Lee, J. & Recker , M. & Choi , H. & Hong , W. & Kim , N. & Lee, K. & Lefler , M. & Louviere , J. & Walker , A. (2016).
Applying Data Mining Methods to Understand User Interactions within Learning Management Systems: Approaches and Lessons Learned. *Journal of Educational Technology Development and Exchange*, 8(2), 99-116.

Applying Data Mining Methods to Understand User Interactions within Learning Management Systems: Approaches and Lessons Learned

Ji Eun Lee
Mimi M. Recker
Hongkyu Choi
Won Joon Hong
Nam Ju Kim
Kyumin Lee
Mason Lefler
John Louviere
Andrew Walker
Utah State University

Abstract: *This article describes our processes for analyzing and mining the vast records of instructor and student usage data collected by a learning management system (LMS) widely used in higher education, called Canvas. Our data were drawn from over 33,000 courses taught over three years at a mid-sized public Western U.S. university. Our processes were guided by an established data mining framework, called Knowledge Discovery and Data Mining (KDD). In particular, we use the KDD framework in guiding our application of several educational data mining (EDM) methods (prediction, clustering, and data visualization) to model student and instructor Canvas usage data, and to examine the relationship between these models and student learning outcomes. We also describe challenges and lessons learned along the way.*

Keywords: Educational Data Mining (EDM), Learning Management System (LMS), Knowledge Discovery and Data Mining

1. Introduction

Institutions of higher education are increasingly turning to learning management systems (LMSs) to help support their instructional missions (Hawkins & Rudy, 2007; Smith, Lange, & Huston, 2012). An LMS provides features to support teaching and learning activities, such as providing access to instructional content, delivering quizzes, conducting assessments, and supporting online communications and collaborations between students and teachers. In recent years, the use of LMSs to support teaching has become

nearly universal in higher education, with 99% of institutions in the U.S. reporting use of an LMS (Dahlstrom, Brooks, & Bichsel, 2014). Use of LMSs to support online teaching is also growing rapidly. As of 2012, 26% of U.S. undergraduate students took at least one course online (National Center for Education Statistics, 2015).

An LMS automatically stores all instructor and student online interactions, collected as a part of natural instructional activity. The increasing availability of these datasets coupled with emerging "big data" and educational data mining (EDM) techniques offer unparalleled opportunities for research on understanding learning and teaching in higher education (Macfadyen & Dawson, 2010). This is a particularly rich educational context, encompassing different modalities (online, blended, and face-to-face), spanning disciplines, and enrolling different kinds of students (full-time, part-time, returning, etc.). These datasets include instructor as well as student usage data and are longitudinal.

The purpose of this research is to address the need to explore a range of analysis techniques and their meaningful application to a complex dataset collected by an LMS widely used in higher education, called Canvas. Our data come from over 33,000 courses taught over four years at a mid-sized public Western U.S. university. We describe our processes, modeling approaches, and results when applying several data mining methods, specifically prediction, clustering, and data visualization. We use these different models to examine student and instructor patterns of activities within (micro) courses and across (macro) courses. We also examine the relationship between model results and student learning outcomes, in particular, student final grades.

In presenting our processes, approaches, and results, we address the following research questions:

1) What were the strengths and weaknesses of our different data mining approaches in terms of modeling usage patterns and predicting outcomes? What were the relationships between these models and student learning outcomes?

2) What were challenges of and lessons learned from employing the various models?

2. Literature Review

2.1. Applying Educational Data Mining to Learning Management System Data

An LMS is typically engineered to record and store all learners and instructors interactions with particular LMS features. For example, students may view an assignment posting, submit assignments, download attachments, etc. This kind of activity, which was once invisible and ephemeral, can now be captured and mined as data points by researchers (Krumm, Waddington, Teasley, & Lonn, 2014; Long & Siemens, 2011). The resulting large data sets, when analyzed in conjunction with increased computing power and advances in analytical tools, have brought about new fields of educational research known as educational data mining (EDM), learning analytics, and academic analytics (Baker & Siemens, 2014; Long & Siemens, 2011). Though there are differences between these fields, they all share the goal of applying analytical tools to large data sets of learning data to help better understand student learning and improve educational experiences (Siemens & Baker, 2012). Given the focus of this article, we use the term EDM.

EDM techniques have been applied

to LMS data sets for many purposes. For example, EDM results have helped identify at-risk learners, examine the relationship between student performance characteristics (e.g., major, year of study) and learning outcomes, monitor student persistence, and inform instructional design (Angeli & Valanides, 2013; He, 2013; Krumm et al., 2014; Picciano, 2014; Siemens, 2013; Long & Siemens, 2011; Xu & Recker, 2011;).

In terms of EDM methods, five methods are widely used in educational research: prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models (Baker & Yacef, 2009; Bienkowski, Feng, & Means, 2012). In our review of EDM studies using LMS data, we found that prediction, clustering, and distillation of data for human judgment were the most commonly used. These are described next.

Prediction refers to developing a model that can predict a response (or outcome) variable, such as student performance, from some combination of predictor variables (Baker & Siemens, 2014). Many studies have used prediction methods in order to find significant predictors of student final grades. In our review of studies, several LMS variables were found to be significant predictors of student final grade including the number of discussion messages posted, the number of assignments completed, the number of quizzes taken (Macfadyen & Dawson, 2010; Thakur, Olama, McNair, Sukumar, & Studham, 2014), the number of interactions with peers, the number of file downloads, regularity in learning intervals (Yu & Jo, 2014), and themes in questions asked online (Abdous, He, & Yen, 2012). Other studies found that variables related to login frequency and the total amount of time online did not significantly predict student final grade (Jo, Kim, & Yoon, 2015; Macfadyen & Dawson, 2010).

