IJMSIT

International Journal of Multidisciplinary Studies and Innovative Technologies

e-ISSN: 2602-4888
dergipark.org.tr/en/pub/ijmsit
Research Article

2025, 9 (2), 201-209
DOI: 10.36287/ijmsit.9.2.3
Received: October 20, 2025; Accepted: November 12, 2025

Who Completes, Who Gets a Certificate? Digital Traces and Learning Outcomes in MOOCs

Rukiye Orman[1*], Hasan Cakir [2] and Nergiz Ercil Cagiltay[3]

[1]Ankara Yıldırım Beyazıt University, Ankara, Turkiye (rukiyeorman@aybu.edu.tr) (ORCID: 0000-0003-1385-0939)*
[2]Gazi University, Ankara, Türkiye (hasanc@gazi.edu.tr) (ORCID: 0000-0002-4499-9712)
[3]Hacettepe University, Ankara, Turkiye (necagiltay@gmail.com) (ORCID: 0000-0003-0875-9276)

Abstract – This study aims to investigate the factors affecting course completion and certification among participants in the context of Massive Open Online Courses (MOOCs). Logistic regression analyses were conducted based on participants' demographic characteristics and learning interactions in the online course "Elements of Structures" offered by the Massachusetts Institute of Technology (MIT). The analyses revealed that variables such as graduate education level, age groups (especially ages 45 and above), number of days of interaction, number of video plays, number of chapters completed, and writing activity had significant effects on the likelihood of both course completion and certification. The contribution of written interaction and content completion to success was found to be particularly significant. Model performance was evaluated using ROC curves and AUC (Area Under the Curve). The AUC values of 0.80 and 0.92, respectively, demonstrated high classification accuracy. The findings reveal that logistic regression analysis is an effective tool in developing predictive models for predicting user success on MOOC platforms and offer strategic recommendations for personalized, interaction-oriented design of learning environments.

Keywords *– Massive Open Online Courses (MOOCs), Educational Data Mining, Logistic Regression, Course Completion, Certification, Student Engagement*

Citation: Orman, R. et. al. (2025). Who Completes, Who Gets a Certificate? Digital Traces and Learning Outcomes in MOOCs. International Journal of Multidisciplinary Studies and Innovative Technologies, 9(2): 201-209.

I. INTRODUCTION

The integration of digital technologies into educational processes has led to significant changes in teaching and learning approaches; within this framework, Massive Open Online Courses (MOOCs) have emerged as an effective learning model that enables access to higher education content on a global scale. Thanks to their scalable structure, low cost, and time-independent delivery capabilities, MOOCs have transformed traditional education approaches, making learning processes more accessible and inclusive [1-3]. These platforms go beyond simply providing access to course materials; they also enable various interactive learning activities, such as taking exams, submitting assignments, and engaging in online discussions. Continuously recording these interactions in a digital environment yields large volumes of data, which has led to the development of an interdisciplinary field called Educational Data Mining (EDM). By analyzing data on student interactions, EDM enables detailed analysis of engagement patterns, achievement trends, and learning behaviors, thus contributing to the design of more effective learning environments [4, 5]. The opportunities offered by EDM are particularly noteworthy in predicting student success, monitoring learning processes, and developing early intervention strategies [6, 7].

Various data mining and machine learning methods, particularly logistic regression, reveal the factors that influence student performance and enable early identification of at-risk individuals. For example, Swacha and Muszyńska discuss using demographic data to predict early dropout rates in programming MOOCs and present logistic regression as an effective method for analyzing such predictors [8]. Liu et al. highlight using logistic regression to determine the impact of students' cognitive assets on their learning achievement, identifying significant predictors aligned with academic achievement [9]. This predictive modeling is vital because it allows educators to identify at-risk students early in the learning process and tailor interventions accordingly, as demonstrated by He et al., who focused on identifying at-risk students based on their behavioral patterns [10]. Such predictive modeling assesses academic achievement and helps educators develop interventions customized to student needs.

MOOCs environments offer rich learning experiences based on intercultural interaction by bringing together learners from different parts of the world on a digital platform. While these structures support equal opportunities in education in terms of accessibility and inclusiveness [11-13], existing literature reveals that learner interactions in MOOCs are mostly addressed at a superficial level and this data is not systematically integrated into instructional design. These studies generally focus on single performance indicators, such as completion rates, video viewing frequency, or forum activities. However, the analysis of multidimensional components, including interaction patterns, pedagogical structure, social context, and technological infrastructure, appears to be limited [13-17]. This suggests that a large portion of MOOCs research focuses on specific behavioral variables rather than a holistic view of learning processes, and therefore, data-based contributions to instructional design remain limited.

This study aims to identify the key factors affecting course completion and certification by analyzing learners' interaction behaviors within the scope of the MITx Elements of Structures (2.01x) course. Among the variables examined are learning behaviors, such as the number of days of interaction (ndays_act), the number of chapters completed (nchapters), the number of videos played (nplay_video), and writing activity (writing_act), which are indicators of interaction. Additionally, demographic variables such as age, education level, and gender are also included. The primary objective of the study is to reveal the significant effects of trainees' demographic characteristics and interaction behaviors on course completion and certification outcomes in online learning environments. The study's unique contribution is that it examines trainees' interaction not only in a single dimension, but also in terms of behavioral (number of active days, content engagement, and writing activity) and demographic (age and education level) dimensions. While previous research has primarily focused on specific types of interaction, this study employs logistic regression analysis to explain course success by incorporating multiple interaction variables into a large-scale MOOC dataset. In this respect, the study aims to provide data-driven, strategic recommendations for MOOC platforms to design personalized learning paths tailored to age groups and learning habits.

