LMS Log Data Analysis from Fully-Online Flipped Classrooms: An Exploratory Case Study via Regularization

Jin Eun Yoo
Korea National University of Education
jeyoo@knue.ac.kr

Minjeong Rho
Korea National University of Education
minjeong019@gmail.com

ABSTRACT
This study illustrated an exploratory study of LMS log data from undergraduate fully-online flipped classrooms. A total of 237 students' instructional video watching behaviors were extracted from LMS, and were analyzed with background variables to predict students' final performance. Regularization was proposed a suitable machine learning technique, as it produces interpretable prediction models. Specifically, Enet (elastic net) and Mnet were employed to handle possible multicollinearity in LMS log data, and the prediction models of Enet and Mnet identified 19 and 21 important predictors of final performance out of 157, respectively. In particular, both regularization models were able to screen lower-performing students as early as the first week of the course. Mere attempts to watch difficult videos after class increased the final scores.

Keywords
LMS log data, machine learning, regularization, flipped classroom, performance modeling

1. INTRODUCTION
The COVID-19 pandemic has changed the education system worldwide. Online learning is no longer an option, and an increasing number of online classes incorporate components of flipped classrooms (FC) in an effort to improve the quality of learning and instruction. Despite varying results regarding the effectiveness of flipped learning in higher education [1, 2, 3, 4], FC has grown rapidly as an innovative pedagogical approach in recent decades. In FCs, students' active involvement in pre-class activities is greatly emphasized as a necessary condition to enhance in-class learning and instruction [5]. However, there has been little empirical research on whether students completed the assigned pre-class activities and whether pre-class activities lead to desired outcomes.

This may relate to analytical limitations of the previous research in terms of data and methods. First of all, learning management system (LMS) log data are a crucial source of information in order to capture students' learning activities. However, not all the studies on FC collected data from LMS, particularly when pre-class assignments do not involve activities in LMS. For instance, the assignment of reading materials cannot be properly recorded outside of LMS. Researchers can ask students ex post facto in a self-report survey. However, self-report questionnaires rely on memories and reflections, and thus are prone to social desirability bias. On the other hand, LMS log data unobtrusively collect near-real-time information; students' activities in LMS are automatically stored in the log files without the students' cognizance [6, 7, 8, 9]. Particularly in the COVID-19 situation, fully-online FCs has emerged. In fully-online FCs both pre-class and in-class activities take place online using platforms such as LMS, and therefore collecting trace data has become much easier than in the original FCs.

Next, there is room for improvement in terms of analysis methods. Despite the aforementioned advantages that log data bring to data analyses, the intractability of log data has been a practical hindrance. Log data are unstructured, which can lead to high-dimensional data (i.e., more variables than observations), depending on data pre-processing and cleaning. Previous research on LMS log data to model students' achievement have analyzed students' behavioral data (e.g., instructional video watching behaviors) [6, 10, 11, 7, 12] as well as background (e.g., gender) [10, 13] and exam data [11, 13]. In particular, behavioral data were used as a tool to measure students' self-regulated learning [6, 10, 11, 7, 13, 12, 14, 15, 16], but aggregate variables such as total login frequencies or average login hours were analyzed with traditional methods [13, 15] or early ML (machine learning) techniques [14, 16]. As traditional methods are likely to result in nonconvergence problems with high-dimensional data, previous research may have used aggregate variables.

However, study time relevant to a specific instructional unit can be traced from log data, which will serve as a better indicator than the sum of study time, a crude measure of time investment in studying [15]. Such detailed information in turn will be conducive to understanding learning and instruction and giving specific, targeted, and timely feedback to students. This relates back to the issue of the previous LMS log data research: lack of empirical research on the relationship between pre-class assignments and students' performance at an instructional unit level. Particularly when behavioral variables at an instructional level are to be ana-
lyzed, ML is a necessary technique to analyze LMS log data from fully-online FCs.