Second, clustering is a method to find data points that naturally group a full dataset into smaller subsets (Bienkowski et al., 2012). The grouping object for clustering can be students, courses, or content. Our review found that most studies used student LMS usage data to group students (e.g., Hung, Hsu, & Rice, 2012; Lust, Elen, & Clarebout, 2013; Romero, Ventura, & García, 2008), whereas a few studies used the course or course content as the object of clustering (e.g., Abdous et al., 2012; Valsamidis, Kontogiannis, Kazanidis, Theodosiou, & Karakos, 2012). In terms of clustering algorithm, K-means was the most commonly used in studies using LMS data (Lust et al., 2013; Romero et al., 2008; Valsamidis et al., 2012).

Finally, distillation of data for judgment refers to depicting data, including data visualization, to enable humans to quickly identify and understand its features. At a micro level, visualization methods such as heatmaps, graphics, and scatter plots, can be used (Baker & Siemens, 2014). In a macro level, EDM results have been used to develop student monitoring and tracking dashboard systems, such as the Course Signals System (Arnold & Pistilli, 2012) and the Context-aware Activity Notification Systems (CANS) (Laffey, Amelung, & Goggins, 2009). Such dashboard tools analyze student interaction patterns in order to display dashboards of performance and identify students at-risk of failing.

However, EDM, as an emerging discipline, has borrowed techniques from other fields (Long & Siemens, 2011), and as such still lacks a standardized set of tools, models, and processes for analyzing these large dataset (Macfadyen & Dawson, 2012; Romero & Ventura, 2010). In their review of EDM studies conducted between 1995 and 2005, Romero and Ventura (2007) identified a lack of standards for how usage data is preprocessed, modeled, and post-processed. As such, to help

to LMS data sets for many purposes. For example, EDM results have helped identify at-risk learners, examine the relationship between student performance characteristics (e.g., major, year of study) and learning outcomes, monitor student persistence, and inform instructional design (Angeli & Valanides, 2013; He, 2013; Krumm et al., 2014; Picciano, 2014; Siemens, 2013; Long & Siemens, 2011; Xu & Recker, 2011;).

In terms of EDM methods, five methods are widely used in educational research: prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models (Baker & Yacef, 2009; Bienkowski, Feng, & Means, 2012). In our review of EDM studies using LMS data, we found that prediction, clustering, and distillation of data for human judgment were the most commonly used. These are described next.

Prediction refers to developing a model that can predict a response (or outcome) variable, such as student performance, from some combination of predictor variables (Baker & Siemens, 2014). Many studies have used prediction methods in order to find significant predictors of student final grades. In our review of studies, several LMS variables were found to be significant predictors of student final grade including the number of discussion messages posted, the number of assignments completed, the number of quizzes taken (Macfadyen & Dawson, 2010; Thakur, Olama, McNair, Sukumar, & Studham, 2014), the number of interactions with peers, the number of file downloads, regularity in learning intervals (Yu & Jo, 2014), and themes in questions asked online (Abdous, He, & Yen, 2012). Other studies found that variables related to login frequency and the total amount of time online did not significantly predict student final grade (Jo, Kim, & Yoon, 2015; Macfadyen & Dawson, 2010).

Second, clustering is a method to find data points that naturally group a full dataset into smaller subsets (Bienkowski et al., 2012). The grouping object for clustering can be students, courses, or content. Our review found that most studies used student LMS usage data to group students (e.g., Hung, Hsu, & Rice, 2012; Lust, Elen, & Clarebout, 2013; Romero, Ventura, & García, 2008), whereas a few studies used the course or course content as the object of clustering (e.g., Abdous et al., 2012; Valsamidis, Kontogiannis, Kazanidis, Theodosiou, & Karakos, 2012). In terms of clustering algorithm, K-means was the most commonly used in studies using LMS data (Lust et al., 2013; Romero et al., 2008; Valsamidis et al., 2012).

Finally, distillation of data for judgment refers to depicting data, including data visualization, to enable humans to quickly identify and understand its features. At a micro level, visualization methods such as heatmaps, graphics, and scatter plots, can be used (Baker & Siemens, 2014). In a macro level, EDM results have been used to develop student monitoring and tracking dashboard systems, such as the Course Signals System (Arnold & Pistilli, 2012) and the Context-aware Activity Notification Systems (CANS) (Laffey, Amelung, & Goggins, 2009). Such dashboard tools analyze student interaction patterns in order to display dashboards of performance and identify students at-risk of failing.

However, EDM, as an emerging discipline, has borrowed techniques from other fields (Long & Siemens, 2011), and as such still lacks a standardized set of tools, models, and processes for analyzing these large dataset (Macfadyen & Dawson, 2012; Romero & Ventura, 2010). In their review of EDM studies conducted between 1995 and 2005, Romero and Ventura (2007) identified a lack of standards for how usage data is preprocessed, modeled, and post-processed. As such, to help

included in the dataset, before and after data cleaning.

Canvas data consists of a record of user activity along a number of features, including both instructors' and students' views of and participation with features such as assignments, quizzes, conferences, discussions, etc. Table 2 shows the Canvas features (or variables) contained in the dataset. These have been categorized into the four general LMS usage categories as defined by Dawson (2008).

Table 1. Canvas dataset from Spring 2014, before and after data cleaning.

