IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol.004, LIR226
DOI: https://doi.org/10.52731/lir.v004.226

Predicting Student Dropout Risk Using LMS Logs

Takaaki Ohkawauchi *, Eriko Tanaka *

Abstract

Traditionally, the prediction of student dropout in university classes has often been based on students' pre-enrollment information or confirmed grade data for each semester after enrollment. However, effective support requires early intervention when signs of dropping out appear. In this study, we propose a model to continuously measure dropout signs using log data accumulated in a learning management system during classes. By applying machine learning to the log data in the learning management system, we could continuously update information on at-risk students with high accuracy from the beginning to the end of the class.

Keywords: learning management system, log data, machine learning, prediction of at-risk student

1 Introduction

Student dropout is a major problem in universities in many countries. Dropout is not only a waste of time and money for students but also undermines universities' ability to secure tuition fees, which are a major source of funding. In addition, Japanese universities are now required to disclose their dropout rates, and a high dropout rate lowers perceived educational quality and hinders student recruitment. According to dropout rate data reported irregularly by the Ministry of Education, Culture, Sports, Science, and Technology, approximately 7% of students at Japanese universities fail to graduate and drop out at some point during their four years of study [1]. Some reasons for withdrawal are positive, such as the desire to attend other universities or find a job. However, most are negative, such as poor financial circumstances, decreased motivation to study, and a lack of course credits.

As indicated above, student attrition can be caused by academic, environmental, economic, or a combination of these factors. However, dropout does not occur suddenly; there will typically be signs and processes leading to it. For example, students who attend classes and perform well rarely suddenly drop out. Typically, students' class attendance rates and test scores decline between when they begin to consider leaving university and when they actually leave. Early detection of such signs would allow university staff to more effectively support students. This study aimed to identify students likely to drop out using Learning Management System (LMS) logs to continuously monitor their learning status and grades.

2 Related Studies

In the U.S., where university dropout rates are high, the factors behind the dropout rates have long been studied. Tinto's "Student Integration Model" is the foundation of dropout research to date. This model is based on the idea that academic and social integration interact and influence

* Nihon University, Tokyo, Japan

each other, leading to dropout [2][3]. According to this theory, not all students begin schooling on equal footing at the time of enrollment; they have their own motivations for enrolling and attitudes toward graduation. When students' motivation to enroll and graduate is reinforced by academics and other social circumstances, dropout risk decreases; without sufficient reinforcement, the risk increases.

Astin [4] proposed the "I-E-O model," which explains student learning in terms of "input," "environment," and "outcome." Robbins [5] categorized the factors related to students' dropout as "academic factors," which focus on grades before enrollment, "non -academic factors," which focus on the learning environment and student life after enrollment, and "other factors," such as parents' educational background and income, and examined how each factor affects the dropout rate. However, his definition of non-academic factors includes data on learning, such as learning attitudes and learning time, corresponding to Astin's concept of the environment. Therefore, the model can be used to examine whether the outcome results in university dropout from the input and environment. For example, one study showed that parental income affects dropout rates. A related study found that non-academic factors such as study habits, motivation, and goals were more important in predicting dropout rate than grades at the time of enrollment [6]. One study found that outcome is affected not only by students' environment and motivation but also subsequent non-academic factors, including motivation [7].

Even at the Open University in the UK, which is known for its advanced online education, as many as 35% of learners dropped out before the first assignment, and there have been reported cases in which nearly 60% of all students eventually dropped out [8]. In Europe and the U.S., attempts have been made not only to analyze the causes of dropouts but also to take measures to prevent them, with Seidman emphasizing the importance of early identification and intervention for dropout prevention [9].

One promising method for dropout prevention is early and accurate measurement of academic and social integration. Tinto et al. used subjective measures based on students' self-report s and questionnaires to measure academic and social integration; however, their methods are undergoing reexamination. Thomas [10] focused on social integration and proposed a more objective measurement method using social networks rather than students' subjective opinions.

