in memory of Luigi Colazzo

Je-LKS

Journal of e-Learning and Knowledge Society

The Italian e-Learning Association Journal

SIe-L

Società Italiana di e-Learning

LEARNING ANALYTICS: FOR A DIALOGUE BETWEEN TEACHING PRACTICES AND EDUCATIONAL RESEARCH

n. 3
2019
SEPTEMBER

Je-LKS
Journal of e-Learning
and Knowledge Society

www.sie-l.it
www.je-lks.org

The Journal is issued online three times per year.

Je-LKS is an Open Access Online publication. This means that everybody can free access online to abstracts and full lenght articles.

Libraries or researchers, can subscribe a Reprint Service Subscription. The Reprint Service Subscription is on a yearly basis and you can receive the three printed year issues for a 121€ annual fee (including taxes and shipping expenses).

In order to join SIe-L:

- segreteria@sie-l.it

- Tel. +39 052 2522521

For more information visit www.sie-l.it

Registration at the Rome Court in the pipeline.

eISSN: 1971 - 8829 (online)

ISSN: 1826 - 6223 (paper)

Resp. dir. Aurelio Simone

ANVUR Ranking: A-Class for Sector 10, 11-D1 and 11-D2

To the authors:

paper can be addressed to:
www.je-lks.org

Je-LKS

Journal of e-Learning and Knowledge Society

The Italian e-Learning Association Journal

Vol. 15, n. 3, 2019

LEARNING ANALYTICS: FOR A DIALOGUE BETWEEN TEACHING PRACTICES AND EDUCATIONAL RESEARCH

Je-LKS
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal
PEER REVIEWED PAPERS
LEARNING ANALYTICS: FOR A DIALOGUE BETWEEN TEACHING PRACTICES AND EDUCATIONAL RESEARCH
Vol. 15, n.3, 2019
ISSN: 1826-6223 | eISSN: 1971-8829

PREDICTIVE MODEL SELECTION FOR COMPLETION RATE IN MASSIVE OPEN ONLINE COURSES

Annamaria De Santis, Katia Sannicandro, Claudia Bellini, Tommaso Minerva

University of Modena and Reggio Emilia
{annamaria.desantis; katia.sannicandro; claudia.bellini; tommaso.minerva}@unimore.it

Keywords: MOOCs, Predictive Model, User Profile, Completion Rate, Learning Analytics

In this paper we introduce an approach for selecting a linear model to estimate, in a predictive way, the completion rate of massive open online courses (MOOCs). Data are derived from LMS analytics and nominal surveys. The sample comprises 722 observations (users) carried out in seven courses on EduOpen, the Italian MOOCs platform. We used 24 independent variables (predictors), categorised into four groups (User Profile, User Engagement, User Behaviour, Course Profile). As response variables we examined both the course completion status and the completion rate of the learning activities. A first analysis concerned the correlation between the predictors within each group and between the different groups, as well as that between all the dependent variables and the two response variables.
The linear regression analysis was conducted by means of a stepwise approach for model selection using the asymptotic information criterion (AIC). For each of the response variables we estimated predictive models

for citations:
De Santis A., Sannicandro K., Bellini C., Minerva T. (2019), *Predictive Model Selection for Completion Rate in Massive Open Online Courses,* Journal of e-Learning and Knowledge Society, v.15, n.3, 145-159. ISSN: 1826-6223, e-ISSN:1971-8829
DOI: 10.20368/1971-8829/1135034

using the different groups of predictors both separately and in combination.
The models were validated using the usual statistical tests.
The main results suggest a high degree of dependence of course completion and completion rate on variables measuring the user's behavioural profile in the course and a weak degree of dependence on the user's profile, motivation and course pattern.
In addition, residual analysis indicates the potential occurrence of interaction effects among variables and non-linear dynamics.

1 Introduction

The three major themes comprised by learning analytics are predictors and indicators, visualisations and interventions. Studies belonging to the first theme aim «to establish a predictive model» and «to identify specific correlations between user actions in online tools and academic performance», as well as among skills, self-regulated learning and learning strategies (Gasevic *et al*., 2019).