Since completing pre-class assignments and preparing for interactive in-class activities is critical in FC, a high level of self-regulated learning (SRL) is necessary for students to succeed. SRL strategies related to students’ academic success such as effective time management, metacognition, effort regulation, and critical thinking have been shown to have a significantly positive effect on students’ academic success [17]. The question is which behaviors indicate SRL. Students carrying out SRL would naturally include more time on attending lectures and self-study which have a positive effect on academic achievement [18]. Previous studies have used variables such as login frequencies, LMS menu usage, material download, content pages viewed, and posted messages [6, 10, 11, 13, 14, 15, 16]. However, aggregate measures of these data display inconsistent effects on student achievement. For instance, login frequencies [13, 16] and LMS menu usage [13, 14, 15] were statistically significant or important indicators to students’ academic achievement in online learning. In contrast, in MOOC (massive open online course) environments, forum variables such as numbers of messages posted, or comments received were found to be not directly related to students’ learning [11].

Constant effort put into preparing for FCs may be difficult to capture with aggregate data. That is, instructional unit based log data would be a better predictor for academic success. A study predicting online student performance [19] demonstrates that the study habits of students with high levels of academic success can even be observed even in the first few weeks of a course. The implication is that instructional unit based analysis could yield richer information about the study patterns of students which eventually leads to timely intervention by the instructor.

Among ML, this study proposes regularization. Although ‘prediction’ is the operative word in ML, learning analytics is one of the fields which needs to be augmented with explanation. Regularization or penalized regression is known to produce explainable prediction models. Based on linear regression, the regression coefficients of regularization can be interpreted in the similar way as those in traditional, non-penalized regression. This is a great advantage in LMS data analysis, as prediction models need to be interpreted under certain educational settings, for instance to plan more effective intervention strategies for at-risk students. There has been little study employing regularization methods in LMS log data analysis. Specifically, this study chose Enet [20, 21] and Mnet [22] among regularization as they handle multicollinearity, a likely challenge in LMS data analysis. The two main research questions were as follows:

1. What are the students’ instructional video watching behaviors like at an instructional unit level? Do students complete pre-class assignments in fully-online undergraduate flipped classrooms?

2. Among students’ behavioral and background variables, which variables are important to predict students’ academic achievement?

2. MACHINE LEARNING

For a Gaussian family, Enet and Mnet are expressed as equations 1 and 2, respectively. The second term on the right-hand side of equation 1 is the penalty function of Enet, consisting of two tuning parameters: $\lambda$ and $\alpha$. Enet is a combination of LASSO and ridge. The parameter $\lambda$ regularizes shrinkage of the coefficients, and the parameter $\alpha$ controls the amount of ridge. When $\alpha$ is 1, equation 1 reverts to the LASSO equation, and when $\alpha$ is 0, it reverts to the ridge equation. Aforementioned, by adding the ridge component to the equation, Enet can handle multicollinearity.

$$\hat{\beta}_{\text{Enet}} = \arg\min_{\beta} \left[ \frac{1}{2n} \left\| y - \sum_{k=1}^{K} X_k \beta_k \right\|^2 + \lambda \sum_{k=1}^{K} (\alpha \left\| \beta_k \right\| + (1 - \alpha) \left\| \beta_k \right\|^2) \right].$$ (1)

$$\beta_{\text{Mnet}} = \arg\min_{\beta} \left[ \frac{1}{2n} \left\| y - \sum_{k=1}^{K} X_k \beta_k \right\|^2 + \lambda_1 \sum_{k=1}^{K} \left| \beta_k \right| + \lambda_2 \sum_{k=1}^{K} \left| \beta_k \right|^2 \right],$$ (2)

where $J(x|\lambda_1, \gamma) = \left\{ \begin{array}{ll} -\frac{x^2}{2\gamma^2} + \lambda_1 |x|, & |x| \leq \gamma \lambda_1 \\ \frac{x^2}{2\gamma^2} \lambda_1^2, & |x| > \gamma \lambda_1 \end{array} \right.$

Enet uses convex penalties, which increase linearly regardless of the coefficient size. By contrast, Mnet uses a concave penalty, which tapers off for coefficients in larger absolute values, yielding nearly consistent coefficient estimates [22]. Mnet has three tuning parameters (equation 2). The parameter $\lambda_1$ has the same regularization function as the $\lambda$ penalty in Enet (equation 1). The $\gamma$ parameter of Mnet controls the concavity of the convex penalty. When the concavity penalty goes to infinity, the MCP penalty reverts back to the LASSO penalty. Mnet also deals with multicollinear data; the penalty associated with $\lambda_2$ adds the ridge component to the equation.