In this context, the research sought to answer the following questions:

1. What factors influence course completion?
2. What factors influence certification in courses?

The following sections explain the research method, dataset, analysis processes, and findings. The results obtained were discussed within the scope of the literature, and the results were interpreted.

II. Materials and Method

This study used data from an online course offered by the Massachusetts Institute of Technology (MIT) to determine the variables influencing trainees' course completion and certification. Logistic regression analysis was used, a method suitable for situations where the dependent variable is binary. Logistic regression is a powerful statistical technique used to analyze the effects of independent variables (continuous, categorical, or both) on a dependent variable and to estimate the probability that the dependent variable belongs to a particular category [18, 19].

The analysis model developed within the scope of the research consists of three main stages: data preprocessing, modeling, and evaluation (Figure 1). The raw dataset was first prepared in the required format for analysis and then processed using algorithms. To overcome processing capacity limitations during the modeling process, we utilized cloud-based resources provided by Google; all studies were conducted in the Google Colaboratory environment.

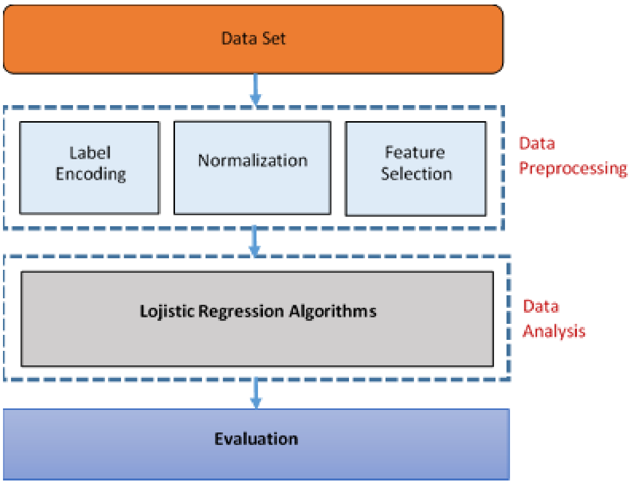

Figure 1: Study Design

Commonly used Python libraries such as Pandas, Scikit-learn, MLib, and PyCharm were used for data processing and analysis. Apache Spark, known for its effectiveness in big data analysis, was also used as the workspace. The analyses statistically revealed the determining factors affecting trainees' course completion and certification.

II.I. Data Set

The dataset used in this study (the EOS dataset) belongs to the Massive Open Online Course (MOOC) titled MITx/2.01x: Elements of Structures, offered by MIT via the edX platform in the spring of 2013. The course is an online adaptation of a face-to-face course offered by MIT's Department of Mechanical Engineering. It investigates the mechanical properties of flexible (deformable) structural elements.

The course content includes a comprehensive assessment system consisting of 81 video recordings, 308 problem activities, 138 HTML content pages, various supplementary learning materials, 9 assignments (36% weighting), and 2 quizzes (64% weighting). Participants must achieve a final grade of at least 60 to receive a certificate of achievement at the end of the course. This structure aims to measure knowledge acquisition and assess students' active participation in the learning process.

The dataset contains learning analytics data from a total of 12,646 participants. These data include multidimensional interactions such as video viewing times, forum posts, homework and exam performances, engagement with interactive materials, click sequences, and MATLAB-based problem-solving activities [20]. In this respect, the dataset allows for a holistic examination of cognitive and behavioral learning processes.

One of the most striking features of the dataset is that learning interactions are recorded synchronously throughout the learning process. This allows for detailed monitoring of the learning process and in-depth analysis using advanced analytical techniques [21, 22]. In addition, the fact that the individuals participating in the course have different socio-demographic and cultural backgrounds strengthens the study's external validity. It allows comparative analyses between student profiles [23].

Datasets like this dataset have been used in many EDM studies before and are frequently referenced in the literature [24-28]. The comprehensive and multilayered nature of the dataset provides a strong foundation for analyses aimed at understanding learner behavior in the MOOC context. Therefore, the selection of this dataset for this study serves the purpose of the research and is consistent with methodological approaches in the existing literature.

II.II. Variables Used in the Research

The variables used in the study were determined based on the individual characteristics of the trainees and their interaction behaviors in the online learning environment. To this end, measurable metrics that could quantitatively represent the learning process were developed, and their definitions are detailed in Table 1. The variables used in the analyses were determined based on the study conducted during the data preprocessing phase, and the variables used in the analyses are highlighted in bold in the table.

To address the research questions, a course completion status variable was derived to determine the factors influencing trainees' course completion status, and a certification variable was derived to determine the factors influencing certification status. To determine course completion status, a completion variable was derived based on final grade; participants with a final grade of zero were classified as "course dropouts" (completion = 0), while others were classified as "completers" (completion = 1). Information regarding whether participants had received a certification was represented by a binary variable called "certified," which was coded with values of 0 (non-certified) and 1 (certified).