In analyzing data on academic integration, some methods utilize subjective measures such as the learner's self-reported degree of understanding of the class. However, objective and quantitative indicators such as students' grade data for each subject have recently become mainstream for incorporation into the analysis. Self-regulation is an important aspect of learning, and the information obtained from the learning process and its results are considered useful for capturing not only academic factors but also non-academic factors, such as attitudes and appetite toward learning [11].

Recently, methods have been developed to constantly analyze learners' behaviors using LMS and eBook log data [12]. For example, some systems predict academic performance from eBook logs and issue early warnings [13], while others intervene with students based on warnings in MOOCs [14]. Other attempts have been made to numerically predict each student's grade score using eBook logs and machine learning methods rather than predicting with or without a student's dropout risk [15].

There are other advantages to using LMS logs in addition to their facilitation of continuous data analysis. Although active student comments and participation are noticeable in in-person classes, it is difficult to capture passive student motivations. Beaudoin [16] suggested that even students with passive attitudes might be fully engaged in learning without losing motivation. We believe that LMS log data can help capture this type of learning behavior and awareness, which does not actively surface.

3 Proposed Method

Following previous studies, this study uses LMS log data to capture factors that lead to certain outcomes, including withdrawal from classes, based on more detailed and objective indicators. The aim is to predict, with high accuracy throughout the class period, whether each student will eventually receive credits. Log data in each class are analyzed using machine learning methods to achieve this. Previous surveys have found that credit completion rate is directly related to graduation and highly correlated with the dropout rate. One academic factor—whether students can earn credits for each course—helps measure the risk of dropping out of university. Additionally, we believe that information on learning attitudes, such as the frequency of LMS access and study time used in the analysis process, will also lead to the measurement of non-academic factors.

3.1 Analyzed Data

The College of Humanities and Sciences of Nihon University (CHS) uses Blackboard [17] as a common LMS within the faculty. The CHS has over 3,500 classes, and all data for all classes are stored on Blackboard. Data such as lecture videos, lecture materials (e.g., PDF), mini-tests, assignments, and scores are accumulated in the database. In addition, when students access each piece of content is recorded each time they click on it within the site.

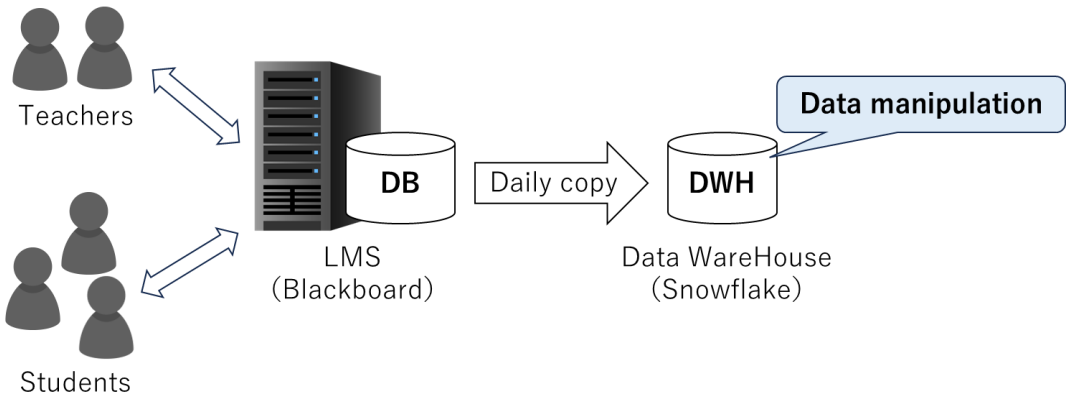


Figure 1: System Configuration

All data in the Blackboard database are sent daily to Snowflake [18], an external DB (Figure 1). By accessing the copied DB, high-load API calls and SQL queries can be executed without overloading the DB server used in the actual class.