Course modality	Face-to-face			Broadcast			Online			Total	% decrease
Course level	Under graduate	Mixed	Graduate	Under graduate	Mixed	Graduate	Under graduate	Mixed	Graduate		
Before data cleaning											
# courses	1,446	178	198	256	23	52	238	26	44	2,461	
#instructors	1,704	239	252	282	25	79	283	31	53	2,948	
# students	56,734	3,577	3,047	7,234	498	1,244	9,608	1,245	675	83,862	
After data cleaning											
# courses	1,178	124	118	153	18	44	185	18	32	1,870	-24.01%
#instructors	1,383	156	150	174	20	71	215	22	38	2,229	-24.39%
# students	47,301	2,836	2,167	3,882	405	1,143	7,587	1,099	521	66,941	-20.18%

Table 2. Canvas features (variables) logged by source and grouped by instructional category.

	Variables	***Definitions***
Administration	announcements_v	# of visits to announcement page (navigation page for all announcements)
	roster_v	# of visits to roster page (navigation page for people enrolled in the course)
	enrollment_v	# of times enrollment viewed (information for a specific person on the roster)
	calendar_v	# of times calendar viewed
Assessment	assignment_v	# of times assignment viewed (viewing instructions or reviewing instructions after submission)
	assignment_p	# of times assignment participated (submission or resubmission)
	quiz_v	# of times quiz viewed (viewing instructions or viewing previous attempts)
	quiz_p	# of times quiz participated (submission of quizzes)
	grades_v	# of visits to grades page (a student's grade page for a course)
	files_v	# of visits to files page (course navigation page for all files)
Content	attachment_v	# of times attachment viewed (downloading or previewing files)
	syllabus_v	# of visits to syllabus page
	topics_v	# of visits to topics page (course navigation page for all discussion topics)
Engagement	discussion_v	# of times discussion viewed
	discussion_p	# of times discussion participated (making comments or reply)
	wiki_v	# of times wiki viewed (viewing or reloading edits)
	wiki_p	# of times wiki edited and saved
	collaboration_p	# of entering into collaboration
	conferences_v	# of visits to conference page (navigation page for all web conferences)

4. Results

The underlying research methodology for our research program is aligned with the three KDD phases, as described above. In this next section, we address the first research question by describing our processes and modeling approaches within each KDD phase. We then address the second research question by documenting the challenges and lessons learned within each KDD phase.

4.1. Data Preprocessing

4.1.1. Data cleaning approaches. As a first step, AS merged student usage data, course data, and grade data from the various campus enterprise systems. Merging these data resulted in many null values and we thus needed to differentiate between meaningful nulls (i.e., when the activity was not possible due to course design) and accurate nulls (i.e., when the feature was present but not used).

We also eliminated courses for a variety of reasons, including courses with no meaningful usage data, no final grades, with fewer than 5 students, as well as courses with low (fewer than 10) instructor/content or student/content interactions. Ultimately, a total of 1,870 courses were included for the analysis. Figure 1 summarizes the data cleaning processes. Table 1 shows the number of courses, students, and instructors enrolled in each course offered in three modalities (face-to-face, online, broadcast) during Spring semester, 2014, and the raw and percent reduction in data as a result of the data cleaning process. MySQL scripts were developed to conduct the data cleaning process.

Then, we converted the data into a suitable format for applying data mining methods. We extracted only relevant variables for the analyses from the full dataset (through queries to the SQL server) and exported the data in CSV format, which is compatible with analytical tools such as SPSS, R, and Tableau. Lastly we created a summarization table (matrix) by displaying each user in a row and variables in columns.

4.1.2. Variable selection and transformation approaches. Since we used three different methods (prediction, clustering using the EM algorithm, and hierarchical clustering analysis), different variable selection and transformation approaches were used for each method. For instance, we used Pearson's Correlation to detect variables that were highly correlated with outcome variables, and to remove or combine predictors highly correlated with each other (multicollinearity).

In addition, we found that many variables had skewed distributions, in particular student final grade (over 30% of final grades were As, with the proportion decreasing for lower grades), thus violating the basic assumptions of many parametric statistics. Thus, we used

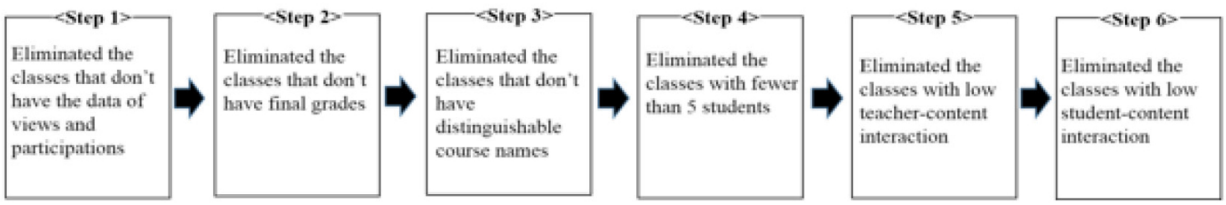

Figure 1. The data cleaning process.

various transformation strategies, such as converting to Z-scores or grouping student final grades into grade bands depending on the purpose of analysis. The details of transformation approaches for each data mining method are described in the next section.

4.2. Data mining, Model building, and Model selection

As described above, educational data mining modeling approaches have different objectives (Bienkowski et al., 2012). In this research, our modeling approach was aimed at examining the extent that Canvas feature use was predictive of student final grades. As described in this section, we used a combination of statistical and machine learning prediction methods, including multinomial logistic regression, clustering using the Expectation-Maximization (EM) clustering algorithm, and clustering using Hierarchical Clustering Analysis (HCA). Various analytical tools were used, including SPSS, Weka, Tableau, and R studio.