Table 1: Variables and Their Descriptions

Nu	Variable	Explanation
1	viewed	Those who accessed the "Courseware" tab within the course
2	explored	Researched: Those who accessed at least half of the course sections
3	certified	Certificate earners
4	continent	Continent/region where trainees are located
5	completed	Course completion status (if grade = 0, completed = 0, else completed = 1)
6	Country Economic group	Developed/developing countries
7	education_ level	Highest level of education completed, 1: High school (el, jhs, hs, o), 2: Associate's degree/Bachelor's degree (a, b), 3: Postgraduate degree (m, p, p_oth, p_se)
8	YoB	Year of birth
9	gender	0: Female (f), 1: Male (m)
10	age	Ages between 16 and 88
11	age_group	1: Age groups 16-30, 2: Age groups 31-44, 3: Age groups 45 and above
12	grade	The final course grade ranges 0 to 1
13	Interaction intensity	Interaction intensity (nevents/ndays_act)
14	nevents	Number of interactions with the course recorded in tracking logs
15	ndays_act	Number of unique days the student interacted with the course
16	nplay_video	Number of video play events
17	writing_act	Writing activity (nforum_posts + nforum_comments)
18	reaction_act	Reacting activity (nforum_votes + nforum_endorsed)
19	nchapters	Number of sections the student interacted with
20	nforum_posts	Number of messages posted to the discussion forum
21	roles	Roles define staff and instructors
22	course_year	Course year
23	Course_level	1: Introductory, 2: Intermediate, 3: Advanced
24	Course_Name	Name of course
25	Course_Category	Course_Category,1: Science, Technology, Engineering, and Mathematics (STEM), 2: Computer Science (CS), 3: Humanities, History, Religion, Design, and Education (HHRDE), 4: Government, and Health and Social Sciences (GHSS)
26	geoip_country	Geographical region country

Note: Variables in bold are those selected during the feature selection process and used in the analyses.

II.III. Data Preprocessing Process

The variables used in this study were determined based on the participants' characteristics and the interaction-based behaviors they exhibited in the online learning environment. Accordingly, measurable metrics were developed to quantify the learning process, and the detailed definitions of the relevant variables are presented in Table 1. Following the preliminary analysis and data preprocessing performed on the obtained multidimensional data, the variables used in the study were determined and marked in bold in the table.

In line with the study's primary objective, the "completion" variable was derived to examine the factors influencing the course completion status of trainees, and the "certification" variable was derived to discuss their certification status. The course completion status variable was created based on participants' final grades. Accordingly, participants with a final grade of zero were classified as "course incomplete" (completion = 0), and participants with a final grade greater than zero were classified as "course complete" (completion = 1).

Certification status was represented by a binary variable called "certified," created based on records obtained from the platform. This variable was coded to indicate whether participants received a certificate of achievement at the end of the course; a value of "0" was assigned to those who did not receive a certificate, and a value of "1" was assigned to those who did. These binary variables were treated as dependent variables in the logistic regression analysis, and the effects of the relevant independent variables on these variables were statistically tested.

Personally identifiable information (user_id, username, ip, etc.) that did not contribute to the analysis, as well as variables that were missing or lacked meaningful information, were removed from the analysis. Additionally, to maintain data integrity and prevent bias, rows containing one or more null values or columns containing inconsistent data (e.g., roles, verified_enroll_time, verified_unenroll_time) were removed. This was achieved by applying the drop function, which removes rows containing null values. Remaining duplicate rows were removed using the drop_duplicates function to ensure the uniqueness of each record in the dataset. Before proceeding to the modeling phase of data mining processes, the algorithms must make the variables in the dataset processable. In this context, the One-Hot Coding (OHE) technique, one of the label encoding methods, was used to convert categorical variables into a numerical format. This method creates a separate binary column for each class of each categorical variable, allowing the algorithms to interpret these variables directly. Then, normalization was applied to the numerical variables, reducing the disproportionate effects of variables with different scales on the model. This process is a critical preliminary step for improving the model's overall performance, especially in distance-based and scale-sensitive algorithms. Normalization prevents variables with significant value ranges from dominating the model, ensuring that all variables contribute equally. Feature selection was performed as part of the data preprocessing process to reduce model complexity, prevent overfitting, and reduce computational costs. By ensuring that only explanatory variables with statistically significant relationships with the target variable are included in the model, feature selection increases model accuracy and speeds up the computational process [29, 30]. In this study, the "Information Gain" method was used in the feature selection process. This method ranks the variables by measuring the explanatory power of each independent variable on the target variable, allowing the most informative variables to be selected. The variables and their corresponding values in the model based on the information gain scores are presented in Table 2.

Furthermore, several new variables were derived to expand the scope of the analyses and examine learner behavior in more detail. For example, to determine whether trainees completed the course, the "completed" variable was created based on the final grade; this variable was defined as 0 for those with a final grade of zero and 1 for those with a grade greater than zero. Similarly, the "certified" variable, which indicates whether participants received a certificate, was coded in a binary format as 0 (not received) and 1 (received). Additionally, composite variables reflecting levels of interaction on the forum, such as "writing activity" (nforum_posts + nforum_comments) and "reaction activity" (nforum_votes + nforum_endorsed), were calculated and integrated into the model. Such derived variables facilitate a multidimensional learning process analysis, contributing to more comprehensive modeling outcomes.

Table 2: Importance Ranking of Variables According to Information Gain

Nu	Attribute	Score
1	grade	1,233726
2	completed	0,520217
3	nevents	0,432478
4	ndays_act	0,348848
5	interaction_intensity	0,299383
6	nplay_video	0,257852
7	certified	0,253111
8	completed	0,243632
9	writing_act	0,077868
10	reaction_act	0,035382
11	continent	0,010995
12	country_economic_group	0,004065
13	education_level	0,000000
14	age_group	0,000000
15	gender	0,000000

II.IV. Data Analysis

The primary objective of this study is to identify the factors influencing course completion and certification among MOOC participants. For this purpose, logistic regression analysis was frequently used when the dependent variable has a binary categorical structure. Logistic regression allows for the relationship between one or more independent variables and a binary target variable and is considered a statistically robust method, particularly for predicting categorical outcomes [31, 32]. In this context, these analyses considered course completion and certification status dependent variables. In contrast, the independent variables were the participants' characteristics and interaction behaviors in the online learning environment.

Classification performance metrics such as the ROC curve (Receiver Operating Characteristic Curve) and AUC (Area Under Curve) were used to evaluate the performance of the resulting models. These metrics measure the model's ability to discriminate between classes and objectively assess its accuracy [33].

