

Improving generalizability of predictive models through course-related variables

Pedro Manuel Moreno-Marcos^{1,*}, Pedro J. Muñoz-Merino¹ and Carlos Delgado Kloos¹

¹Department of Telematic Engineering, Universidad Carlos III de Madrid, 28911 Leganés (Madrid), Spain, ROR: 03ths8210

Abstract

Students' dropout and academic failure are two of the main challenges in educational contexts. Researchers have made significant efforts to develop predictive models to detect students at risk. However, one of the main limitations is that these models are trained with data from one course but they do not work well when they are used in a different course (sometimes even in another edition of the same course) due to the impact of the course context. In this direction, this work aims to analyze how generalizability of the models could be improved by using global models that contain data about many courses and whether or not it is possible to enhance the models by using course-related variables that could capture information about the context. In order to that, data from 16 Small Private Online Courses (SPOCs) are used to develop the models to predict dropout and students' success. Results show that while it is possible to achieve accurate predictions at global level when training using several courses, these models do not properly fit all individual courses. Particularly, there is a drop in Area Under the Curve (AUC) higher than 0.1 in 17-40% of the courses, depending on the variable to predict. Moreover, it is possible to enhance the predictive models (up to 0.08 in AUC) by adding course-related variables that capture the main features of the course context. Among these variables, the most relevant ones are the average length of videos, and the number of videos and exercises in the course. These results add new insights about the variables that should be used in the models to improve the generalizability, which is crucial for real implementations.

Keywords

Analytics, Generalizability, Retention, Blended learning, Higher Education

1. Introduction

Prediction is one of the main research areas in learning analytics because there is a high interest in detecting students who will drop out the course or fail [1]. In this direction, many works have been carried out in different contexts at both degree-level (i.e., predict who will drop out the degree [2, 3]) and course-level (i.e., predict who will drop out the course [4, 5]). However, one important limitation of these works is that the models are usually trained using data from one course and they may lose accuracy when using the models in other contexts [6].

In order to analyze this problem, several researchers have tried to train models using data from different populations and for example, Ocumpaugh et al. [6] experienced difficulties when generalizing affect detectors in different populations. In addition, other works have analyzed how models can be transferred to the subsequent editions of the course. In this line, and Veeramachaneini [7] reported a drop of 0.1 or more in AUC when using the dropout predictive models trained with data from previous edition of the course. Similarly, Moreno-Marcos et al. [8] analyzed generalizability using data from two SPOCs to prepare the university entrance exams in two cohorts. They found that models worked well when transferring them to the other SPOC with the same students, and they were still acceptable when transferring them to the same course in a subsequent edition. However, they faced difficulties when modifying both the students and the course.

LASI Spain'25: Learning Analytics Summer Institute, May 26–27, 2025, Vitoria, Spain

*Corresponding author.

✉ pemoreno@it.uc3m.es (P. M. Moreno-Marcos); pedmume@it.uc3m.es (P. J. Muñoz-Merino); cdk@it.uc3m.es (C. Delgado Kloos)

🌐 <https://www.it.uc3m.es/pemoreno/> (P. M. Moreno-Marcos); <http://www.it.uc3m.es/~pedmume/> (P. J. Muñoz-Merino); <http://www.it.uc3m.es/cdk/> (C. Delgado Kloos)

🆔 0000-0003-0835-1414 (P. M. Moreno-Marcos); 0000-0002-2552-4674 (P. J. Muñoz-Merino); 0000-0003-4093-3705 (C. Delgado Kloos)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Given the relevance of the context, other authors, such as Bote-Lorenzo and Gómez-Sánchez [9] proposed in-situ models, which are trained with previous data in the same course (e.g., use data up to chapter i-1 to analyze what happens in chapter i). In addition, some have tried to quantify the impact of the generalization and there have been even special issues with a focus on this problem [10]. Among the works in that line, some works have shown that specific predictors [11] or regularization techniques could improve generalizability [12], and that the sample could also have an impact so that generalization to future editions may be better when training with a small number of students at risk [13].

Nevertheless, further work is needed around this line as there is not consensus about the how to improve generalization and the best ways to achieve it [14], and this problem is still considered a key limitation of current learning analytics models [15]. In this direction, one possibility to improve generalization is by creating global models that capture the information of many courses and including course-related variables that add information about the course context beyond the students' information. With this idea, the objective of this paper is to analyze (1) the impact of global models that include data from many courses to capture the variability of contexts and (2) the impact of course-related variables to discover whether or not they could serve to mitigate the generalizability issue.