4.2.1. Descriptive statistics. Before engaging in modeling activities, we examined descriptive statistics to get a sense of the larger dataset. Figure 2 shows the proportion of features used in courses offered in different modalities. We computed the proportion by dividing the number of courses that used a particular feature by the total number of courses. The figure suggests that course modality may influence feature use, and that certain features are used more heavily than others. This observation informed subsequent analyses described below.

4.2.2. Prediction of student final grade. As noted, we were interested in examining to what extent usage of various Canvas feature influenced student final grade. We first considered the use of Hierarchical Linear Modeling (HLM) given the nested nature of the data, however the data violated the independence assumption as students may be simultaneously enrolled in multiple courses. We then considered multiple linear regression but the data failed to meet assumptions due to skewed distributions that were still present even after performing suggested transformations. Next, ordinal logistic regression was considered and rejected because regression coefficients were not similar across the dependent variable, the

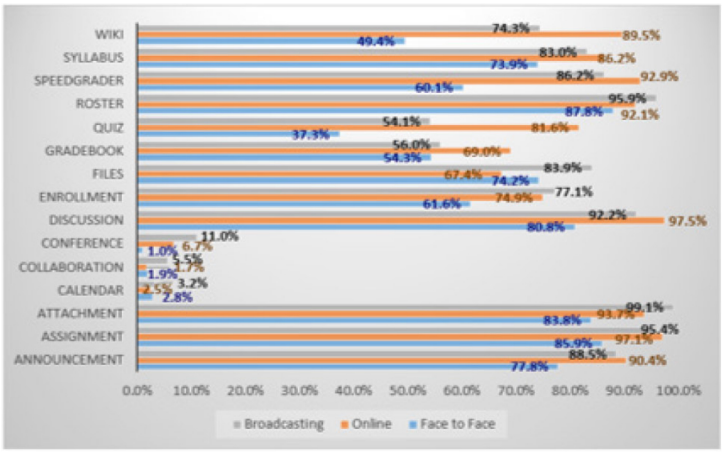

Figure 2. Instructors' usage rate of Canvas features by course modalities.

student final grade.

We thus selected multinomial logistic regression to examine the influence of various LMS features on final grades. In particular, without the assumption of normality, linearity, or homoscedasticity, multinomial logistic regression offers a way to predict the probability of students' membership in one or the other of the final grade categories, based on their use of LMS features. In our dataset, the dependent variable – student final grade – consisted of 11 grade categories, meaning that regression coefficients needed to be interpreted for all variables in 10 comparisons, making interpretation complex. In addition, as noted above, the distribution of final grades was highly skewed toward higher grades. Therefore, in order to equalize the number of final grade observations and to simplify interpretation, we categorized the final grades into four bands: highest (A), high (A-, B+, and B), low (B-, C+, C, C-, D+, D), and lowest (F).

We then transformed raw frequencies of Canvas feature counts into a proportion of total possible activity in order to account for courses with different levels of activity. The proportion was calculated by dividing the number of student views and participation by the total number of content pages posted by the instructors. Next, in order to identify the most meaningful independent variables, we excluded variables which contained too many zeros and nulls since they offered no meaningful data. To address model fit, a series of predictive models were tested using backward elimination, and the final model was selected based on significance of the independent variables. Using SPSS, we conducted two multinomial logistic regressions on final grade (in 4 grade bands) for face-to-face and online courses, respectively.

4.2.3. Clustering at the macro level. In order to identify groups of courses in which instructors and students exhibited similar online patterns, we clustered courses based on instructor and student use of Canvas features. We extracted the student Canvas data (N = 15,255) from 1,040 undergraduate, face-to-face classes.

First, we extracted 7 instructor and 18 student variables (Canvas features) from each course based on Pearson's Correlation, removing features with a value over 0.7. Then, we applied the Expectation Maximization (EM) clustering algorithm using Weka (Ferguson et al., 2006). Advantages of using the EM clustering algorithm are that internally it determines the best number of clusters and also outputs clusters using cross validation. The EM algorithm determines the number of clusters using the following steps:

1. Set the number of clusters to 1;
2. Split the training set randomly into 10 folds;
3. Run EM 10 times with the 10 folds cross validation;
4. Calculate the average log-likelihood over all 10 results;
5. Increase the number of clusters by 1 and repeat from step 2 when the log-likelihood value increases.

When the number of clusters was 3, clusters achieved the highest log likelihood. Thus, the EM clustering algorithm internally stopped when the number of clusters was 3. We also applied the K-means clustering algorithm to validate the EM results. The K-means clustering algorithm also found 3 as an optimal number of clusters. In section 4.3.2, we analyze these three clusters of courses in terms of their relationship with student learning outcomes (i.e., final grade).

4.2.4. Clustering at the micro level. In order to investigate student patterns of activities at the micro level and their relationship with final grade, we selected a course offered by the same instructor in both face-to-face (N = 33) and online (N = 36) formats. As these courses were taught by the same instructor and had similar enrollments, it became easier to compare different course modalities.

To visually analyze student LMS usage patterns (clusters) within course modality, we built a clustergram (Bowers, 2010), which combines hierarchical cluster analysis (HCA) with a heatmap using R studio with the "ComplexHeatmap" and "GetopotLong" packages. The heatmap represents each participant's row of data across each of the columns of variables as a color block, ranging from colder blue for -3 SD below the mean to a hotter red for value +3 SD above the mean, with zero values in white. As such, the heatmap, as a form of visual analytics, enables the human eye to examine the different intensities in patterns across the entire dataset.