III.RESULTS

This section includes the findings of logistic regression analysis within the scope of the research questions.

1. Factors affecting the course completion status of trainees

As part of this study's first research question, a logistic regression analysis was conducted to identify the factors influencing participants' online course completion. In this analysis, course completion was the dependent variable, while the independent variables were participants' demographic characteristics and interaction behaviors in the online learning environment. Descriptive statistics for the variables included in the analysis are presented by type. Information on categorical variables is presented in Table 3, and information on continuous variables is presented in Table 4.

Table 3: Descriptive Statistics of Categorical Variables

Variable	Groups	f	%
gender	0: female	1.627	13,0
	1: male	8.828	70,4
	Not specified	2.085	16,6
	Total	12.540	100
Age group	1:16-30	7.793	62,1
	2: 31-44	1.904	15,2
	3: 45 and above	693	5,5
	Not specified	2.150	17,1
	Total	12.540	100
Education status	1: High school (el, jhs, hs, o),	3.612	28,8
	2: Associate's degree/Bachelor's degree (a, b),	4.025	32,1
	3: Postgraduate degree (m, p, p_oth, p_se)	2.691	21,5
	Not specified	2.212	17,6
	Total	12.540	100

Table 4 Descriptive Statistics of Continuous Variables

Variable	count	Mean	Std	Min	Max
completed	12.474	0,21	0,41	,0	1,0
ndays_act	12.094	9,66	17,77	1,0	312,0
nplay_video	12.094	78,45	239,60	,0	6.180,0
nchapters	8.679	3,92	3,76	1,0	15,0
writing_act	12.540	,83	14,86	,0	1.473,0

To determine the variables affecting trainees' online course completion, the following variables were included in the logistic regression analysis: gender, educational status (LoE_new2), age (Age), number of days of interaction (ndays_act), number of video plays (nplay_video), number of chapters (nchapters), and writing activity (writing_act). The analysis results are presented in Table 5. The logistic regression analysis indicated that the overall fit of the model was statistically significant ($\chi2(9) = 3803.27$, $p< .05$), with a Nagelkerke R^2 value of 0.435, demonstrating that the model explained approximately 43% of the variance in the dependent variable and exhibited a high level of goodness of fit. In addition, the model correctly classified 85.5% of the observations, as shown in Figure 3.

Table 5: Logistic Regression Analysis Results of Factors Affecting Course Completion Status

Variable	B	SE B	Wald	df	sig	Odds Ratio
(Constant)	-2,799	,117	571,022	1	,000	,061
gender	-,169	,103	,103	1	,101	,844
Associate Degree	,032	,087	,087	1	,713	1,033
Postgraduate	,348	,099	,099	1	,000*	1,416
Age2	-,375	,103	,103	1	,000*	,688
Age3	-,605	,166	,166	1	,000*	,546
Number of days (ndays_act)	,088	,007	,007	1	,000*	1,092
Number of video plays (nplay_video)	,005	,000	,000	1	,000*	1,005
Number of chapters (nchapters)	,175	,018	,018	1	,000*	1,192
Writing activity (writing_act)	,210	,039	,039	1	,000*	1,233

*p< .05

According to the results in Table 5, significant relationships were observed between course completion status and several participant characteristics and interaction behaviors. Specifically, individuals with a postgraduate degree (Wald $\chi^2(1) = 0.099$, $p < .05$), those aged 31–44 (Wald $\chi^2(1) = 0.10$, $p < .05$), and participants aged 45 and older (Wald $\chi^2(1) = 0.17$, $p < .05$) showed a significant difference in the likelihood of course completion. Additionally, variables related to learning behaviors such as the number of days interacted (Wald $\chi^2(1) = 0.07$, $p < .05$), the number of video plays (Wald $\chi^2(1) = 0.00$, $p < .05$), the number of sections studied (Wald $\chi^2(1) = 0.018$, $p < .05$), and writing effectiveness (Wald $\chi^2(1) = 0.039$, $p < .05$) were also found to have significant effects on course completion. In contrast, gender and associate/bachelor's degree level education variables did not significantly contribute to course completion.

The model's classification performance was examined using various evaluation metrics in Table 6 and the ROC curve in Figure 2. The AUC (Area Under the Curve), which measures the model's ability to distinguish between two classes, was calculated as 0.80, demonstrating the model has a strong discriminatory capacity. Additionally, the accuracy rate calculated from the confusion matrix in Figure 3 is 0.86, indicating that the model correctly predicted 85.5% of the observations.

Table 6: Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-score	Total
0	0,85	0,86	0,94	0,90	1.514
1	0,85	0,83	0,65	0,73	642

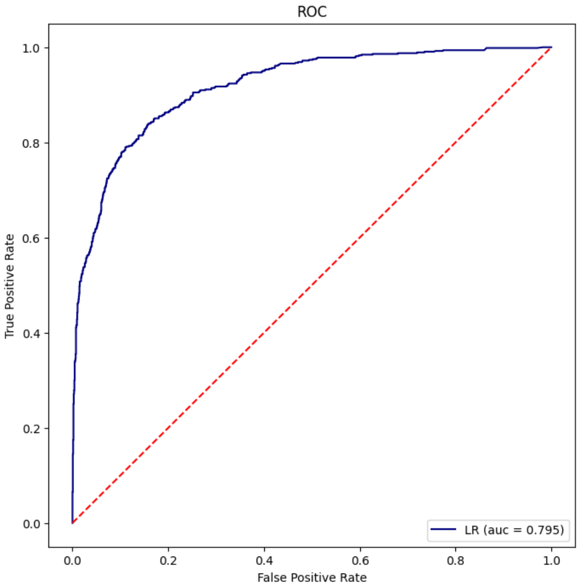

Figure 2. ROC curve

Consequently, standard classification metrics such as accuracy, precision, recall, F1 score, and Kappa statistic were utilized to evaluate model performance in binary logistic regression analysis. These metrics reflect different aspects of the model and contribute significantly to predicting learner behavior in the context of MOOCs (Massive Open Online Courses). As seen in the confusion matrix in Figure 3, the model achieved an accuracy value of 0.86, correctly classifying 85.5% of the cases.