The remainder of the paper is as follows. Section 2 details the methodology of the paper, including a description of the data, the list of variables, and the analytical methods. Section 3 presents the results obtained in this paper in relation to the objectives. Finally, Section 4 provides the conclusions of this paper, as well as the main limitations and future research directions.

2. Methodology

The study was carried out using data from 16 SPOCs offered by a Spanish university and hosted in a local Open edX instance. These courses serve to support face-to-face courses (offered in a synchronous way) although there are three ways the SPOCs could be used: (1) SPOCs needed to pass the course, with a weight in the final grade; (2) SPOCs that are part of the course and they could be combined with flipped classroom, although they are not part of the summative evaluation; and (3) SPOCs that are only used as support materials. However, this information is unknown.

Considering these scenarios, the prediction goal is to forecast student success and dropout. Regarding success, a student is considered successful when the average grade in the SPOC activities is higher or equal than 5 out of 10 (non attempted activities count as 0). Regarding dropout, two possible definitions are considered: (1) dropout related to activity, which means that students dropout when they do not interact for two consecutive weeks (excluding weeks where less than 10% of students interact), and (2) dropout related to completion, which means that students dropout when they do not complete 75% of the activities at least.

In order to carry out these models, several variables are used. The full list of variables is presented in Table 1. For the implementation of the models, Random Forest is used (as from previous studies, e.g., [16, 17], this is one of the most successful methods for prediction in similar contexts), and models are evaluated using the Area Under the Curve (AUC), since this is a well-known metric and it is generally appropriate for classification problems involving students behaviors [18].

3. Results

A first analysis was conducted to evaluate the potential of models trained with many courses and how these models work in specific courses. Particularly, models were trained with 15 courses and evaluated with the remaining one using data from week 8 out of 16. Results of the 16 trained models with each set of 15 courses are presented in Table 2. When conducting this analysis, it was observed that the dropout rate or failure was 100% in some courses, probably because these SPOC were designed as a support material and the expected engagement was different. This occurs in one course considering the dropout definition related to activity, in five courses with the dropout definition related to completion,

Table 1

List of variables involved in the analyses

Variable	Description
Variables about interactions with videos	
perc_vopen	It indicates the percentage o videos the student has opened over the total
perc_vtotal	It indicates the percentage of visualized content in videos considering the total duration of all videos and without counting repetitions of the same segments of the video
perc_compl	It indicates the percentage of videos the student has fully watched over the total
avg_rep	It is the relationship between the total time invested watching videos and the total time of different content that the student has watched
avg_pauses	It indicates the average number of times the students pause the video considering the opened videos
Variables about interactions with exercises	
perc_attempt	It indicates the percentage o exercises the student has attempted
avg_attempt	It is the average number of attempts in the exercises the student has attempted
avg_attempted	It is the average grade of the student considering all the attempts in only the attempted exercises
avg_attempted_fr	It is the average grade of the student considering the first attempt in only the attempted exercises
perc_correct	It indicates the percentage of exercises the student has solved correctly (100%) over the exercises they have attempted
CFA	It indicates the percentage of exercises the student has solved correctly (100%) at first attempt over the exercises they have attempted
steak_ex	It indicates the longest streak of consecutive exercises with a score of 100% (correct exercises)
Variables about platform use	
steak_acc	It indicates the maximum number of consecutive days the student has accessed to the platform to do course activities
perc_days	It indicates the percentage of days the student has accessed to the platform
avg_con	Average number of consecutive days the student accesses to the platform. For example, if the student acceses 3 consecutive days, and later 5 consecutive days, the value is $(3+5)/2=4$
Variables related to the course characteristics	
area	It indicates the thematic area of the course, including (1) Humanities, (2) Social Sciences, (3) Natural Sciences, (4) Formal Sciences, and (5) Professions and Applied sciences
synchronous	It indicates whether the course is instructor-paced or self-paced (synchronous or asynchronous)
nexercises	It indicates the total number of exercises available in the course
nvideos	It indicates the total number of videos available in the course
avg_duration	It indicates the average number of seconds of the videos in the course
num_staff	It indicates the total number of staff (mainly instructors) involved in the course
english	It indicates whether the course is delivered in English (1) or Spanish (0)

and four courses considering success. Given that it is not possible to compute AUC for those cases as there is a single category, they appear with a dash.