To consider the bias resulting from data-splitting in model validation, this study executed subsampling techniques for variable selection [23, 24]. The following three steps were repeated 100 times with random data-splitting. First, the whole data were randomly divided with the ratio of 7:3 to get the training and test data, respectively. Second, for a value of the penalty parameter, the training data were split with the ratio of 4:1 to execute 5-fold CV. For a value of $\lambda$, the prediction error is calculated, which was referred to as the CV error of the $\lambda$ [20]. Third, the second step was repeated for every $\lambda$ in range, and the $\lambda$ of the lowest CV error served as the penalty value of the regularization. That $\lambda$ value was applied to the test data in step 1, which yielded prediction measure.

The selection or non-selection of each variable from step 2 was counted in the 100 iterations, which served as the selection counts of the study. Particularly, this study presented
variables selected 1, 25, 50, 75 times or more, and all 100 times [25, 26]. All the programs were written in R 3.6.2. Specifically, the grpreg library [27] was used for regularization.

3. MATERIALS
In the Fall semester of 2020, 242 undergraduate students in pre-service teacher program enrolled in 8 fully-online undergraduate classes titled Measurement and Evaluation. The classes of the Fall semester were mandatory for sophomores majoring in Liberal Arts and Social Sciences. Three instructors (A, B, C) including a head-instructor (A) taught the 8 classes. All the 8 classes scheduled a simultaneous final at the end of the course, and shared the same class materials including instructional videos, textbooks, and syllabus. The instructional videos were pre-recorded PowerPoint presentations with the head-instructor talking, with content based on a book also written by the head-instructor. There were a total of 34 video clips covering 11 instructional topics in the corresponding 11 instructional weeks (refer to the videos 01_1 to 11_4 in Appendix A).

On the orientation day of the first week, the importance of the weekly assignments of instructional video watching before class was emphasized, particularly because students were asked to create and complete class projects within groups based on the contents of the assigned videos. During class, interactions in small groups of 4 to 6 students were greatly encouraged. The groups were engaged in discussions on team projects and SPSS exercises in Zoom breakout rooms. A non-mandatory quiz of 4-5 short questions was presented for each week in LMS. Students were told that the quizzes would serve as formative assessments and the quiz scores did not count toward grades.

In total, 21,589 rows of video watching activities as well as 5,107 rows on board-posts readings were recorded in the log file. As many of the students indicated that they used the double-speed option of the LMS in video watching, this study used 50% of the video length as a criterion. If a student watched a video 50% of the length or more, the student is counted to have completed watching the video, and vice versa.

As the first research question was to investigate students’ video watching behaviors at an instructional unit level, this study counted the frequencies of each video, separating before/after and attempted but incomplete/completed video watching. Specifically, 4 variables were created for each video: BI (incomplete attempt before class), BC (complete watching before class), AI (incomplete attempt after class), and AC (complete watching after class). Six Aggregate variables were also obtained for comparison purposes to previous research: BI, BC, and B (before-combined (I+C)) for before class counts; and AI, AC, and A (after-combined (I+C)) for after class counts.

The response variable of this study was final. The final test consisted of 35 multiple-choice items, and was given simultaneously to all the 242 students at the last week of the course. There were 5 students who missed the final, and those students’ data were excluded from further analysis. The background and response variable were merged to the variables from LMS data, which resulted in the final dataset of 157 predictors of 237 students.

4. RESULTS
4.1 Students’ Video Watching Behaviors
Table 1 summarizes the descriptive statistics of students’ instructional video watching behaviors. The 6 groups of cells present the summary results of the aggregate variables. Students watched the videos more often after class than before class. Throughout the course students on average attempted to watch and completed watching each video about 1.03 and 1.08 times after class, respectively, while the values dropped to 0.20 and 0.23 before class (Table 1). The mean values smaller than 1 indicate that the students on average did not watch all the videos. Attempts and completions combined (I+C), students on average clicked about half of the videos before class (0.42), but they clicked each of the videos more than twice after class (2.11).

The range of students’ video watching frequencies was quite wide. Some students clicked none of the videos after class (AI min= 0.00), while others after class clicked and finished watching each video as many as 4.20 (AI max= 4.20) and 2.47 times (AC max= 2.47), respectively. The maximum frequencies of before class watching were also less than those of after class, 1.38 and 1.00 for incomplete and complete watching, respectively.