Following the recommendations of the HCA literature, in order to standardize variance, all student data were transformed to Z-scores (Bowers, 2010). HCA was applied to cluster both rows (students) and columns (Canvas features). In the present study, we used Euclidean distance measure, which is the most commonly used type when analyzing ratio or interval scale data. For the clustering algorithm, we chose average linkage, which defines the distance between two clusters based on "the average distance between all pairs of the two clusters' members" (Mooi & Sarstedt, 2011, p. 250).

In clustergrams, the rows represent data for each student, and the HCA reordered students in terms of the similarity of their LMS usage patterns. The columns represent the Canvas features and the HCA clustered LMS features in terms of their similarity. The final column shows student final grades (not included in the HCA calculations) to help visualize how usage patterns relate to a student's final grade. Student final grades were coded into 5 grade categories (A, A-, B, C/D, F) to reduce the complexity of interpretation.

4.3. Data Evaluation, Interpretation, and Presentation

In the final phase of KDD, models and data are visualized and interpreted. Here, we discuss the results from the multinomial logistic regression, and cluster analysis at the macro and micro levels.

4.3.1. Prediction of student final grade. The first multinomial logistic regression model (see Table 3) used the data from face-to-face students (N = 19,162), with the "Lowest" grade category as the reference group. The results showed that nine LMS features were significant predictors of student final grades. As shown in Table 3, the "assignment participated" variable had the largest odds ratio: after holding other variables constant, with each unit increase in the "assignment participated" variable, the multinomial odds ratio for a student in the "Highest" grade category relative to the "Lowest" category would be expected to increase by 589%, while the multinomial odds ratio for a student in the "High" grade category relative to the "Lowest" category would be expected to increase by 248%. However, for this variable, the differences between the "Low" grade category and the "Lowest" grade category were not statistically different.

The second multinomial logistic regression model (see Table 4) was based on data from online students (N = 5,279), again using the "Lowest" grade band as the reference category. The results showed that nine LMS features were significant predictors

Table 3. Results of multinomial logistic regression in the face-to-face courses with "Lowest" as the reference group.

Proportional activity (v= views, p= participation)	Highest vs. Lowest (comparison 1)			High vs. Lowest (comparison 2)			Low vs. Lowest (comparison 3)		
	OR	SE	*P*	OR	SE	*P*	OR	SE	*p*
announcements_v	.98*	.00	.00	.99*	.00	.00	1.01	0.00	.05
assignment_v	1.15*	.03	.00	1.08*	.03	.01	1.09*	0.04	.01
assignment_p	6.89*	.16	.00	3.48*	.16	.00	0.83	0.20	.37
grade_v	1.01*	.00	.00	1.01*	.00	.00	1.01*	0.00	.00
quiz_p	1.55*	.03	.00	1.47*	.05	.00	1.41*	0.05	.00
attachment_v	1.08*	.02	.00	1.07*	.02	.00	1.01	0.02	.70
syllabus_v	1.30*	.04	.00	1.23*	.04	.00	0.98	0.05	.73
discussion_v	.55*	.14	.00	.58*	.14	.00	0.79	0.18	.20
discussion_p	.98*	.00	.00	.99*	.00	.00	1.01	0.00	.05
N	19,162								
X^2	1680.806								
Prediction	39.8%								

*$p < .05$

Table 4. Results of multinomial logistic regression in the online course with "Lowest" as the reference group.

Proportional activity (v= views, p= participation)	Highest vs. Lowest			High vs. Lowest			Low vs. Lowest		
	OR	SE	*p*	OR	SE	*p*	OR	SE	*p*
announcemnts_v	1.00	.01	.46	1.01	.01	.22	1.01	.01	.29
assignment_v	0.94	.04	.12	1.07	.04	.10	1.10*	.05	.05
assignment_p	6.12*	.25	.00	2.09*	.24	.00	1.12	.30	.71
grade_v	1.01*	.00	.00	1.01*	.00	.00	1.00	.00	.14
quiz_v	.84*	.05	.00	.91	.05	.08	.94	.06	.31
quiz_p	8.22*	.24	.00	4.89*	.24	.00	3.32*	.29	.00
attachment_v	.99*	.00	.00	.99*	.00	.00	.98*	.01	.00
syllabus_v	1.00	.01	.48	1.01	.01	.12	1.01	.01	.48
discussion_v	1.38*	.06	.00	1.10	.06	.09	1.01	.07	.89
discussion_p	.83	.16	.22	1.10	.16	.56	1.60*	.19	.01
wiki_v	1.18*	.02	.00	1.11*	.02	.00	1.03	.02	.28
N	5,279								
X^2	797.971								
Prediction	47.1%								

of student final grades. In particular, the "quiz participated" variable had the largest odds ratio. After holding other variables constant, for each unit increase in "quiz participated", the multinomial odds ratio relative to the "Lowest" category increased as follows: 722% among students in the "highest" category, 389% among students in the "high" category, and 232% among students in the "Low" category. In sum, the multinomial logistic regression results suggested that the "assignment participated" variable was the strongest predictor in face-to-face classes whereas the "quiz participated" variable was

the strongest predictor in online classes.

4.3.2. Clustering at the macro level. As explained above, the EM algorithm output 3 clusters based on both instructor and student Canvas feature use. Each sub-figure in Figure 3 shows the distribution of feature use for 4 instructor and 4 student features (out of a total of 7 instructor and 18 student features) in each of the three clusters. The red line indicates the median value for courses in each cluster. A greater dispersion of the blue color indicates greater use of that feature in that cluster.