Results show that despite obtaining an accurate AUC at global level in general, the AUC for each individual course may vary. When analyzing the training set of 15 courses, results are very similar in all cases, but when testing specific courses, higher differences are observed. For the first case of dropout related to inactivity, the global AUC obtained with the 15 courses with cross-validation was between 0.76-0.79, and there was a drop higher than 0.1 in AUC in 40% of the individual courses (6 courses).

Curso	Dropout (activity)		Dropout (course completion)		Success	
	AUC_GLB	AUC_CUR	AUC_GLB	AUC_CUR	AUC_GLB	AUC_CUR
Course 1	0.78	-	0.92	1.00	0.94	1.00
Course 2	0.76	0.74	0.94	0.71	0.95	0.75
Course 3	0.78	0.63	0.92	-	0.93	-
Course 4	0.78	1.00	0.92	0.86	0.93	1.00
Course 5	0.77	0.71	0.92	-	0.93	-
Course 6	0.77	0.82	0.93	0.93	0.93	0.98
Course 7	0.78	0.62	0.92	-	0.93	-
Course 8	0.77	1.00	0.93	-	0.93	-
Course 9	0.78	0.68	0.92	0.90	0.93	0.93
Course 10	0.79	0.68	0.92	0.95	0.93	0.96
Course 11	0.78	0.56	0.92	0.90	0.93	0.93
Course 12	0.78	0.70	0.93	1.00	0.94	1.00
Course 13	0.78	0.57	0.93	0.87	0.94	0.92
Course 14	0.75	0.92	0.92	0.95	0.93	0.99
Course 15	0.76	0.40	0.92	-	0.94	0.54
Course 16	0.79	0.64	0.94	0.80	0.95	0.85

AUC_GLB: AUC of global model. AUC_CUR: AUC of current course

Table 2

Results of the predictive models in specific courses

Moreover, there are 3 courses (20%) where there is an increase in AUC higher than 0.1.

When analyzing the dropout with the alternative definition, the global AUC was between 0.92-0.94 and a drop in AUC higher than 0.1 was observed in only 18% of the courses (two courses). Among those two courses, the AUC was below 0.8 in one course, which suggests that global models offer a high performance in most of the cases. Similarly, a global AUC between 0.93-0.95 was observed when predicting students' success and only 17% of the courses (two courses) experienced a drop in AUC higher than 0.1. Thus, the global models trained with several courses may be useful in most of the cases and this could be a valid approach when having a large set of courses. Nevertheless, there might be courses where this approach does not work, and more specific models should be developed, which is consistent with the literature, which suggests that one-size-fits-all solutions are not possible in learning analytics models [19].

A second analysis was conducted to analyze whether or not course-related variables could improve global models. For this analysis, global models were computed with and without course-related variables throughout the course. Thus, models without course-related variables include the other three categories in Table 1 (videos, exercises and platform use) related to the students' interactions. In addition, a model with just course-related variables was implemented to analyze the predictive power of these variables by themselves (course model). Results of these models are presented in Figure 1.

From this figure, it can be seen that there was an average improvement with course-related variables between 0.03-0.04 in AUC for all the three dependent variables, and this value went up to 0.07-0.08 at the beginning of the course, when the difference was higher. This may entail that course-related variables contain meaningful information that could improve predictive models. In addition, it is observed that the AUC of only course-related variables is fair even when there is not information about the students, which reinforces the importance of the course context.

In order to delve into the variables, the importance of variables was computed using the Mean Decrease of Gini [20]. For the case of dropout related to activity, the variable that stood out was the percentage of attempted exercises, followed by the average number of repetitions per video and the percentage of days that the student accesses to the platform. Regarding the course-related variables, they were not the most important ones although their relative importance was similar in several variables. Particularly, the most important ones were the average duration of videos and the number of exercises. For the case of dropout related to completion, the most relevant variables were the percentage of

