too late (over 1 month after the course has started), it is again better for students who enrol closer to the starting date (thus earlier), but this difference is significant only without the Bonferroni correction.

Based on these results, we can further address **RQ2** as follows. The teacher could analyse the data very early on, and give specific customised feedback to students. As the registration time is known before or just after the start date, students could be advised to only register when they are quite sure about attending the course, and as close as possible to the start of the course. They should possibly be also given the choice to deregister, if they have lost the interest, to give a teacher a better and more realistic picture of the actual cohort to follow the course. For students who are registering too late, they should possibly be notified of further times the course is run, and let known in advance that they would have to put in an additional effort, if they really want to complete the course - possibly offering them simplified material, or other type of support, to catch up.

Alternatively (or at the same time with the tutor intervention), an intelligent tutoring MOOC extension could implement some rules to automatically deliver such messages to the students (e.g., Table 3). The table shows also that students can not only be given messages, but also be supported with additional resources, or more tailored resources, when they are starting, for instance, late. This is especially interesting, as the majority of the students belong to this category, according to our studies. For early registration, storing the course information as soon as possible in the agenda of the students and ensuring that no other overlap is occurring by omission is important (of course, other overlaps outside the influence of the students may still exist). Further development of such adaptation rules remains for future research, although defining periods centred on the start date, and moving standard deviations from it, seems promising. Finally, whilst the research questions posed in section 1 are answered, further answers can be sought in future work, as briefly mentioned in section 5.

Table 3. Rules in pseudo-code based on registration date.

IF registration_date < 3 months before start date (BSD) **THEN** recommend (“Please consider registering closer to the start of the course. Would you like to be reminded of this a couple of months before the course start? Would you like to have the date automatically registered to your Google calendar?”)
IF registration_date in (2 months BSD to 1 week BSD) **THEN** recommend (“Please consider confirming your registration closer to the course start. Would you like a reminder a week before the course starts? Would you like to have the date automatically registered to your Google calendar?”)
IF registration_date in (1 week BSD to 1 month after start date (ASD)) **THEN** recommend (“As you have registered to the course after its start, please note that you would have to put in much more focussed work in order to complete. If you prefer to opt to enrol the next time this course starts, please let us know. Please also consider visiting these links for additional support.”)
IF registration_date > 1 month ASD **THEN** recommend (“You have registered very late to the module. You are strongly recommended to consider taking this course at a later date.”)

4.1. Limitations

The results above are interesting and, to the best of our knowledge, not tackled before.

However, they come, as in any research, with some caveats which need mentioned. Firstly and importantly, the variances for the five periods (P1 to P5) in
Table **2** (especially for P1 and P5), as well as in Table 1, are very large. This is consistent with the data spread, as can be seen in Fig. 2. In the latter, it can be clearly seen that, especially for non-completers, the spread of the registration date is quite wide. This is less so for the completers, however, as can be seen on the left side of Fig. 2. This large variance could diminish the value of the statistical significance of the results obtained; this could be further indirectly affected by the large size of our sample (see, e.g., [10]). However, as [10] recommends, we have also visualised the data (as in Fig. 2), and have been able to see that completers are less spread than non-completers. What this actually means is that statements about completion are more likely to be statistically significant than statements about non-completers. Having however a binary range makes things easier, as this means that we could, in principle, make statements about non-completers, just by using the opposite as would be recommended for completers. Possible further research could look into eliminating the outliers; however, this needs done with care, as important information should not be lost in the process. For the latter reason, and for avoiding sampling errors, we have opted for this paper to keep all students in.