By examining the median values in a cluster within each sub-figure in Figure 3, we noted that instructors in cluster B and C were more active than instructors in cluster A in terms of posting assignments, quizzes, discussion pages, and wiki pages. When we examined student activities within each cluster, we found that students in cluster B were the most active users in terms of participation with assignments, quizzes, discussions, and wikis.

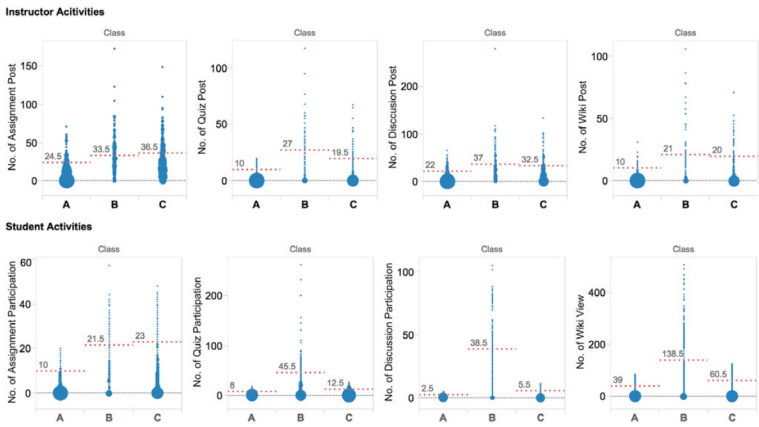

Figure 3. Distribution of use for 4 instructor and 4 student features within the 3 clusters identified by the EM clustering algorithm at the macro level.

Table 5. Average/mean values final grades in each cluster.

	Cluster A	***Cluster B***	***Cluster c***
# of students (%)	14,459 (43%)	5,120 (15%)	13,924 (42%)
Average	3.05	3.21	3.12
Median	3.33	3.67	3.67

**Maximum values shown in bold*

We then examined how instructor and student activities were related to student final grades. We first measured the average and median values of student final grades in each cluster (see Table 5). Students in cluster B achieved the highest average and median grades, while students in cluster C outperformed students in cluster A. The clustering results thus suggest that instructors and students with higher levels of Canvas activity also had higher student final grades.

4.3.3. Clustering at the micro level. Figure 4 presents the clustergram for the face-to-face course (left) and the online course (right). In the face-to-face course, in terms of students (rows), the figure reveals that student clusters with "hotter" colors (higher LMS usage) tended to receive higher final grades, whereas student clusters with "colder" colors (lower LMS usage) tended to have lower final grades. In terms of Canvas features (columns), several features were clustered together that followed the taxonomy shown in Table 2 (e.g., "quiz viewed' and "quiz participated" clustered; and "" files viewed" and "attachment viewed" both clustered together). In addition, patterns in the online course were similar to the face-to-face course in that students' LMS usage aligned with their final grades.

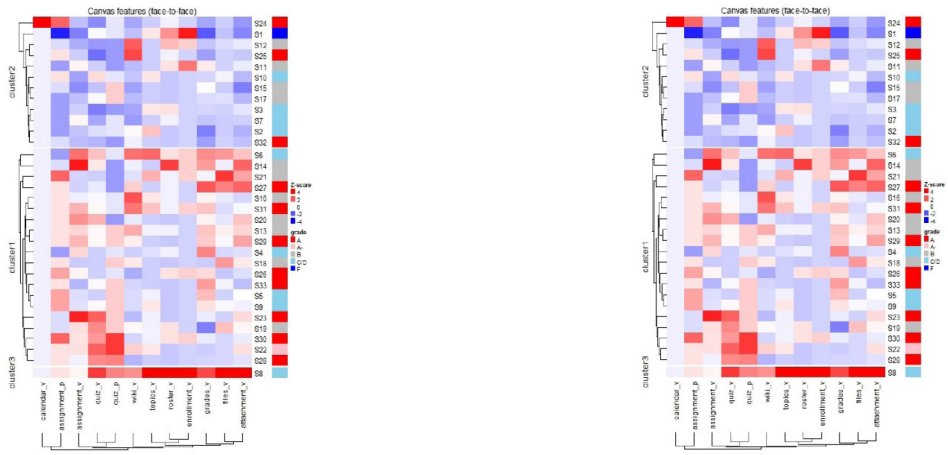

Figure 4. A clustergram of the face-to-face course (left) and a clustergram of the online course (right).

Table 6. Descriptive statistics of final grades in each cluster.

Course modalities	*Cluster*	*N*	*Mean*	*SD*	*Min.*	*Max.*
Face-to-face	1	20	3.21	.75	2.00	4.00
	2	12	2.58	1.24	.00	4.00
	3	1	1.67	-	-	-
Online	1	6	2.66	.84	1.67	4.00
	2	15	3.31	.67	1.67	4.00
	3	15	1.93	1.63	0.00	4.00

For a closer interpretation of the clustergrams, we divided the rows (students) into three overall clusters through visual inspection, and compared final grades in each cluster for both courses (see Table 6). In the face-to-face course, the mean value for final grades in cluster 1 (M = 3.21, SD =.75) with hotter colors was higher than for cluster 2 (M = 2.58, SD = 1.24) with colder colors. Similarly, in the online course, the mean value for final grades in cluster 2 (M = 3.31, SD = .67) was higher than for cluster 3 (M = 1.93, SD = 1.63), but this time the difference was statistically significant ($U = 52.00$, $p < .05$). Thus, student clusters within both courses appear related to their final grades, something that has been noted in the past HCA heatmap literature (Bowers, 2010).