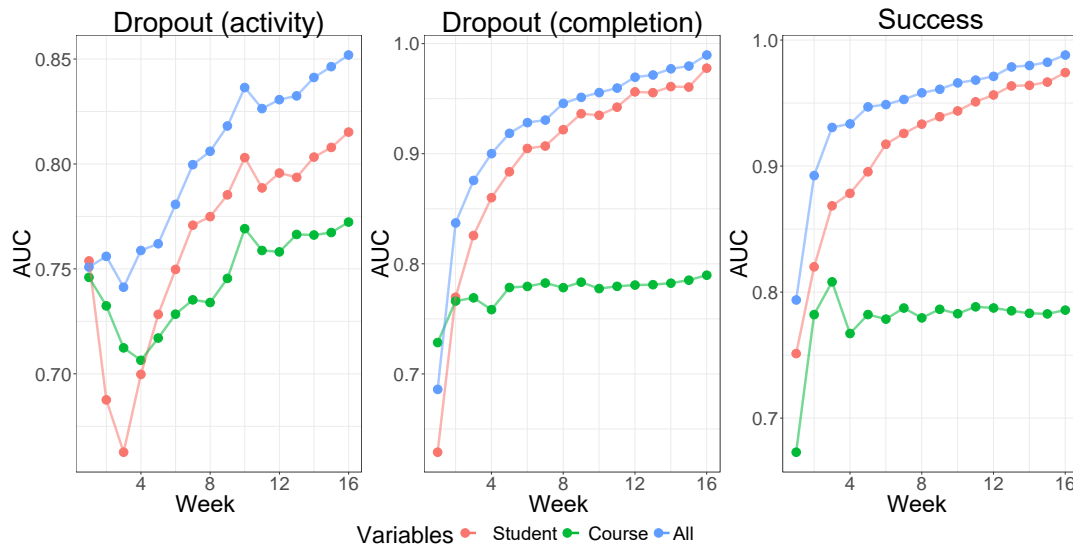


Figure 1: Distribution of the number of days students deliver the exams before the deadlines

attempted exercises, followed by the percentage of correct exercises at first attempt and the percentage of visualized videos. It is noteworthy that the fourth most significant variables was the average length of videos, which is a course-related variables. Regarding the other course related variables, their relative importance was smaller although the number of exercises had higher values of importance. As for success prediction, the percentage of attempted exercises was the most relevant variable, which is reasonable as engaging with exercises is crucial to complete the course and be successful. After this variable, the most relevant one is the percentage of correct questions at first attempt and the average duration of videos. The latter variable reflects how relevant the course design might be for student success and the fact that an inappropriate duration may lead to dropout or failure. In this case, the number of exercises and videos appear in positions four and six, which also highlight how crucial a good course design is. In summary, variables related to students' interactions such the percentage of attempted exercises are the most relevant ones, although the course-related variables also have a significant impact in the models.

4. Conclusions

This study has analyzed the impact of global models to improve generalization and the use of course-related variables. Results have shown the potential of the models trained with many courses as they generalized in most of the cases. Particularly, for the dropout related to activity, there was a drop in AUC higher than 0.1 in 40% of the courses, when analyzing dropout related to completion, that drop occurred in 18% of the courses, and when analyzing success, the drop occurred in 17% of the courses. For the latter two cases, the global AUC was above 0.9, which means that global models work very well in more than 80% of the courses. For the case of dropout related to inactivity, results were worse although global models were not accurate enough. Further research should be done in that case as inactivity might be high dependent on the context (the expected activity of the MOOC may vary depending on the methodology).

When analyzing the impact of course-related variables, results showed that they can improve the predictive models, with an average improvement of 0.03-0.04 in AUC and up to 0.07-0.08 at the beginning of the course. Moreover, the predictive power of these variables by themselves was not very strong, but managed to achieve fair predictive results. Finally, the analysis of variable importance showed that the percentage of attempted exercises was the variable that stood out, and the average duration of videos and the number of exercises and videos were the most relevant course-related variables.

Despite the aforementioned findings, there are some limitations that are worth mentioning. One key

limitation was the lack of information of the course methodology. This had a significant effect on the dropout related to inactivity as the expected activity was unknown. In addition, these analyses were done based only on the SPOC, but the summative assessments of the course and final grades are not covered. Thus, students who are considered as dropouts may not engage with the SPOC but still pass the course. Furthermore, the way dropout is defined may also have an effect on the results, and the sample could also be influential. While several courses are analyzed, more courses would be needed to improve the generalizability of the findings.

As future work, it would be interesting to gather more information about the course methodology to create a new category of variables related to the methodology and analyze whether or not they could even improve the performance of global models. Moreover, it would be relevant to further analyze the generalizability of the predictive models with more contexts and analyze the specific reasons why models do not fit well in some cases. Finally, it would be relevant to put these models into practice in active courses and analyze the possible interventions and their impact to improve academic success and reduce dropout.