4.2 Machine Learning Results
4.2.1 RMSE and Selection Counts
RMSE (root mean square error) was the prediction measure of the response variable of this study. The RMSE averages of Enet and Mnet were 5.58 and 5.69 with SDs of 0.50 and 0.46, respectively.

Consistent with literature [28, 22], Mnet always selected fewer variables than Enet. Of note, 103 and 94 predictors were selected out of 157 at least once with Enet and Mnet, respectively. This signifies the importance of running multiple iterations and employing selection counts, particularly when the research purpose is variable selection via regularization [25, 26]. In other words, employing selection counts considers the bias resulting from random data-splitting in model building.

Applying 25 or more selection counts resulted in 33 and 21 predictors for Enet and Mnet, respectively. A total of 19 and 3 predictors were selected at least 1 out of 2 runs of Enet and Mnet, respectively. This signifies the importance of running multiple iterations and employing selection counts, particularly when the research purpose is variable selection via regularization [25, 26]. In other words, employing selection counts considers the bias resulting from random data-splitting in model building.

4.2.2 Selected Variables
This study on log data analysis presents the summary of predictors selected 50 or more for Enet and 25 or more for Mnet in Table 2. Due to space limit, part of the results are discussed. Student gender and grade were selected important. When the other variables were held constant, male students had lower final score than female students. Interestingly, on-grade students, sophomores, tended to have lower scores. Students’ attitudes toward measurement and evaluation (attitudes) also resulted in higher scores.
Among variables extracted from log data, the total number of clicks on SPSS material postings (spss.sum) and the numbers of quiz-taking (test.M and test.P) were important predictors to final. More clicks on SPSS postings lead to higher scores on final. Specifically, one more click on the SPSS material increased students’ final scores by 0.11 and 0.16 for Enet and Mnet, respectively. Similarly, although students knew that the scores on quizzes did not count toward the final grade, simply taking the quizzes increased the final scores regardless of the device (mobile or PC). Students who watched the instructional videos mobile (lecture.M) also tended to have higher scores in final.

Among the 142 variables on video watching, 12 to 13 variables were selected as important depending on the regularization method (Table 2). Findings from the 13 selected variables are as follows. First, the very first video turned out to convey crucial information in predicting students’ achievement, although it covered the easiest contents on formative assessment. The more the students completed watching the first video before class (BC01_1), the higher their final scores were. Specifically, one more completed watching of the first video before class increased students’ final score by 0.57 in Enet and by 0.68 in Mnet. By contrast, the more the students completed watching the first video after class (AC01_1), the lower the final scores were. One more completed watching of the first video after class decreased students’ final scores by 0.45 in Enet and by 0.53 in Mnet. By contrast, the more the students completed watching the first video after class (AC01_1), the lower the final scores were. One more completed watching of the first video after class decreased students’ final scores by 0.45 in Enet and by 0.53 in Mnet. By contrast, the more the students completed watching the first video after class (AC01_1), the lower the final scores were. One more completed watching of the first video after class decreased students’ final scores by 0.45 in Enet and by 0.53 in Mnet. By contrast, the more the students completed watching the first video after class (AC01_1), the lower the final scores were. One more completed watching of the first video after class decreased students’ final scores by 0.45 in Enet and by 0.53 in Mnet. By contrast, the more the students completed watching the first video after class (AC01_1), the lower the final scores were.

Second, with the exception of AC01_1, the other AC variables (e.g., AC03_1, AC03_2, AC04_3, AC09_3, AC10_2, AC11_4) had positive relations with the final. The selected AC variables covered either the earlier technical contents or the most difficult concepts at the end. Particularly, the earlier technical contents included the first SPSS practice (AC03_1, AC03_2) and Ebel and Angoff standard setting (AC04_3). Cronbach’s alpha (AC09_3), reliability with SPSS (AC10_2), and the relationship between reliability and validity (AC11_4) covered the most difficult concepts in the last weeks of the course. Students who completed watching these videos multiple times after class were more likely to obtain higher final scores.

Third, the relationship of AI variables to the final seems to depend on the class progress. Students who attempted but failed to complete watching the video on Ebel and Angoff standard setting covered in the fourth instructional week had lower scores on final (AI04_3). By contrast, incomplete watching of some videos on the last topic (covered in the last instructional week) were positively related to the final (AI11_2 and AI11_4). Of note, both AC and AI variables on video 11_4, the last video, had positive correlation coefficients.