Furthermore, even if there is a measurable difference in the distribution of the registration time and completion data, the possibility exists that this is not causal, i.e., it is possible that the date of submission is not the cause for the completion or non-completion of the students. For instance, it could be that a certain type of students, more inclined to complete the course, tends to register at a certain time. Thus, suggesting to students to alter their registration behaviour might not be enough. Our paper already takes this into account, by including suggestions to visit additional material, etc. This could be further analysed with cause-effect relationship modelling methods. For example, the partial least squares path modelling method (PLS-SEM [12]) could be used to determine a causal relationship. We certainly do not expect, for instance, for the registration date to be the only reason for student dropout, as previous research has already shown [3, 14, 18, 20]. Nevertheless, we have clearly found some statistically significant correlation, which can be used to raise awareness of both students and teachers, and to be able to intervene early on. These findings can be applied together with other findings from research on other parameters influencing dropout, once the course has started. However, as can be seen from our research, the actual synchronous course delivery only takes a very brief proportion of the time when students are registering (and interacting) with a MOOC, and thus other mechanisms of detection of issues need to be implemented.

5. Conclusion and Future Research

This paper tackles the important and challenging issue of predicting student dropout as well as completion, which are the most targeted issues in research relating to MOOCs. However, most studies (rather predictably) analyse the course whilst it is running. We argue here that, in some cases, this might be too late. Thus, importantly, this paper presents the results of a study aiming to discover *if there are factors that can be identified before the students even start the course*, to predict which enrolled participants will not complete the MOOC and, possibly, take actions. The study is based on the analysis of a *large data set of FutureLearn MOOC users over several courses, each with several runs*. Our results show that *completion can be predicted based on the date of the registration.* We perform a fine-grain analysis on this phenomenon, based on our preliminary findings. Interestingly, we *detect specific periods* for which it is more likely to complete for students registered (relatively) early, as well as periods for which the opposite is true. We show that these periods are intrinsically linked to the course start-date. We show how these findings can lead to personalisation strategies based on the *earliest possible detection of potential issues*. Additionally, this research is applied to a less explored MOOC platform, FutureLearn. Unlike many of its counterparts in other parts of the world, FutureLearn has been arguably based, from the start, on solid pedagogical foundations, which make it specifically

interesting for education-related research[4]. However, for this paper, the results we obtain are founded on features shared by all MOOCs, such as the information on the date of the registration of the students, as well as the information on their having completed all allocated tasks. Thus, we can claim that our results have a more generic impact. Furthermore, as we address the research question via a genuinely large-scale experiment involving several subjects, in a truly longitudinal study, reaching over several iterations of all courses considered, we further ensure the generality of our claims.

Potential future work has several dimensions: firstly, different parameters we have collected but not discussed have been [7, 23] and will continue to be analysed, to predict the behaviour of students within further longitudinal, data-rich studies; secondly, machine learning algorithms will be considered (including the popular deep neural networks) for predicting relations which may be less obvious from a statistical data analysis; thirdly, specific rules will be refined and new ones defined for adaptation in MOOCs, based on our data-driven approach and the findings from it, and their implementation will be pursued and evaluated – starting with the evaluation of the rules suggested in this paper.

Acknowledgements

We would like to thank Nigel Smith from FutureLearn and Ray Irving from Warwick Business School for providing the Warwick FutureLearn data and encouraging us to explore it.

References

1. Atapattu, T., K. Falkner, and H. Tarmazdi. Topic-wise Classification of MOOC Discussions: A Visual Analytics Approach. in EDM. 2016.
2. Atenas, J., Model for democratisation of the contents hosted in MOOCs. International Journal of Educational Technology in Higher Education, 2015. 12(1): p. 3-14.
3. Balakrishnan, G. and D. Coetzee, Predicting student retention in massive open online courses using hidden Markov models. Electrical Engineering and Computer Sciences University of California at Berkeley, 2013.
4. Barba, P.G.d., G.E. Kennedy, and M.D. Ainley, The role of students' motivation and participation in predicting performance in a MOOC. Journal of Computer Assisted Learning, 2016. 32(3): p. 218-231.
5. Chua, S.M., et al., Discussion Analytics: Identifying Conversations and Social Learners in FutureLearn MOOCs. 2017.
6. Coffrin, C., et al. Visualizing patterns of student engagement and performance in MOOCs. in Proceedings of the fourth international conference on learning analytics and knowledge. 2014. ACM.
7. Cristea, A. I. et al. How is learning fluctuating? FutureLearn MOOCs fine-grained temporal Analysis and Feedback to Teachers, ISD 2018.
8. Daniel, J., Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. Journal of interactive Media in education, 2012. 2012(3).
9. Dillon, J., et al. Student Emotion, Co-occurrence, and Dropout in a MOOC Context. in EDM. 2016.
10. Figueiredo Filho, D.B., et al., When is statistical significance not significant? Brazilian Political Science Review, 2013. 7(1): p. 31-55.
11. Guo, P.J., J. Kim, and R. Rubin. How video production affects student engagement: An empirical study of mooc videos. in Proceedings of the first ACM conference on Learning@ scale conference. 2014. ACM.
12. Hair, J.F., C.M. Ringle, and M. Sarstedt, PLS-SEM: Indeed a silver bullet. Journal of Marketing theory and Practice, 2011. 19(2): p. 139-152.