Therefore, to carry out such research, we need to identify the data set and analysis methods.

Malcolm Brown (2012) identifies three kinds of data to design an LA application:

- *dispositional indicators*, which are features that students have before the course that can predict his/her involvement in the activities, including age, gender, learning experiences, financial status, psychological measures, "learning power", learning styles and personality types.
- *activities and performance indicators*. The author defines these as «digital breadcrumbs left by learners as they engage in their learning activities and make their way through the course sequence» (p. 2). Some examples of these types of data can be logins, time spent, forum posts, grades and quiz scores.
- *student artefacts*, namely essays, forum posts, media productions and other objects produced by students while attending the course.

In explanatory and predictive modelling, linear regression represents one of the conventional approaches used for building predictive models, together with logistic regression, nearest neighbour classifiers, decision trees, neural networks and so forth (Brooks & Thompson, 2017).

In the analysis of data coming from massive open online courses (MOOCs), linear regression has frequently been used in previous studies to estimate the relationship among data coming from LMS, surveys or students' accounts.

In 4 edX MOOCs, Philip Guo and Katharina Reinecke (2014) have analysed correlations and conducted multiple linear regression among three categories of

variables: i) demographics, including age, years of education, country, student-teacher ratio from UNESCO documents (number of students divided by number of teachers); ii) motivation, comprising certificate, grade, coverage (number of learning sequences visited by students), discussion forum events; and iii) navigation, specifically backjumps and textbook events.

«Age, gender, education level, motivation for taking the MOOC, working in groups, and intention of completing the course» (Zhang *et al*., 2019, p.143) are the independent variables in research realised on MOOCs offered on the Coursera platform. This research aimed to identify learners' profiles and their preferences in group working and in attending MOOC, as well as to predict if demographic and motivational elements affect course completion. Other studies have focused on the influence of the instructional design of courses (Jung *et al*., 2019) or on participation and motivation (Brooker *et al*., 2018) across different disciplines (Williams *et al*., 2017).

These investigations tell us, among other things, that countries and age can affect the means of navigating among learning activities. Working in groups does not affect course completion. Motivation varies in courses related to Humanities or STEM, and interaction with course content can help predict student learning.

In this study, we perform a regression model selection to define the relationship between students' and courses' profiles and course completion. We used data from EduOpen, the Italian MOOC platform, and those collected through a survey.

2 Materials and methods

We performed an empirical study to understand how the features of MOOCs and users' profiles, motivation and behaviour affect course completion in order to define a linear model to predict completion rates, starting from analysed phenomena regarding courses and learners.

2.1 Data

The data come from EduOpen, the Italian MOOCs platform, including 22 universities. This project, funded by the Italian Ministry of Education, was launched in 2016. Today, the users registered to the portal number more than 55,000. EduOpen is a Moodle-based platform; the courses published until the present day are more than 250, are divided into six categories and offered in two fruition modalities: self-paced or tutored.

This study involves seven of the courses that show differences in category, level, effort, language and fruition mode. Three of the courses selected belong to

the category "Science", four are tutored, and the same number are in the Italian language and for beginners. 1,508 learners were enrolled in these courses, and the mean completion rate was 23%, a higher value than the usual percentage revealed in other portals and researches. The courses deal with very different themes (from ethnobotany to gender violence, robots, nanoparticles, and history of sports) and provide qualitative and quantitative assessments, as appropriate.

We collected users' data through:

- Moodle reports that give us information about users' log, single activity completion and general course completion;
- a questionnaire administered to users before starting the course, composed of 15 closed questions, of which the last two are designed with multiple items. The survey investigates the demographics and motivations of learners.

922 students (61.1% of enrollers) replied to the survey. Removing N/A, we obtained a data set with 722 observations corresponding to 722 users (students).

2.2 Variables

For each student, we collected variables from a survey, courses, and log data.

Independent variables (predictors) were divided into four groups: User Profile, User Engagement, User Behaviour and Course Profile.