2. Factors affecting certification status in courses

A binary (logistic) regression analysis was conducted to determine the factors influencing the certification status of MOOC participants. In this analysis, certification status was considered the dependent variable. In contrast, gender, education level (LoE_new2), age (Age), number of days of interaction (ndays_act), number of video plays (nplay_video), number of chapters reviewed (nchapters), and writing activity (writing_act) were considered as independent variables. Descriptive statistics for the relevant variables are presented in Tables 7 and 8.

Table 7: Descriptive Statistics of Categorical Variables

Variable	Groups	f	%
gender	0: female	1.627	13,0
	1: male	8.828	70,4
	Not specified	2.085	16,6
	Total	12.540	100
Age group	1:16-30	7.793	62,1
	2: 31-44	1.904	15,2
	3: 45 and above	693	5,5
	Not specified	2.150	17,1
	Total	12.540	100
Educational Status	1: High school (el, jhs, hs, o),	3.612	28,8
	2: Associate's degree/Bachelor's degree (a, b),	4.025	32,1
	3: Postgraduate degree (m, p, p_oth, p_se)	2.691	21,5
	Not specified	2.212	17,6
	Total	12.540	100

Table 8 Descriptive Statistics of Continuous Variables

Variable	Count	Mean	Std	Min	Max
certified	12.540	0,07	0,25	,0	1,0
ndays_act	12.094	9,66	17,77	1,0	312,0
nplay_video	12.094	78,45	239,60	,0	6.180,0
nchapters	8.679	3,92	3,76	1,0	15,0
writing_act	12.540	,83	14,86	,0	1.473,0

According to the analysis results, the model was found to be statistically significant overall ($\chi^2(9) = 3301.07$, $p < .05$) and explained 75.26% (R^2) of the variance in certification status. The model's classification accuracy is relatively high; as seen in Figure 4, 96.3% of the observations were correctly classified.

Table 9: Logistic Regression Analysis Results of Factors Affecting Certification Status

Variables	B	SE B	Wald	df	sig	Odds Ratio
(Constant)	-9,55	,499	367,3	1	,000	,000
gender	,144	,197	3,260	1	,572	1,154
Associate's degree/ Bachelor's degree	-,355	,213	,121	1	,071	,701
Postgraduate degree	,074	,212	1,034	1	,728	1,077
Age2	-,216	,299	9,610	1	,309	,806
Age3	-,927	,005	77,815	1	,002*	,396
ndays_act	,045	,000	5,563	1	,000*	1,046
nplay_video	,000	,044	263,649	1	,018*	1,000
nchapters	,713	,003	4,703	1	,000*	2,040
writing_act	-,005	,197	3,260	1	,030*	,995

*p< .05

According to the findings in Table 9, significant variables affecting certification status include participants aged 45 and over (Wald $\chi^2(1) = 77.82$, $p < .05$), number of days of interaction (Wald $\chi^2(1) = 5.56$, $p < .05$), number of video plays (Wald $\chi^2(1) = 263.65$, $p < .05$), number of sections reviewed (Wald $\chi^2(1) = 4.7$, $p < .05$), and writing activity (Wald $\chi^2(1) = 3.26$, $p < .05$). Specifically, the number of days of interaction and the number of sections completed were observed to have more significant effects on certification status compared to other variables. This suggests that students' active participation in online environments and regular content consumption increase their likelihood of certification [34-39].

However, gender and education level were insignificant in predicting the model's certification status. This result is consistent with previous studies [40, 41]. For example, Zhang et al. (2019) reported that gender and education level had no decisive effect on online learning success [42]. In this context, individual success in online learning environments is more closely related to interaction intensity and active participation in the learning process.

Regarding the age variable, individuals aged 45 and over were found to be more likely to receive a certification. This can be interpreted as the experience and motivation that come with age positively contributing to the learning process [42-45]. This result indicates the importance of online learning systems providing content specific to age groups.

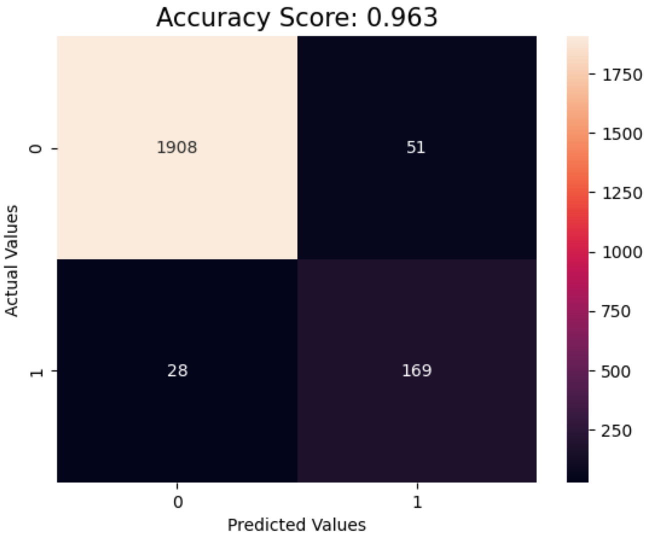

Figure 4. Confusion Matrix

The overall performance of the model was evaluated using the statistical metrics presented in Table 10 and the ROC curve presented in Figure 5. The area under the ROC curve (AUC) was calculated as 0.92, indicating that the model successfully distinguished between the two classes. AUC is a frequently used metric in evaluating the performance of classification models, and as the value approaches 1, the model's discriminatory power increases. Therefore, the high AUC value obtained in this study demonstrates that the logistic regression model can reliably predict certification status.