Students' grades are given after accessing each piece of content and working on mini-tests and assignments. However, we believe that the stage at which grades are determined is too late to support early intervention to prevent dropouts. Data on grades must be available while or immediately after work is graded to allow for faster and more efficient assistance for at-risk students.

3.2 Target Classes

The CHS has 18 departments comprising humanities, social sciences, and natural sciences, with approximately 2,000 students per academic year (AY). For the analysis, we used data from the undergraduate course "Information Literacy." This class was designed to provide students with a

broad overview of basic ICT, including e-mail usage, academic writing, data analysis, presentation skills, operation of Microsoft Office products, information security, hardware, and software.

The CHS offered 15 classes per subject, all of which were conducted face-to-face until AY 2019. However, owing to COVID-19, PDFs of teaching materials and recorded lecture videos could be viewed on-demand during AY 2020–2021. Since AY 2022, approximately half of the classes have been offered as on-demand classes and the other half as in-person classes (Table 1). Which classes are face-to-face or on-demand varies annually because the order of topics covered in class changes every year.

Table 1: Information literacy class format（P: In-Person Class、D: On-Demand Class）

AY	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2021	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D
2022	P	P	D	D	P	P	P	P	P	D	P	D	D	D	D
2023	P	P	P	P	P	P	P	P	D	D	D	D	D	D	P

Owing to the large number of students in the CHS, class instruction is shared by 5–6 faculty members and 15 classes are offered each year. Before 2020, each faculty member set up their own teaching materials and tests after the content was determined among them. However, after 2021, courses on the LMS were integrated, with everyone using the same teaching materials and lecture videos delivered in on-demand class sessions. Therefore, the quality of the class content and lectures was almost the same in all classes.

Based on the above, this study uses data from the LMS for the same subject from 2021 to 2023 to predict which students will not receive credits.

3.3 Prediction of Credit Completion

Following previous research, we used LightGBM, a machine learning method, to estimate whether a student received credits [19][20]. First, the label information on whether credits were earned was used as the objective variable. In our preliminary study, we used the number of students accessing course materials, mini-test scores, and attendance at face-to-face classes through the fifth week to estimate whether the students would eventually earn credits. Consequently, we were able to estimate whether each student would acquire credits with approximately 94% accuracy before the remaining ten classes. However, some students that the AI predicted would be able to earn credits ultimately failed the course. A detailed analysis of the data showed that such students had good grades and no problems accessing the LMS until the fifth class, after which the situation changed.

Therefore, this study proposes a model that continuously adds explanatory variables starting from the first week of class to predict whether final credits will be earned (Figure 2). As a specific explanatory variable, in light of research showing that the level of engagement with the assignment affects the quality of learning [21], we implemented a machine learning model comprising the following four items.

- Whether a student took a mini-test for the week in question
  ex.) “submission_01” shows 1st week’s submission status
- Percentage of a student’s scores on the mini-test for the week in question
  ex.) “score_01”
- Number of times a student accessed the week’s materials
  ex.) “access_01_contents”

・How many times students accessed other week's materials and communications
ex.) "access_01_all"