As response variables we considered a dichotomic variable (Certificate Download) reporting if the user completed the course and downloaded the course certificate and the completion rate of the tracked activities in the course. The full set of variables, together with their summary statistics, is reported in Table 1.

Table 1
VARIABLES LIST

GROUP	PREDICTOR	NOTES	GROUP	PREDICTOR	NOTES
User Profile	GENDER	The gender of the user	**User Engagement** (all values refer to an individual estimation)	EFFORT	Estimated effort (hours) to complete the course
	DEGREE	The highest level degree		PRE. KNOWLEDGE	Estimated level of knowledge in the field of the course
	LANGUAGE	Native language		DROPOUT_ TOT	Level of disposition to abandon the course
	AGE	The age of the user/student		DROPOUT_ INT	Level of disposition to abandon the course by lack of interactions with instructors/peers
	MARRIED	Married or common-law partner		DROPOUT_ LEA	Level of disposition to abandon the course by lack of learning design
	CHILDREN	Has children		DROPOUT_ NAV	Level of disposition to abandon the course by lack of navigation
	TRAINING	Attending an official degree		MOTIVATION	Level of motivation to attend the course
	WORKING	Working status	**Course Profile**	CTUTORED	Whether the course is tutored or self-paced
	SECTOR	Working sector		CCAT	Course category
	DIGITAL	Digital competencies		CLANG	Course language
User Behaviour	CLICKS_ TRACKED	Rate of clicks on tracked activities		CHOUR	Estimated effort (from the instructor) to complete the course
	CLICKS_TOTAL	Rate of clicks on overall activities		CLEVEL	Difficulty level of the course
RESPONSE VARIABLES					
CERTIFICATE	The user completed the course AND downloaded the certificate (binary variable)		**CRATE**	Rate of the tracked activities completed by the user	

2.3 Analysis methods

After depicting the data set through conventional descriptive statistical tools, we examined the correlation within each group of predictors and between each

predictor and the two response variables. We then conducted a full stepwise analysis to select and fit the linear regression models. We considered two cases, one in which the response variable was certificate download, and the other in which it was completion rate.

The stepwise approach is both backward and forward. The stepwise selection algorithm adds and removes the predictors to obtain a stable set of variables and the optimal final regression model based on the maximisation of the asymptotic information criterion (AIC). AIC is an indicator that balances the number of observations, the number of independent variables introduced and the variance of the residuals in a model with independent variables (Akaike 1969, 1978; Paterlini & Minerva, 2010).

For each stepwise linear regression model, we reported the R-squared adjusted, Residual standard error, F-statistics and model p-value.

Within each linear regression model, we evaluated the value of the intercept and the predictors' coefficients, and for each of them we estimated the standard error, t test, and p-value.

The usual residual analysis was carried out to analyse the model's goodness. We used the Shapiro-Wilk test, the Anderson-Darling test and the Lilliefors (Kolmogorov-Smirnov) test together with the graphical Q-Q plot to test the normal distribution of the model residuals.

In the last part of the study, based on the previous results, further gradual stepwise regression analyses were carried out, including only variables from selected groups or sets of groups.

As a computational environment, we used R/R-Studio and the following R libraries: tidyverse, caret, leaps, MASS, kableExtra, data.table and summarytools.

The full dataset and a R-Markdown script file are available as supplementary material to this paper.

3 Results

3.1 Overview of the sample

The gender representation of the sample contained two groups of almost the same size (55.1% women, 44.9% men). Nearly 90% of students spoke Italian, 42.4% were married/cohabiting and 31.0% had one or more children.

The mean student age in our sample was 38 years; 53.6% of learners had stable work, 22.6% were occasional workers and 13.4% were unemployed. 37.5% had finished secondary school and 58.6% had a tertiary educational qualification (equal or more than a bachelor's degree). At the time of investigation, 41.6% were not attending a university course. Instead, 30.6% were working towards Bachelor's, Master's or doctoral degrees.

Based on these results, we may expect on EduOpen the presence of two different sets of students. The first shows younger university students: occasional workers with no family responsibilities. The second (and larger) contains adults more committed to taking care of family and lifecare issues.