Table 10 Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-score	Total
0	0,96	0,99	0,97	0,98	1.959
1	0,96	0,77	0,86	0,81	197

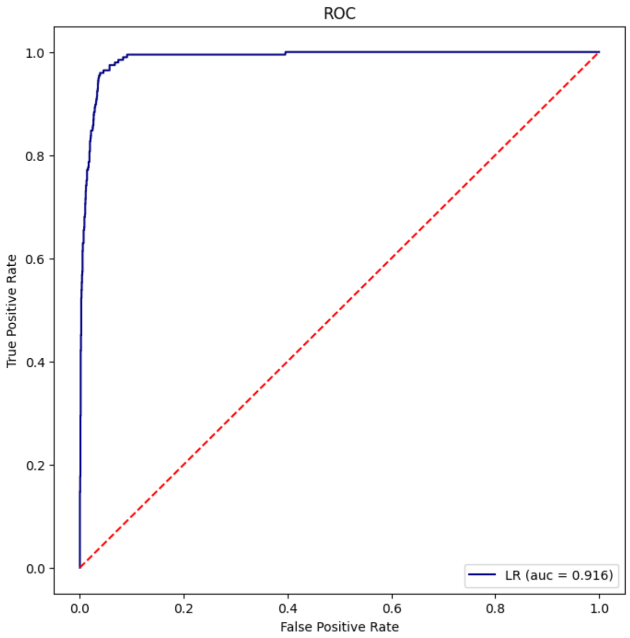

Figure 5. ROC curve

According to the research findings, there are significant relationships between certification status and the age group 45 and older, the number of days of interaction, video plays, sections reviewed, and writing activity. Specifically, the number of interaction days and sections completed positively impacts participants' likelihood of receiving a certification. A study by Chao et al. (2021) also highlighted that frequency of interaction and the number of days of activity positively impact certification and course completion [46]. However, some studies [47, 48] indicate that how a course is divided into sections also impacts success. The course in this study had 15 sections and was structured to be completed one per week; furthermore, the sections were kept as short as possible.

IV. DISCUSSION

1. Results of Factors Affecting Participants' Course Completion Status

This study examined variables affecting the course completion status of individuals participating in MOOCs through logistic regression analysis. The findings indicate that the model explained 43.5% of the variance in course completion status and correctly classified 85.5% of participants.

According to the results of the analysis, graduate education level, age groups 31-44 and 45 and above, number of days spent interacting, video playback frequency, level of section completion, and writing activity exhibit statistically significant effects on the likelihood of course completion. Of these variables, graduate education level and writing activity are potent predictors of course completion.

These findings are consistent with previous literature. Kizilcec et al. (2017) stated that interactive behaviors such as written content production and forum participation increase online course completion rates. Oswald et al. (2020) found that individuals with higher education levels demonstrate a higher commitment to online learning. However, it is also known that the literature on the effect of education level on course completion has yielded differing results [40, 45, 49-51].

2. Results of Factors Affecting Certificate Obtaining Status in Courses

Factors influencing MOOC participants' certification status were also evaluated using logistic regression analysis. The resulting model explained 75.26% of the variance related to certification status and correctly classified 96.3% of participants.

The analysis results indicate that for individuals aged 45 and over, the number of days spent interacting, videos watched, chapters completed, and writing activity significantly impact certification. Among these factors, interaction duration and the number of chapters completed substantially increase the likelihood of certification. However, gender and education level were not found to affect certification significantly.

These results are consistent with the findings of Chao et al. (2021) regarding the positive effects of interaction frequency and duration on learning success. Furthermore, studies conducted by Bingöl et al. (2020) and Çağıltay et al. (2020) have indicated that courses with a larger number of chapters may lower participant completion rates; therefore, dividing content into shorter, more manageable segments may increase learning efficiency [47, 48]. This study divided the course structure into 15-week modules, with one section completed each week.

Using models with high AUC (Area Under Curve) values increases the effectiveness of predicting the likelihood of obtaining a certification in MOOC environments. The area under the ROC curve of 0.92 demonstrates that the developed model can distinguish between highly accurate classes and has high predictive power. This finding highlights the value of logistic regression models as a strategic tool for personalizing users' learning experiences and delivering targeted content.

V. CONCLUSION

User interaction and duration of participation in online learning environments significantly impact individuals' success in completing courses and earning certifications. Age, in particular, stands out as a significant factor in this context; the higher success rates of individuals aged 45 and over demonstrate the contribution of experience and motivation to learning performance [42-45]. Accordingly, it is recommended that MOOC platforms offer differentiated, personalized content based on age groups.

Furthermore, applications that enhance written interaction and make forums more functional can increase participation and achievement levels. Well-designed discussion forums reinforce learners' sense of belonging and strengthen their motivation to learn by providing peer support comparable to that found in physical classrooms. However, considering that there are student profiles that actively participate in forums but do not complete the course, the nature of interaction patterns and their relationship to learning should be examined more thoroughly.

In conclusion, models that can accurately predict course completion and certification in MOOC environments contribute to improving learning processes and the development of personalized education strategies. Logistic regression analysis is a frequently used and powerful method in this regard. When supported by performance indicators such as AUC, it is considered an effective tool for both predicting user success and platform design. This study offers several strategic recommendations for MOOC designers and instructors to increase participant success, such as encouraging written interaction, structuring forums, and developing content appropriate for the target audience.