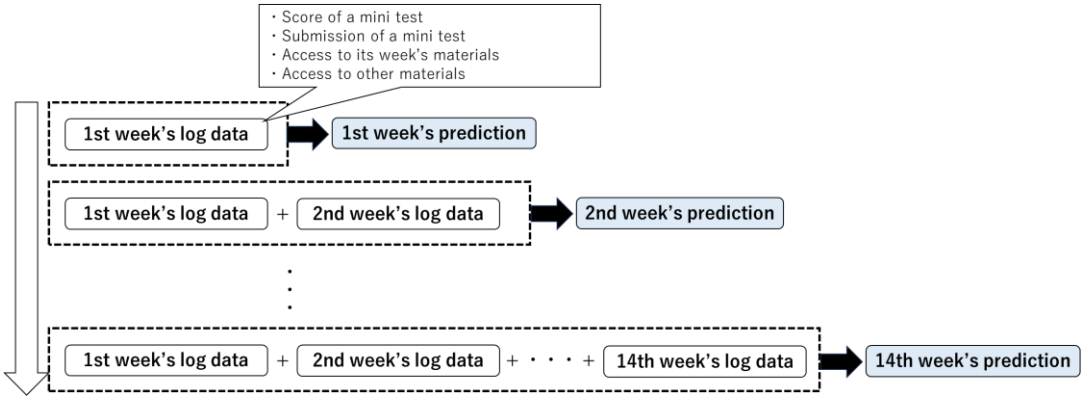
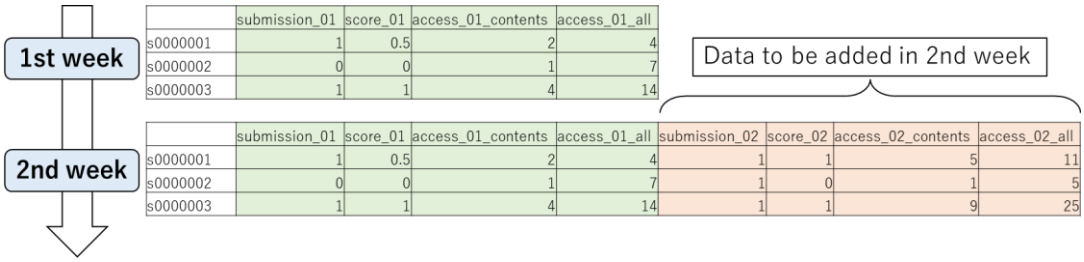


Figure 2: Proposed Model

An example of specific data is shown in Figure 3.



	submission_01	score_01	access_01_contents	access_01_all
s0000001	1	0.5	2	4
s0000002	0	0	1	7
s0000003	1	1	4	14

	submission_01	score_01	access_01_contents	access_01_all	submission_02	score_02	access_02_contents	access_02_all
s0000001	1	0.5	2	4	1	1	5	11
s0000002	0	0	1	7	1	0	1	5
s0000003	1	1	4	14	1	1	9	25

Figure 3: Examples of Specific Data to be Used in a Proposed Model

4 Result

We describe the aggregate data of three years of grades in the Information Literacy course and the prediction of credit acquisition by machine learning.

4.1 Aggregate Data

Table 2 shows the number of students, the number of credits earned, and the total number of LMS accesses to course content in the class for each year.

Table 2: Total Number of Accesses to Teaching Materials and Credit Acquisition Rate

AY	Students	Total number of times content was accessed	Students earning credits	Students not earning credits
2021	1,948	328,364 (Avg.168.6)	1,645 (84.4%)	303 (15.6%)
2022	2,005	343,396 (Avg.171.3)	1,784 (89.0%)	221 (11.0%)
2023	2,185	332,952 (Avg.152.4)	1,951 (89.3%)	234 (10.7%)

Although there is a slight difference in the total number of times class content was accessed, the difference is little because the number of teaching materials (videos and PDFs) created by each faculty member is also slightly different. The non-credit rate is slightly higher in AY 2021. This year's class was fully on-demand, which we believe was appropriate in light of previous

research showing that dropout rates are higher in online classes [8].

4.2 Evaluation of Prediction of Credit Completion

As shown in Figure 2, we estimated whether each student would eventually acquire credits using LightGBM, with the number of times the course material was accessed and mini-test scores for each week as explanatory variables. The values evaluated by cross-validation by year are listed in Table 3. The values in the table are the averages of the cross-validation with 100 random seeds. For example, in AY 2021, the accuracy was 0.852, and the F-measure was 0.916 when using log data from only the first week. By contrast, the accuracy and F-measure improved to 0.915 and 0.951, respectively, when log data up to week 10 were used. The estimation accuracy improved with each week of data accumulation for all years.