5. DISCUSSION

This study predicted students’ final scores with as few as 19 to 21 predictors out of 157 with regularization techniques. Of note, the prediction models of this study are explainable, as we employed regularization, which is based on linear regression. Specifically, Enet and Mnet were employed to handle multicollinear data. Surprisingly, the prediction models differentiated lower-performing students as early as the first instructional week, right after the orientation week. Instructors now can invest their efforts in intervention without waiting until a quiz or an exam. Completing difficult videos multiple times after class also lead to higher scores in the final. Moreover, mere attempts to watch them after class also increased the scores.

Despite its importance in FC, it has been a foggy area whether students completed the pre-class activities or not and whether the pre-class activities lead to desired outcomes. This study also contributed to partly uncover what was going on behind the curtain of FC. The students on average completed at most 1/5 of the videos before class. Stronger links need to be established between pre-class assignments and in-class team projects.
6. REFERENCES

# APPENDIX

## A. VIDEO IDS AND LABELS

<table>
<thead>
<tr>
<th>video ID</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1_1 formative assessment</td>
</tr>
<tr>
<td>2</td>
<td>2_1 variables and scales</td>
</tr>
<tr>
<td>3</td>
<td>2_2 sampling</td>
</tr>
<tr>
<td>4</td>
<td>3_1 descriptive statistics</td>
</tr>
<tr>
<td>5</td>
<td>3_2 descriptive statistics (SPSS)</td>
</tr>
<tr>
<td>6</td>
<td>4_1 norm-referenced evaluation</td>
</tr>
<tr>
<td>7</td>
<td>4_2 criterion-referenced evaluation</td>
</tr>
<tr>
<td>8</td>
<td>4_3 Ebel and Angoff standard setting</td>
</tr>
<tr>
<td>9</td>
<td>5_1 measuring affective domains</td>
</tr>
<tr>
<td>10</td>
<td>5_2 observation</td>
</tr>
<tr>
<td>11</td>
<td>5_3 interviews</td>
</tr>
<tr>
<td>12</td>
<td>5_4 survey</td>
</tr>
<tr>
<td>13</td>
<td>6_1 performance assessment: definition</td>
</tr>
<tr>
<td>14</td>
<td>6_2 performance assessment: scoring</td>
</tr>
<tr>
<td>15</td>
<td>7_1 test construction steps</td>
</tr>
<tr>
<td>16</td>
<td>7_2 multiple-choice items</td>
</tr>
<tr>
<td>17</td>
<td>7_3 constructed-response items</td>
</tr>
<tr>
<td>18</td>
<td>7_4 scoring caveats</td>
</tr>
<tr>
<td>19</td>
<td>8_1 item difficulty and discrimination I</td>
</tr>
<tr>
<td>20</td>
<td>8_2 covariance and correlation</td>
</tr>
<tr>
<td>21</td>
<td>8_3 item difficulty and discrimination II</td>
</tr>
<tr>
<td>22</td>
<td>8_4 item difficulty and discrimination (SPSS)</td>
</tr>
<tr>
<td>23</td>
<td>9_1 introduction to reliability</td>
</tr>
<tr>
<td>24</td>
<td>9_2 types of reliability</td>
</tr>
<tr>
<td>25</td>
<td>9_3 Cronbach’s alpha</td>
</tr>
<tr>
<td>26</td>
<td>9_4 standard error of measurement</td>
</tr>
<tr>
<td>27</td>
<td>9_5 factors influencing reliability</td>
</tr>
<tr>
<td>28</td>
<td>10_1 objectivity and reliability</td>
</tr>
<tr>
<td>29</td>
<td>10_2 reliability (SPSS)</td>
</tr>
<tr>
<td>30</td>
<td>10_3 objectivity (SPSS)</td>
</tr>
<tr>
<td>31</td>
<td>11_1 content validity</td>
</tr>
<tr>
<td>32</td>
<td>11_2 criterion-related validity</td>
</tr>
<tr>
<td>33</td>
<td>11_3 construct validity</td>
</tr>
<tr>
<td>34</td>
<td>11_4 reliability and validity</td>
</tr>
</tbody>
</table>