ACKNOWLEDGEMENTS

The data were used with permission from Prof. Dr. Kürşat ÇAĞILTAY of MIT. We are grateful to MITx for allowing us to obtain the MOOC data.

Authors' Contributions

The authors' contributions to the paper are equal.

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The authors declare that this study complies with Research and Publication Ethics.

REFERENCES

[1] P. Diver and I. Martinez, "MOOCs as a massive research laboratory: Opportunities and challenges.," *Distance Education,* vol. 36, no. 1, pp. 5-25, 2015.

[2] J. Goopio and C. Cheung, "The MOOC dropout phenomenon and retention strategies," *Journal of Teaching in Travel & Tourism,* vol. 21, no. 2, pp. 177-197, 2020, doi: 10.1080/15313220.2020.1809050.

[3] C. T. Swai and S. E. Mangowi, "Mining school teachers' MOOC training responses to infer their face-to-face teaching strategy preference," The International Journal of Information and Learning Technology, vol. 39, no. 1, pp. 82-94, 2022, doi: 10.1108/ijilt-07-2021-0102.

[4] S. Joksimović et al., "How Do We Model Learning at Scale? A Systematic Review of Research on MOOCs," Review of Education*al Research,* vol. 88, no. 1, pp. 43-86, 2017, doi: 10.3102/0034654317740335.

[5] B. C. Padilla Rodriguez, A. Armellini, and M. C. Rodriguez Nieto, "Learner engagement, retention and success: why size matters in massive open online courses (MOOCs)," *Open Learning: The Journal of Open, Distance and e-Learning,* vol. 35, no. 1, pp. 46-62, 2019, doi: 10.1080/02680513.2019.1665503.

[6] A. Heusler, D. Molitor, and M. Spann, "How Knowledge Stock Exchanges can increase student success in Massive Open Online Courses," *PLoS One,* vol. 14, no. 9, p. e0223064, 2019, doi: 10.1371/journal.pone.0223064.

[7] H. B. Shapiro, C. H. Lee, N. E. Wyman Roth, K. Li, M. Çetinkaya-Rundel, and D. A. Canelas, "Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers," *Computers & Education,* vol. 110, pp. 35-50, 2017, doi: 10.1016/j.compedu.2017.03.003.

[8] J. Swacha and K. Muszyńska, "Predicting dropout in programming MOOCs through demographic insights," *Electronics,* vol. 12, no. 22, p. 4674, 2023.

[9] Z. Liu, X. Kong, S. Liu, Z. Yang, and C. Zhang, "Looking at MOOC discussion data to uncover the relationship between discussion pacings, learners' cognitive presence and learning achievements," *Education and information technologies,* vol. 27, no. 6, pp. 8265-8288, 2022.

[10] J. He, J. Bailey, B. Rubinstein, and R. Zhang, "Identifying at-risk students in massive open online courses," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015, vol. 29, no. 1.

[11] A. Bozkurt *et al.*, "A global outlook to the interruption of education due to COVID-19 pandemic: Navigating in a time of uncertainty and crisis," *Asian journal of distance education,* vol. 15, no. 1, pp. 1-126, 2020.

[12] I. Despujol Zabala, L. Castaņeda, V. I. Marín, and C. Turro, "What do we want to know about MOOCs? Results from a machine learning approach to a systematic literature mapping review," 2022.

[13] G. Veletsianos and S. Houlden, "An analysis of flexible learning and flexibility over the last 40 years of Distance Education," *Distance Education,* vol. 40, no. 4, pp. 454-468, 2019.

[14] E. Anghel, J. Littenberg-Tobias, and M. von Davier, "What Did We Learn About Massive Open Online Courses for Teachers? A Scoping Review," *International Review of Research in Open and Distributed Learning,* vol. 26, no. 2, pp. 130-161, 2025.

[15] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success," *The Internet and Higher Education,* vol. 28, pp. 68-84, 2016.

[16] S. Joksimović *et al.*, "How do we model learning at scale? A systematic review of research on MOOCs," *Review of Educational Research,* vol. 88, no. 1, pp. 43-86, 2018.

[17] M. Zhu, A. R. Sari, and M. M. Lee, "Trends and issues in MOOC learning analytics empirical research: A systematic literature review (2011–2021)," *Education and Information Technologies,* vol. 27, no. 7, pp. 10135-10160, 2022.

[18] S. Menard, *Applied logistic regression analysis*. SAGE publications, 2001.

[19] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research,* vol. 96, no. 1, pp. 3-14, 2002.

[20] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard, "Who does what in a massive open online course?," *Communications of the ACM,* vol. 57, no. 4, pp. 58-65, 2014.

[21] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses," presented at the Proceedings of the third international conference on learning analytics and knowledge, 2013.

[22] M. C. Almodiel, "Assessing Online Learners' Access Patterns and Performance Using Data Mining Techniques," *International Journal in Information Technology in Governance Education and Business,* vol. 3, no. 1, pp. 46-56, 2021, doi: 10.32664/ijitgeb.v3i1.87.

[23] J. DeBoer, A. D. Ho, G. S. Stump, and L. Breslow, "Changing "course": Reconceptualizing educational variables for massive open online courses," *Educational Researcher,* vol. 43, no. 2, pp. 74-84, 2014.

[24] B. P. Cohen and M. Nycz, "Learning analytics in MOOCs: A literature review," *International Journal of Learning Analytics and Artificial Intelligence for Education,* vol. 1, no. 1, pp. 33-48, 2017.

[25] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller, "Understanding in-video dropouts and interaction peaks in online lecture videos," presented at the Proceedings of the first ACM conference on Learning@ scale conference, 2014.