Regarding the engagement variables, the survey respondents indicated their estimated effort to complete the courses as being on average 29 hours. This value was higher than that assigned by the EduOpen instructional designers, ranging from 14 to 25 hours. The mean motivation level to enroll on courses was 23.0 (SD=6.7, Range: 1.0-40.0) and the mean motivation level to drop out was 25.2 (SD=7.6, Range: 3.0-45.0).

Regarding students' behaviour, the mean number of clicks per tracked activities was 3.5 (SD=5.5, range: 0.1-28.3). If we consider all the activities and materials (tracked and not tracked), the mean number of clicks per activity/ document was 2.6 (SD=2.3, range: 0.1-14.3), and thus lower, as expected.

Moving to dependent variables, we can pinpoint that 34.5% of learners completed the course and downloaded the certificate.

On the other hand, we can see that more than 44.3% of users completed at least 90% of learning activities. We have about 10% of users/students who completed most of the course but did not finalize it by downloading the certificate. About 38.9% of students covered less than 20% of learning activities. About 16.8% of users completed more than 20% and less than 90% of the course.

The frequency distribution of the completion rate was nearly bimodal. The modal bin was between 90% and 100% (320 obs.), but the option related to a completion rate of less than 10% showed a frequency of 208 users. Therefore, we can distinguish students who even if enrolled did not attend courses at all and users who after completing at least 50% tended to finish their activities and acquire a certificate. A complete description is in supplementary material attached to this paper.

3.2 Intragroup correlation

As expected, the correlation coefficient (ρ) assumes high values between CRATE and CERTIFICATE (0.77), CLICKS_TOTAL and CLICKS_ TRACKED (0.92) and among variables of the group about course features that can be common to more than one course in the research.

In the other two groups, the correlation shows few significant associations.

In demographic phenomena, we can observe a correlation between AGE and MARRIED (0.50), CHILDREN (0.51), DEGREE (0.26), WORKING (-0.33), SECTOR (-0.34) and between MARRIED and CHILDREN (0.64), WORKING (-0.18), SECTOR (0.17). This evidence confirms that, as we assumed earlier,

older users are more likely to have a stable job, a tertiary education degree and a family.

In the block of engagement variables, correlations are significant among the four variables related to reasons to abandon a MOOC (DROPOUT_TOT, DROPOUT_INT, DROPOUT_LEA, DROPOUT_NAV). The explanation is in the fact that we evaluated these values by the items of the same group of questions in the survey.

However, ρ values deviate from 0 also between the 4 DROPOUT_x variables and MOTIVATION (ρ is between 0.18 and 0.34). Moreover, PRE. KNOWLEDGE of course themes slightly correlates to DROPOUT_INT (0.18) and MOTIVATION (0.32). Even if the ρ values are not far from 0, we can say that:

- the higher students' expectations for participating in a course, the more numerous the reasons for abandoning it;
- the more a student knows the course topics, the more he/she is motivated to enroll and to discuss with teachers and classmates.

3.3 Response vs predictors correlation

We present here the correlation between dependent and independent variables in Table 2.

In most cases, the ρ values are close to 0 and we can observe a weak linear relationship between variables. Except for the group User Behaviour, where the correlation coefficient has values between 0.64 and 0.86 (higher for CLICKS_ TRACKED than CLICKS_TOTAL), in the other groups ρ is between -0.20 e 0.17. The correlation with GENDER, DEGREE, AGE and CHILDREN tell us that men, Italian students, adults and people with children have a slightly higher chance of completing courses. The values related to CERTIFICATE for these groups are slightly stronger than CRATE. The highest ρ values in the block User Engagement are recorded by variables PRE.KNOWLEDGE (CRATE -0.09, CERTIFICATE -0.09) and MOTIVATION (CRATE -0.17, CERTIFICATE -0.15).