Acknowledgments

This work was supported by Universidad Carlos III de Madrid (UC3M) through the Grants for the Research Activity of Young Doctors of the UC3M's Own Research and Transfer Program (ASESOR-IA project). Moreover, it was supported by FEDER / Ministerio de Ciencia, Innovación y Universidades - Agencia Estatal de Investigación through the grant PID2023-146692OB-C31 (GENIE Learn project) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/UE, by the UNESCO Chair of "Scalable Digital Education for All" at UC3M and by the grant RED2022-134284-T funded by MICIU/AEI/10.13039/501100011033.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

- [1] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, C. Delgado Kloos, Prediction in MOOCs: A review and future research directions, *IEEE transactions on Learning Technologies* 12 (2018) 384–401.
- [2] C. Olivares Rodríguez, P. M. Moreno Marcos, E. Scheihing García, P. J. Muñoz Merino, C. Delgado Kloos, An actionable learning path-based model to predict and describe academic dropout (2024).
- [3] A. A. Jiménez Macías, P. M. Moreno Marcos, P. J. Muñoz Merino, M. Ortiz Rojas, C. Delgado Kloos, Analyzing feature importance for a predictive undergraduate student dropout model (2023).
- [4] J. Chen, B. Fang, H. Zhang, X. Xue, A systematic review for MOOC dropout prediction from the perspective of machine learning, *Interactive Learning Environments* 32 (2024) 1642–1655.
- [5] K. Niu, J. Cai, Y. Zhou, W. Tai, X. Feng, Hybrid neural network model for MOOC dropout prediction, *Complex System Modeling and Simulation* (2025).
- [6] J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan, C. Heffernan, Population validity for educational data mining models: A case study in affect detection, *British Journal of Educational Technology* 45 (2014) 487–501.
- [7] S. Boyer, K. Veeramachaneni, Transfer learning for predictive models in massive open online courses, in: *International conference on artificial intelligence in education*, Springer, 2015, pp. 54–63.

- [8] P. M. Moreno-Marcos, T. De Laet, P. J. Muñoz-Merino, C. Van Soom, T. Broos, K. Verbert, C. Delgado Kloos, Generalizing predictive models of admission test success based on online interactions, *Sustainability* 11 (2019) 4940.
- [9] M. L. Bote Lorenzo, E. Gómez Sánchez, et al., An approach to build in situ models for the prediction of the decrease of academic engagement indicators in massive open online courses (2018).
- [10] C. Romero, S. Ventura, Guest editorial: Special issue on early prediction and supporting of learning performance, *IEEE Transactions on Learning Technologies* 12 (2019) 145–147.
- [11] J.-L. Hung, B. E. Shelton, J. Yang, X. Du, Improving predictive modeling for at-risk student identification: A multistage approach, *IEEE Transactions on Learning Technologies* 12 (2019) 148–157.
- [12] D. M. Olive, D. Q. Huynh, M. Reynolds, M. Dougiamas, D. Wiese, A quest for a one-size-fits-all neural network: early prediction of students at risk in online courses, *IEEE Transactions on Learning Technologies* 12 (2019) 171–183.
- [13] N. Gitinabard, Y. Xu, S. Heckman, T. Barnes, C. F. Lynch, How widely can prediction models be generalized? performance prediction in blended courses, *IEEE Transactions on Learning Technologies* 12 (2019) 184–197.
- [14] D. Zhidkikh, V. Heilala, C. Van Petegem, P. Dawyndt, M. Jarvinen, S. Viitanen, B. De Wever, B. Mesuere, V. Lappalainen, L. Kettunen, et al., Reproducing predictive learning analytics in cs1: Toward generalizable and explainable models for enhancing student retention., *Journal of Learning Analytics* 11 (2024) 132–150.
- [15] N. Sghir, A. Adadi, M. Lahmer, Recent advances in predictive learning analytics: A decade systematic review (2012–2022), *Education and information technologies* 28 (2023) 8299–8333.
- [16] P. M. Moreno-Marcos, P. J. Muñoz-Merino, C. Alario-Hoyos, C. Delgado Kloos, Re-defining, analyzing and predicting persistence using student events in online learning, *Applied Sciences* 10 (2020) 1722.
- [17] A. Jiménez-Macías, P. J. Muñoz-Merino, P. M. Moreno-Marcos, C. Delgado Kloos, Evaluation of traditional machine learning algorithms for featuring educational exercises, *Applied Intelligence* 55 (2025) 1–25.
- [18] R. Pelánek, Metrics for evaluation of student models., *Journal of Educational Data Mining* 7 (2015) 1–19.
- [19] D. Gašević, S. Dawson, T. Rogers, 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* 28 (2016) 68–84.
- [20] G. Louppe, L. Wehenkel, A. Suter, P. Geurts, Understanding variable importances in forests of randomized trees, *Advances in neural information processing systems* 26 (2013).