We also found several differences between the two course modalities. First, although the same instructor designed both courses, different LMS features were used in each course. For example, the quiz feature was used only in the face-to-face course, while announcements, syllabus, and discussion tool features were used only in the online course. Secondly, we found differences in the relationship between student final grades and LMS usage. In the face-to-face course, assessment features (specifically the "assignment participated" and "quiz viewed" variables) showed similar color patterns as students' final grade. In the online course, the "wiki viewed' and "grades viewed' variables had similar color patterns as the final grade.

4.4. Challenges and lessons learned.

The previous sections addressed the first research question by presenting results from applying different data mining techniques and examining their relationships with student learning outcomes. This section addresses the second research question: challenges and lessons learned encountered in each KDD phase.

First, in data preprocessing phase, transforming from long format (how data is logged by the LMS) to wide (how data is analyzed in conventional statistical packages) format generated a number of missing or NULL variables, as discussed above. We also struggled with listwise case deletion, and either used techniques that did not require listwise deletion or removed variables with small amounts of data. Finally, skewness is often an issue that is present in log data (Recker & Pitkow, 1996).

In the data mining phase, the first challenge was addressing the skew present in the outcome variable (final grade), as well as the large number of possible grade bands. This was addressed by collapsing final grades into a smaller number of categories, or converting to z-scores. In addition, using clustering as a modeling approach required choosing the (sometimes arbitrary) number of clusters. This choice was validated by using two different clustering algorithms.

Finally, in the data post-processing phase, we found that it was important to consider analyses at multiple levels of analysis and to use multiple data mining methodologies. While more labor and computationally intensive, results can be triangulated in order to gain insights about patterns of usage at the individual student as well as the course level.

5. Conclusion

To address the first research question, we applied three data mining methods, prediction (multinomial logistic regression), clustering at a macro level (EM algorithm), and clustering at a micro level (HCA with heatmap). The result of prediction method

showed that the "assignment participated" variable was the strongest predictor of student outcomes in face-to-face classes, while the "quiz participated" variable was the strongest predictor in online classes. These findings correspond with previous studies, which found that the "number of assignments completed" and the "number of quizzes taken" were significant predictors of student final grade (Macfadyen & Dawson, 2010; Thakur et al., 2014). Second, we used clustering algorithms to group both instructors and students based on their usage patterns. We found that both student and instructor LMS usage patterns were associated with student final grade. Lastly, we applied HCA and created clustergrams to investigate student LMS usage patterns at a micro level. We found that the clustergrams provided a rich contextual portrait of individual students' interaction patterns and how they relate to other students and learning outcome, rather than simply focusing on overall group averages.

Our second research question examined challenges and lessons learned. Similar to other researchers (e.g., Baker & Siemens, 2014), we found that a large proportion of our efforts were devoted to data preprocessing, especially data cleaning. We also found that the application of KDD process was iterative, as we moved from simple to more complex analyses (Blikstein et al., 2014). For example, we often revisited data cleaning when we were unable to perform certain analysis techniques because of faulty data assumptions. Some activities (such as data transformation) were performed during both data preprocessing and modeling. Lines were further blurred with clustergrams, which represent both an analysis technique as well as a rich means of model presentation. We also illustrated how our modeling approaches drew from both statistics and machine learning, and focused on different levels of analysis.

Table 7 summarizes the EDM methods, challenges, and lessons learned as we used the KDD framework to analyze this dataset.

Table 6. Descriptive statistics of final grades in each cluster.

KDD phase	***EDM methods***	***Challenges***	***Lessons learned***
Data processing	- Data cleaning - Variable selecting and transforming	- Highly skewed variables - Listwise case deletion - Long to wide format - Null variables	- Grouping outcome variables into bands - Triangulating multiple techniques
	- Predicting	- Selecting appropriate analysis method	- Grouping outcome variables into bands - Transforming frequencies into proportions
Data mining, model building and selection	Clustering: - EM Clustering (macro level) - HCA (micro level)	-Selecting appropriate number of clusters - Differences in ranges of values across features	- Converting raw scores to z-scores - Using different clustering algorithms
Data evaluating, interpreting, and presenting	- Model fit - Heatmap visualization	- Interpreting model - Interpreting visualization	- Conducting analyses at multiple levels

We note that what data is captured by software tools is generally driven by the underlying technology rather than by educational questions. As such, "big data" and accompanying EDM methods are only useful to the extent that interesting and important educational questions can be addressed (Blikstein et al., 2014; Gašević, Dawson, & Siemens, 2015; Macfadyen & Dawson, 2010). More data is not necessarily better data, and while algorithms can do heavy computational work, researchers need to have a clear vision of the different questions or problems they wish to examine with their data to help inform model building and selection, be willing to go back to prior KDD phases when necessary, and pay close attention to the assumptions of various analysis techniques. Insights from this research contribute to advancing the field of EDM by examining several analysis techniques as approaches for understanding and modeling the increasing and voluminous amount of LMS usage data collected in higher education settings.

References

Abdous, M., He, W., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. Journal of Educational Technology & Society, 15(3), 77–88.

Angeli, C., & Valanides, N. (2013). Technology mapping: An approach for developing technological pedagogical content knowledge. Journal of Educational Computing Research, 48(2), 199–221.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12), Vancouver, BC, Canada, 267–270. doi:10.1145/2330601.2330666

Baker, R., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (2nd ed.; pp. 253–274). New York, NY: Cambridge University Press.