Table 2
RESPONSES AND PREDICTORS CORRELATION COEFFICIENT (P)

GROUP	PREDICTOR	CERTIFICATE	CRATE	GROUP	PREDICTOR	CERTIFICATE	CRATE
User Profile	GENDER	-0.20	-0.13	**User Engagement**	EFFORT	0.00	0.03
	DEGREE	-0.03	-0.01		PRE. KNOWLEDGE	0.09	0.09
	LANGUAGE	-0.13	-0.09		DROPOUT_ TOT	0.04	-0.02
	AGE	0.16	0.12		DROPOUT_ INT	0.07	0.04
	MARRIED	0.07	0.04		DROPOUT_ LEA	0.02	-0.02
	CHILDREN	0.11	0.10		DROPOUT_ NAV	0.05	0.02
	TRAINING	0.04	0.04		MOTIVATION	0.15	0.17
	WORKING	-0.02	-0.01	**Course Profile**	CTUTORED	-0.08	0.02
	SECTOR	0.00	-0.02		CCAT	0.04	0.03
	DIGITAL	0.03	0.01		CLANG	-0.04	-0.03
User Behaviour	CLICKS_ TRACKED	0.64	0.77		CHOUR	-0.05	0.03
	CLICKS_ TOTAL	0.73	0.86		CLEVEL	-0.04	-0.03

These data show that if a relationship between the dependent and independent variables exists, it seems to be non-linear; the elements that most influence completion seem not to be found in the features of students but in their use of the portal (expressed in clicks).

3.4 Stepwise analysis: the general model

We used a stepwise approach, with AIC as model fitness function, to select the predictor set to consider in fitting a regression model among all examined variables. In this first stage, we performed a selection from the whole set of variables, considering the two cases for CERTIFICATE and CRATE prediction model. The selected models presented different predictors for the two response variables: seven for CERTIFICATE, 16 for CRATE.

Table 3 reports the regression results for the model. We show for each selected predictor the estimated value of the regression coefficient (β), t-test and p-value. The legend in Table 3 indicates that only a subset of variables reaches the required significance levels at 95% (starred).

Both models reached a level of significance of 95%; the CERTIFICATE model explains 57% of variations among variables (adjusted $\mathbf{R}^2$=0.5726), while the CRATE model explains the 75% (adjusted $\mathbf{R}^2$=0.7504).

To validate the two regression models, we performed the analysis of residuals. Normality tests on the residuals are not satisfactory because the p-value in Shapiro-Wilk, Anderson-Darling, Lilliefors (Kolmogorov-Smirnov) normality tests indicated that we can reject the hypothesis of normality.

Table 3

STEPWISE SELECTED REGRESSION MODEL FOR CERTIFICATE AND CRATE

REGRESSION MODEL FOR CERTIFICATE Residual standard error: 0.311 on 714 DF Adjusted R-squared: 0.5726 F-statistic: 139 on 7 and 714 DF, p-value: < 2.2e-16					**REGRESSION MODEL FOR CRATE** Residual standard error: 0.2165 on 705 DF Adjusted R-squared: 0.7504 F-statistic: 136.5 on 16 and 705 DF, p-value: < 2.2e-16				
Variable	**Coefficient**	**SE**	**t-test**	**p-value**	**Variable**	**Coefficient**	**SE**	**t-test**	**p-value**
(Intercept)	-0.084	0.069	-1.216	0.224	(Intercept) *	-0.247	0.085	-2.897	0.004
GENDER *	-0.112	0.024	-4.579	0.000	GENDER *	-0.049	0.017	-2.804	0.005
LANGUAGE *	-0.095	0.040	-2.369	0.018	DEGREE *	0.013	0.006	2.065	0.039
DIGITAL *	0.027	0.013	2.070	0.039	AGE *	-0.011	0.004	-2.446	0.015
DROPOUT_TOT *	0.004	0.002	2.536	0.011	CHILDREN	0.031	0.021	1.514	0.130
CLICKS_TOTAL*	0.151	0.005	29.477	0.000	SECTOR	-0.003	0.002	-1.526	0.127
CTUTORED *	-0.124	0.030	-4.147	0.000	EFFORT	0.001	0.000	1.715	0.087
CCAT *	-0.061	0.030	-2.034	0.042	DROPOUT_TOT	-0.013	0.007	-1.865	0.063
LEGEND: DF = Degree of Freedom; SE = Standard Error * = variable with p-value < 0.05 at 95% significance level					DROPOUT_INT	0.012	0.008	1.504	0.133
					DROPOUT_LEA	0.014	0.008	1.700	0.090
					DROPOUT_NAV	0.012	0.008	1.544	0.123
					MOTIVATION *	0.003	0.001	2.043	0.041
					CLICKS_TRACKED *	-0.042	0.009	-4.802	0.000
					CLICKS_TOTAL *	0.220	0.013	16.919	0.000
					CTUTORED *	0.064	0.026	2.517	0.012
					CCAT *	0.126	0.029	4.288	0.000
					CHOUR *	0.016	0.003	4.899	0.000