[4] https://about.futurelearn.com/terms/research-ethics-for-futurelearn

13. Jordan, K., Massive open online course completion rates revisited: Assessment, length and attrition. The International Review of Research in Open and Distributed Learning, 2015. 16(3).
14. Kloft, M., et al. Predicting MOOC dropout over weeks using machine learning methods. in Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. 2014.
15. Koller, D., Ng, A., Do, C., Chen, Z.: Retention and intention in massive open online courses: In depth. Educause review 48, 62-63 (2013).
16. Liang, J., C. Li, and L. Zheng. Machine learning application in MOOCs: Dropout prediction. in Computer Science & Education (ICCSE), 2016 11th International Conference on. 2016. IEEE.
17. Liñán, L.C. and Á.A.J. Pérez, Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. International Journal of Educational Technology in Higher Education, 2015. 12(3): p. 98-112.
18. 1st Int. Conference on Learning Analytics and Knowledge, Banff, Alberta, February 27–March 1, 2011, cited in George Siemens and Phil Long, "Penetrating the Fog: Analytics in Learning and Education," EDUCAUSE Review, 46(5) (Sep./Oct. 2011).
19. Mohamed, M.H. and M. Hammond, MOOCs: a differentiation by pedagogy, content and assessment. Int. J. of Information and Learning Technology, 2018. 35(1) 2-11.
20. Rosé, C.P., et al. Social factors that contribute to attrition in MOOCs. in Proceedings of the first ACM conference on Learning@ scale conference. 2014. ACM.
21. Shi, L. and Cristea, A. I., Demographic Indicators Influencing Learning Activities in MOOCs: Learning Analytics of FutureLearn Courses, ISD 2018.
22. Sinha, T., et al., Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. arXiv preprint arXiv:1407.7131, 2014.
23. Sneddon, J., et al., Development and impact of a massive open online course (MOOC) for antimicrobial stewardship. Journal of Antimicrobial Chemotherapy, 2018.
24. Vigentini, L., M. León Urrutia, and B. Fields. FutureLearn data: what we currently have, what we are learning and how it is demonstrating learning in MOOCs. 7th International Learning Analytics & Knowledge Conference. 2017. ACM.
25. Vivian, R., K. Falkner, and N. Falkner, Addressing the challenges of a new digital technologies curriculum: MOOCs as a scalable solution for teacher professional development. Research in Learning Technology, 2014. 22(1): p. 24691.
26. Watkins, D. Open educational resources movement gains speed. 2017 opensource.com/article/17/10/open-educational-resources-alexis-clifton. [last accessed 19/06/2018]
27. Wen, M., D. Yang, and C. Rose. Sentiment Analysis in MOOC Discussion Forums: What does it tell us? in Educational data mining. 2014.
28. Zhu, M. Sari, A., Lee, M.: A systematic review of research methods and topics of the empirical MOOC literature (2014–2016). The Internet and Higher Education, Elsevier, https://doi.org/10.1016/j.iheduc.2018.01.002 (2018).