Table 3: Evaluate the Model through Cross-Validation

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
2021	Accuracy	0.852	0.855	0.866	0.874	0.890	0.893	0.899	0.911	0.909	0.915	0.920	0.929	0.933	0.947
	F-Value	0.916	0.917	0.923	0.928	0.937	0.938	0.941	0.948	0.947	0.951	0.954	0.958	0.961	0.969
2022	Accuracy	0.886	0.900	0.907	0.917	0.927	0.931	0.937	0.944	0.952	0.953	0.955	0.955	0.959	0.961
	F-Value	0.938	0.945	0.949	0.954	0.960	0.962	0.965	0.969	0.973	0.974	0.975	0.975	0.977	0.978
2023	Accuracy	0.894	0.892	0.892	0.910	0.916	0.926	0.933	0.934	0.941	0.942	0.948	0.957	0.959	0.961
	F-Value	0.943	0.941	0.941	0.951	0.954	0.960	0.963	0.964	0.968	0.968	0.971	0.976	0.977	0.978

Although the AI by lightGBM could more accurately estimate credit completion as students progressed to later stages of the course, it still fails to predict whether students will earn credits. Particular attention should be paid to cases in which the AI estimates that a student can earn credits but the student fails to do so.

Below is a discussion of the students whose outcomes differed from the AI predictions. Each week's mini-test accumulated to 60% of the total grade. The remaining 40% was divided evenly between the submission of reports and the final exams. In this study, we estimated whether students could earn credits based solely on their scores on mini-tests. Therefore, although the mini-test score alone showed no problems, the AI missed the predicted credits earned by students who did not submit reports or take the final exams in the last week of class. If we had included the final-week exam, it would have already been too late in the intervention; therefore, we excluded it as an explanatory variable.

Although the submission status and score of the report assignment were important, it was difficult to incorporate them as explanatory variables because the timing of the report assignment differed annually.

Next, an AI model was generated using log data from AY 2021 and 2022 as training data to estimate whether students could earn credits in AY 2023 (Table 4).

Table 4: Projection of Student Credit Completion in AY2023

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
2023	Accuracy	0.897	0.897	0.898	0.913	0.918	0.916	0.928	0.932	0.937	0.941	0.915	0.919	0.942	0.930
	F-measure	0.945	0.945	0.945	0.953	0.955	0.954	0.961	0.963	0.966	0.968	0.952	0.953	0.967	0.960

It can be confirmed that the prediction is made with relatively high accuracy, mainly in the first half. However, compared with Table 4, where the same dataset was used for training and evaluation, the rate of increase in accuracy in the second half is not stable and is slightly lower.

5 Conclusion

Some students who fail to acquire credits in a course show signs early, while others suddenly show signs in the middle or later stages of the course. Therefore, to provide effective assistance to a larger number of students, such students must be identified at an early stage and on an ongoing basis.

This study focused on LMS logs as data and clues to consistently capture the signs of student dropout before grades were assigned. We obtained highly accurate estimates from an early stage using the number of times teaching material was accessed and weekly mini-test data as explanatory variables in a required undergraduate course to predict students at risk of not earning credits. Furthermore, by continuing to make predictions by adding log data and grades accumulated each week, we increased the accuracy and captured students who showed signs of dropping out midway through the course.

Our study had some limitations which should be addressed in future research. In this study, we estimated student dropout with high accuracy by applying machine-learning methods to a single subject. However, classes vary in style, and it is conceivable that some classes may not have weekly assignments. Therefore, future research should gather data from more courses to use as training data to make the model generalizable across courses. We will work toward a more versatile prediction model for at-risk students that uses only data that occur regardless of the class style, such as the number of times a course was accessed.

References

[1] Ministry of Education, Culture, Sports, Science and Technology, https://www.mext.go.jp/en/index.htm (accessed 2023-10-23).

[2] Tinto,V., "Dropout from Higher Education: A Theoretical Synthesis of Recent Research," Review of Education Research, Vol.45, No.1, pp.89-125, 1975.