We obtained the same outcome by graphic display in Q-Q plot (Figure 1). We must refuse the hypothesis that model residuals follow a normal distribution. Consequently, we can assert that a predictive linear regression

model can partially explain the completion rate of courses and provide evidence for non-linear or interaction or for missing variables effects.

The fact that the residuals' distribution was not normal and that the regression model does not explain all observations in the data set requires further and in-depth analysis.

The explanations of these results can be seen in one or more factors:

- the variables are not exhaustive of the phenomena described in each block and are not calculated through significant parameters or scales;
- the relationship between the dependent and independent variables is not linear;
- there are interactions among the variables that we have not taken into account and that require further studies.

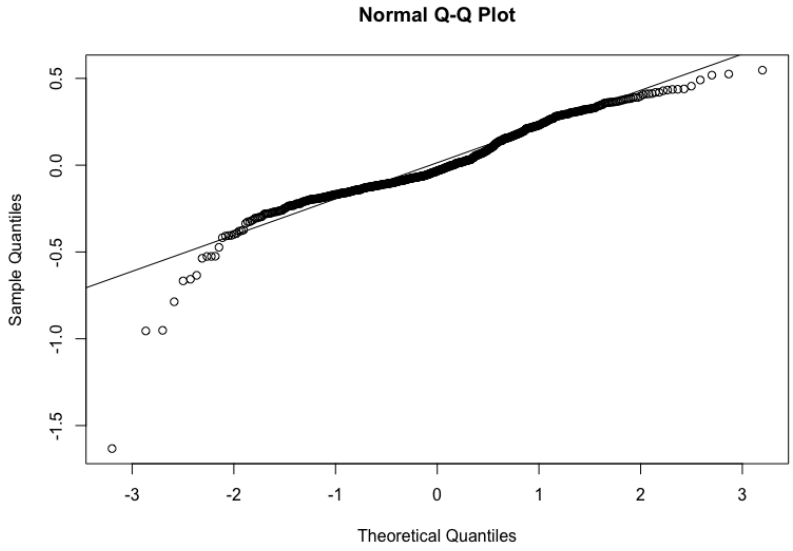

Fig. 1 - Normal Q-Q plot for residuals in CRATE selected model.

3.5 Stepwise analysis: partial models

As a final step and before planning further analysis, we re-ran the regression model selection by including one or more groups of predictors (Table 4). The goal was to better understand the role of each group of variables.