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–16.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Washington, DC: U.S. Department of Education, Office of Educational Technology. Retrieved from https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., & Koller, D. (2014). Programming pluralism: Using learning analytics to detect patterns in the learning of computer programming. Journal of the Learning Sciences, 23(4), 561–599.

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. Practical Assessment Research & Evaluation, 15(7), 1–18.

Dahlstrom, E., Brooks, D. C., & Bichsel, J. (2014). The current ecosystem of learning management systems in higher education: Student, faculty, and IT perspectives. Louisville, CO: ECAR. Retrieved from https://net.educause.edu/ir/library/pdf/ers1414.pdf

Dawson, S. (2008). A study of the relationship between student social networks and sense of community. Journal of Educational Technology & Society, 11(3), 224–238.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37–54.

Ferguson, K., Arroyo, I., Mahadevan, S., Woolf, B., & Barto, A. (2006). Improving intelligent tutoring systems: Using expectation maximization to learn student skill levels. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), Lecture Notes in Computer Science: Vol. 4053. Intelligent tutoring systems (pp. 453–462).

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. TechTrends, 59(1), 64–71.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Waltham, MA: Morgan Kaufmann.

Hawkins, B. L., & Rudy, J. A. (2007). EDUCAUSE Core data service: Fiscal year 2006 summary report. Boulder, CO: EDUCAUSE. Retrieved from https://net.educause.edu/ir/library/pdf/pub8004.pdf

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior. 29(1), 90–102.

Hung, J.-L., Hsu, Y.-C., & Rice, K. (2012). Integrating data mining in program evaluation of K-12 online education. Journal of Educational Technology & Society, 15(3), 27–41.

Jo, I.-H., Kim, D., & Yoon, M. (2015). Constructing proxy variables to measure adult learners' time management strategies in LMS. Journal of Educational Technology & Society, 18(3), 214–225.

Krumm, A. E., Waddington, R. J., Teasley, S. D., & Lonn, S. (2014). A learning management system-based early warning system for academic advising in undergraduate engineering. In J. A. Larusson & B. White (Eds.), Learning analytics: From research to practice (pp. 103–119). New York, NY: Springer.

Laffey, J., Amelung, C., & Goggins, S. (2009). A context awareness system for online learning: Design based research. International Journal on E-Learning, 8(3), 313–330.

Long, G., & & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. Educause Review, 46(5), 30–32.

Lust, G., Elen, J., & Clarebout, G. (2013). Regulation of tool-use within a blended course: Student differences and performance effects. Computers & Education, 60(1), 385–395.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers & Education, 54(2), 588–599.

Macfadyen, L. P., & Dawson, S. (2012). Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan. Journal of Educational Technology & Society, 15(3), 149–163.

Mooi, E., & Sarstedt, M. (2011). Cluster

analysis. In A concise guide to market research: The process, data, and methods using IBM SPSS statistics (pp. 237–284). Heidelberg, Germany: Springer.

National Center for Education Statistics. (2015). Digest of education statistics: 2013. Retrieved from http://nces.ed.gov/programs/digest/d13

Picciano, A. G. (2014). Big data and learning analytics in blended learning environments: Benefits and concerns. International Journal of Artificial Intelligence and Interactive Multimedia, 2(7), 35–43.

Recker, M., & Pitkow, J. E. (1996). Predicting document access in large multimedia repositories. ACM Transactions on Computer-Human Interaction (TOCHI), 3(4), 352–375.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135–146.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 40(6), 601–618.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), 368–384.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. American Behavioral Scientist, 57(10), 1380–1400. doi:10.1177/0002764213498851

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12), Vancouver, BC, Canada, 252–254. doi:10.1145/2330601.2330661.

Smith, V. C., Lange, A., & Huston, D. R. (2012). Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. Journal of Asynchronous Learning Networks, 16(3), 51–61.

Thakur, G. S., Olama, M. M., McNair, A. W., Sukumar, S. R., & Studham, S. (2014). Towards adaptive educational assessments: Predicting student performance using temporal stability and data analytics in learning management systems. Proceedings of the 20th ACM SIGKDD conference on knowledge discovery and data mining (ACCESS '14), New York, NY, USA.

Valsamidis, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., & Karakos, A. (2012). A clustering methodology of web log data for learning management systems. Journal of Educational Technology & Society, 15(2), 154–167.

Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Xu, B., & Recker, M. (2011). Understanding teacher users of a digital library service: A clustering approach. Journal of Educational Data Mining, 3(1), 1–28.

Yu, T., & Jo, I.-H. (2014). Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education. Proceedings of the 4th International Conference on Learning Analytics and Knowledge (LAK '14), Indianapolis, IN, USA, 269–270. doi:10.1145/2567574.2567594

Contact the Author

Ji Eun Lee
Utah State University
Email: jieun.lee@aggiemail.usu.edu

Mimi M. Recker
Utah State University
Email: mimi.recker@usu.edu

Hongkyu Choi
Utah State University
Email: hongkyu.choi@aggiemail.usu.edu

Won Joon Hong
Utah State University
Email: wonjoon.hong@aggiemail.usu.edu

Nam Ju Kim
Utah State University
Email: namju1001@gmail.com

Kyumin Lee
Utah State University
Email: kyumin.lee@usu.edu

Mason Lefler
Utah State University
Email: masonlefler@hotmail.com

John Louviere
Utah State University
Email: john.louviere@usu.edu

Andrew Walker
Utah State University
Email: andy.walker@usu.edu