[3] Tinto,V, "Classrooms as Communities: Exploring the Educational Character of Student Persistence," The Journal of Higher Education, Vol.68, No.6, pp.599-623, 1997.

[4] ASTIN, A.W., "Preventing students from dropping out, " Jossey-Bass Inc Pub, San Francisco, 1975.

[5] Robbins, S.B., Lauver, K., Le, H., Davis, D., Langley, R., Carlstrom, A., "Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis," Psychological Bulletin, Vol.130, No.2, pp.261-288, 2004.

[6] Mangold,W.D., Bean,L.G.,Adams,D. J., Schwab,W.A.,Lynch,S.M., "Who Goes Who Stays: An Assessment of the Effect of a Freshman Mentoring and Unit Registration Program on College Persistence," Journal of College Student Retention: Research, Theory & Practice, Vol.4, No.2, pp.95-122, 2002.

[7] Brown, G. T. L., Peterson, E. R., & Yao, E. S. , "Student Conceptions of Feedback: Impact on Self-Regulation, Self Efficacy, and Academic Achievement.," British Journal of Educational Psychology,Vol. 86, pp.606-629, 2016.

[8] Salmon,G., "E-moderating: The Key to Teaching and Learning Online," Routledge, London, 2021.

[9] Seidman,A. "College Student Retention: Formula for Student Success Second Edition," Rowman & Littlefield Publishers. 2012.

[10] Thomas, S. L., "Ties that Bind," The Journal of Higher Education, Vol.71, No.5, pp.591-615, 2000.

[11] Zimmerman, B.J., & Moylan, A.R. , "Self-Regulation: Where Metacognition and Motivation Intersect.," Handbook of Metacognition in Education, pp.299-316, 2009.

[12] 11 Oi, M., Okubo, F., Shimada, A., Yin, C., Ogata, H., "Analysis of Preview and Review Patterns in Undergraduates' e-Book Logs," Proceedings of the 23rd International Conference on Computers in Education, pp.166–171, 2015.

[13] Akcapınar,G., Hasnine, M. N., Majumdar, R., Flanagan, B., & Ogata, H., "Developing an Early-Warning System for Spotting At-Risk Students by using eBook Interaction Logs," Smart Learning Environments, Vol.6, No.4, pp.1–15, 2019.

[14] Wan, H., Liu, K., Yu, Q., & Gao, X, "Pedagogical Intervention Practices: Improving Learning Engagement based on Early Prediction," IEEE Transactions on Learning Technologies, Vol.12, No.2, pp.278–289, 2019.

[15] Flanagan,B., Majumdar,R. & Ogata, H., "Early-Warning Prediction of Student Performance and Engagement in Open Book Assessment by Reading Behavior Analysis," Int J Educ Technol High Educ, Vol.19,No.41, 2022.

[16] Beaudoin, M. F., "Learning or lurking?: Tracking the "invisible" online student," The Internet and Higher Education,Vol.5, No.2, pp.147–155, 2002.

[17] Educational Technology Services - Blackboard, Blackboard.Inc, https://www.blackboard.com (accessed 2023-10-23).

[18] The Snowflake Data Cloud - Mobilize Data, Apps, and AI, Snowflake Inc, https://www.snowflake.com/en/ (accessed 2023-10-23).

[19] Kary.N, Philip.L, "A Lightweight Method using LightGBM Model with Optuna in MOOCs Dropout Prediction," Proceedings of the 6th International Conference on Education and Multimedia Technology, pp.53-59, 2022.

[20] Welcome to LightGBM's documentation, Microsoft Corporation, https://lightgbm.readthedocs.io/en/v4.1.0/ (accessed 2023-10-23).

[21] Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M., "Penetrating the Black Box of Time-on-Task estimation," Proceedings of the 5th International Conference on Learning Analytics and Knowledge, pp.184–193, 2015.