Table 4
PARTIAL REGRESSION MODELS FOR CERTIFICATE AND CRATE

Partial regression models for CERTIFICATE (* = variable p-value < 0.05 at 95% significance level)				**Partial regression models for CRATE** (* = variable p-value < 0.05 at 95% significance level)			
MODEL	**FORMULA**	**Adj-R2**	**F-statistic** ***(p-value)***	**MODEL**	**FORMULA**	**Adj-R2**	**F-statistic** ***(p-value)***
I - Target vs Profile + Course	CERTIFICATE ~ GENDER* + LANGUAGE* + AGE* + SECTOR	0.073	15.16 (6.6E-10)	I - Target vs Profile + Course	CRATE ~ GENDER* + LANGUAGE + AGE* + MARRIED + CHILDREN + TRAINING + CTUTORED* + CCAT*	0.044	5.181 (2.6E-03)
II - Target vs Profile + Course + Engagement	CERTIFICATE ~ GENDER* + LANGUAGE* + AGE* + SECTOR + DROPOUT_TOT + DROPOUT_INT + DROPOUT_NAV + MOTIVATION*	0.097	10.73 (2.5E-11)	II - Target vs Profile + Course + Engagement	CRATE ~ GENDER* + LANGUAGE* + AGE* + CTUTORED* + CCAT* + DROPOUT_TOT* + DROPOUT_INT* + DROPOUT_LEA* + DROPOUT_NAV* + MOTIVATION*	0.090	8.171 (1.37E-12)
III - Target vs Engagement + Behaviour	CERTIFICATE ~ CLICKS_TOTAL* + CLICKS_TRACKED* + PRE.KNOWLEDGE* + DROPOUT_TOT*	0.549	220.3 (< 2.2E-16)	III - Target vs Engagement + Behaviour	CRATE ~ CLICKS_TOTAL* + CLICKS_TRACKED* + PRE.KNOWLEDGE + DROPOUT_TOT* + DROPOUT_INT* + DROPOUT_LEA* + DROPOUT_NAV + MOTIVATION	0.737	253.2 (<2.2E-16)
IV - Target vs Engagement	CERTIFICATE ~ MOTIVATION	0.021	16.61 (5.1E-2)	IV - Target vs Engagement	CRATE ~ DROPOUT_TOT* + DROPOUT_INT* + DROPOUT_LEA* + DROPOUT_NAV* + MOTIVATION*	0.046	8.008 (2.3E-07)

V - Target vs Behaviour	CERTIFICATE ~ CLICKS_ TOTAL* + CLICKS_ TRACKED *	0.544	430.8 (< 2.2 E-16)	V - Target vs Behaviour	CRATE ~ CLICKS_ TOTAL* + CLICKS_ TRACKED *	0.733	988.9 (< 2.2E-16)

We first considered only groups related to variables whose values were known before beginning the courses. In the second row of Table 4, it is possible to see the regression models calculated with User Profile and Course Profile as predictors, while the third row contains results related to models that add the group of User Engagement to the two previous ones. Both the models for CERTIFICATE and CRATE explain a percentage of observations of less than 7%. Including the engagement variables, the value of Adjusted $\mathbf{R}^2$ increases by a very low percentage (2% for CERTIFICATE and 5% for CRATE).

Therefore, excluding the personal features of users and the general characteristics of the courses, we focused on groups related to the engagement and behaviour of users (third model of the tables): the results of Adjusted $\mathbf{R}^2$ show a condition similar to the total regression models described in the previous paragraph. After distinguishing the model contingent on User Engagement by that depending on User Behaviour (fifth and sixth rows), we can see that the last regression model of the tables (the one related to behaviours) explains almost the same completion rate of the model as that considering all the variables in our research.

The only factor that at the end of the analysis of regression models was in a stronger relationship with CERTIFICATE or (better) CRATE is represented by the number of clicks, which can be seen as students' participation in learning activities.

Conclusions

The aim of this study was to define a predictive (and adaptive) system that in a MOOC platform estimates (predicts) the completion of courses and percentage of completed activities according to students' demographics, engagement and behaviours, as well as the courses' features.

We performed an analysis of data collected from a survey and the reports of EduOpen LMS. We identified 24 independent variables in four blocks: User Profile, User Engagement, User Behaviour and Course Profile. The responses measured course completion (binary variable) and learning activities' completion rates (quantitative variable).

The findings of the study suggest that we can outline two learner profiles on EduOpen: on the one hand, young university students and, on the other, adult professionals. This result is confirmed by intragroup correlation, which

moreover highlights the relationship between pre-knowledge and motivation. The correlation between responses and predictors tells us that there is a weak linear relationship between variables, except for numbers of clicks (tracked and total), a factor more widely confirmed by stepwise regression run among all predictors and selected groups.