[26] N. H. Ling, C. J. Chen, C. S. Teh, D. S. John, L.-C. Ch'ng, and Y. F. Lay, "Global Trends of Educational Data Mining in Online Learning," *International Journal of Technology in Education,* vol. 6, no. 4, pp. 656-680, 2023, doi: 10.46328/ijte.558.

[27] M. T. Sarıtaş, C. Börekci, and S. Demirel, "Quality Assurance in Distance Education Through Data Mining," *International Journal of Technology in Education and Science,* vol. 6, no. 3, pp. 443-457, 2022, doi: 10.46328/ijtes.396.

[28] A. Urban, "How Hands-on Assessments Can Boost Retention, Satisfaction, Skill Development, and Career Outcomes in

Online Courses," *Ai Computer Science and Robotics Technology,* vol. 2, 2023, doi: 10.5772/acrt.23.

[29] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Education and Information Technologies,* vol. 28, no. 1, pp. 905-971, 2023.

[30] C. Romero and S. Ventura, "Educational Data Mining: A Survey from 1995 to 2005," *Expert Systems with Applications,* pp. 135-146, 2017.

[31] Ş. Büyüköztürk, "Sosyal bilimler için veri analizi el kitabı," *Pegem Atıf İndeksi,* pp. 001-214, 2018.

[32] J. Pallant, *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. Routledge, 2020.

[33] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters,* vol. 27, no. 8, pp. 861-874, 2006.

[34] J. Littenberg-Tobias, J. A. Ruipérez-Valiente, and J. Reich, "Studying learner behavior in online courses with free-certificate coupons: Results from two case studies," *International Review of Research in Open and Distributed Learning,* vol. 21, no. 1, pp. 1-22, 2020.

[35] V. Sherimon, P. Sherimon, L. Francis, D. Devassy, and T. K. George, "Factors associated with Student enrollment, completion, and dropout of massive open online courses in the Sultanate of Oman," *International Journal of Learning, Teaching and Educational Research,* vol. 20, no. 11, pp. 154-169, 2021.

[36] L. Caprara and C. Caprara, "Effects of virtual learning environments: A scoping review of literature," *Education and Information Technologies,* vol. 27, no. 3, pp. 3683-3722, 2022.

[37] A. Eltayar, S. R. Aref, H. M. Khalifa, and A. S. Hammad, "Prediction of graduate learners’ academic achievement in an online learning environment using a blended trauma course," *Advances in Medical Education and Practice,* pp. 137-144, 2023.

[38] S. Gunes, "Design entrepreneurship in product design education," *Procedia-Social and Behavioral Sciences,* vol. 51, pp. 64-68, 2012.

[39] A. Wald, P. A. Muennig, K. A. O'Connell, and C. E. Garber, "Associations between healthy lifestyle behaviors and academic performance in US undergraduates: a secondary analysis of the American College Health Association's National College Health Assessment II," *American Journal of Health Promotion,* vol. 28, no. 5, pp. 298-305, 2014.

[40] A. S. Schulze, *Massive open online courses (MOOCs) and completion rates: are self-directed adult learners the most successful at MOOCs?* Pepperdine University, 2014.

[41] V. Sherimon, L. Francis, D. Devassy, and W. Aboraya, "Exploring the Impact of Learners’ Demographic Characteristics on Course Completion and Dropout in Massive Open Online Courses," *International Journal of Research - Granthaalayah,* vol. 10, no. 1, pp. 149-160, 2022, doi: 10.29121/granthaalayah.v10.i1.2022.4469.

[42] Q. Zhang, F. C. Bonafini, B. B. Lockee, K. W. Jablokow, and X. Hu, "Exploring demographics and students’ motivation as predictors of completion of a massive open online course," *International Review of Research in Open and Distributed Learning,* vol. 20, no. 2, 2019.

[43] L. R. Gorfinkel, A. Giesler, H. Dong, E. Wood, N. Fairbairn, and J. Klimas, "Development and evaluation of the online addiction medicine certificate: free novel program in a Canadian setting," *JMIR medical education,* vol. 5, no. 1, p. e12474, 2019.

[44] P. J. Guo, J. Kim, and R. Rubin, "How video production affects student engagement: An empirical study of MOOC videos," presented at the Proceedings of the First ACM Conference on Learning@ Scale Conference, 2014.

[45] N. P. Morris, B. Swinnerton, and S. Hotchkiss, "Can demographic information predict MOOC learner outcomes?," in *Experience track: proceedings of the European MOOC stakeholder*, 2015: Leeds.

[46] Y.-P. Chao *et al.*, "Using a 360 virtual reality or 2D video to learn history taking and physical examination skills for undergraduate medical students: pilot randomized controlled trial," *JMIR serious games,* vol. 9, no. 4, p. e13124, 2021.

[47] I. Bingol, E. Kursun, and H. Kayaduman, "Factors for success and course completion in massive open online courses through the lens of participant types," *Open Praxis,* vol. 12, no. 2, pp. 223-239, 2020.

[48] N. E. Cagiltay, K. Cagiltay, and B. Celik, "An analysis of course characteristics, learner characteristics, and certification rates in MITx MOOCs," *International Review of Research in Open and Distributed Learning,* vol. 21, no. 3, pp. 121-139, 2020.

[49] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton, "Studying learning in the worldwide classroom research into edX's first MOOC," *Research & Practice in Assessment,* vol. 8, pp. 13-25, 2013.

[50] A. D. Ho *et al.*, "Harvardx and mitx: The first year of open online courses—fall 2012–summer 2013 (harvardx and mitx working paper# 1)," *EducationXPress,* vol. 2014, no. 2, pp. 1-1, 2014.

[51] Q. Zhang *et al.*, "Exploring the communication preferences of MOOC learners and the value of preference-based groups: Is grouping enough?," *Educational Technology Research and Development,* vol. 64, pp. 809-837, 2016.