This first step leads us to continue the research with a further selection of variables to be included in the regression models and with an in-depth study of the types of function that express the relationships among variables. The use of genetic algorithms for regression modelling, including genetic algorithms for regressors' selection (GARS), or better yet genetic algorithms for regressors' selection and transformation (GARST) is proposed to determine not only the most adequate variables, but also the most appropriate mathematical transformations (Paterlini & Minerva, 2010). This will allow us to understand if a relationship among phenomena exists and what are its characteristics.

At the same time, this first finding shifts our attention from the profiles of courses and learners to the learning activities and materials within courses. The design of courses, interaction with contents, assessments and time spent represent at this point the elements to investigate in order to gain clearer explanations of the phenomena that data describe and to develop the long-term potential to intervene on the elements that we, as the portal administrators, manage in the production of MOOCs. These may include the audio-video quality and the length of the videolectures, the design of assessments, the automatic reminders, the completion indicators and the tools to support self-regulated learning, among other factors.

Learners usually attend MOOCs following autonomous and independent learning paths, sometimes in the list proposed by teachers, but in other cases according to an order chosen by themselves. The design of this particular typology of online courses must be planned very carefully, paying attention to automatic processes that are necessary for massive courses. However, at the same time and in order to reply to different learning styles, this should permit the students to participate in activities, learning with a high level of freedom.

Therefore, this research places the management and quality of the MOOCs at the centre of the debate.

The next variables to include in future studies are scores and assessments, the number of interactions with different materials in the portal (such as video lectures, documents, links, forums and collaborative activities) and time used to carry out each activity. These new indicators should provide a more comprehensive description of "what happens in our virtual classroom", providing explanations regarding the number of clicks recorded in this study as the fundamental element that can predict MOOC completion.

REFERENCES

Akaike A. (1969), Statistical predictor identification, *Annals of the Institute of Statistical Mathematics*, 22, 203-217.

Akaike A. (1978), Bayesian analysis of the minimum AIC procedure, *Annals of the institute of Statistical Mathematics*, 30, part A, 9-14.

Brooker A., Corrin L., de Barba P., Lodge J. & Kennedy G. (2018), A tale of two MOOCs: How student motivation and participation predict learning outcomes in different MOOCs, *Australasian Journal of Educational Technology*, 34 (1), 73-87.

Brooks C. & Thompson C. (2017), *Predictive Modelling in Teaching and Learning*, in: Lang C., Siemens G., Wise A. & Gašević D. (eds.), *Handbook of Learning Analytics*. 61-68, SOLAR: Society for Learning Analytics Research.

Brown M. (2012), *Learning Analytics: Moving from Concept to Practice* (EDUCAUSE Learning Initiative Briefs), Louisville (CO), EDUCAUSE.

Gašević D., Tsai Y.S., Dawson S. & Pardo A. (2019), How do we start? An approach to learning analytics adoption in higher education, *The International Journal of Information and Learning Technology*.

Guo P.J. & Reinecke K. (2014), Demographic differences in how students navigate through MOOCs, in: *Proceedings of the first ACM conference on Learning@ scale conference*. 21-30, New York, ACM.

Jung E., Kim D., Yoon M., Park S. & Oakley B. (2019), The influence of instructional design on learner control, sense of achievement, and perceived effectiveness in a supersize MOOC course, *Computers & Education*, 128, 377-388.

Paterlini S., & Minerva, T. (2010). Regression Model Selection using Genetic Algorithms. In V., Munteanu, R., Raducanu, G., Dutica, A., Croitoru, V.E., Balas, & A. Gravilut (Eds.), *Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing* (pp. 19-28). USA: WSEAS Press.

Williams K.M., Stafford R.E., Corlissb S.B. & Reillyc E.D. (2018), Examining student characteristics, goals, and engagement in Massive Open Online Courses, *Computers & Education*, 126, 433-442.

Zhang Q., Bonafini F.C., Lockee B.B., Jablokow K.W. & Hu X. (2019), Exploring Demographics and Students' Motivation as Predictors of Completion of a Massive Open Online Course, *International Review of Research in Open and Distributed Learning*, 20 (2), 140-161.