A Comparison of Educational Statistics and Data Mining Approaches to Identify Characteristics that Impact Online Learning

L. Dee Miller and Leen-Kiat Soh and Ashok Samal
Department of Computer Science and Engineering
University of Nebraska
Lincoln, NE 68588
{lmille, lksoh, samal}@cse.unl.edu

Kevin Kupzyk and Gwen Nugent
Center for Research on Youth, Family and Schools
University of Nebraska
Lincoln, NE 68588
kevin.kupzyk@unmc.edu, gnugent1@unl.edu

Learning objects (LOs) are important online resources for both learners and instructors and usage for LOs is growing. Automatic LO tracking collects large amounts of metadata about individual students as well as data aggregated across courses, learning objects, and other demographic characteristics (e.g. gender). The challenge becomes identifying which of the many variables derived from tracked data are useful for predicting student learning. This challenge has prompted considerable research in the field of educational data mining and learning analytics. This work advances such research in four ways. First, we bring together two approaches for finding salient variables from separate research areas: hierarchical linear modeling (HLM) from education and Lasso feature selection from computer science. Second, we show that these two approaches have complimentary and synergistic results with some variables considers salient by both and others salient by only one. Third, and most importantly, we demonstrate the benefits of a combined approach that considers a variable salient when either HLM or Lasso consider that variable salient. This combined approach both improves model predictive accuracy and finds additional variables considered salient in previous datasets on student learning. Lastly, we use the results to provide insights into the salient variables to the learning outcome in undergraduate CS education. Overall, this work suggests a combined approach that improves the identification of salient variables in big data and also improves the design of LO tracking systems for learning management systems.

1. INTRODUCTION

In the last 20 years, there has been an explosion of online instructional material of various forms—from simple passive content in the form of web pages to sophisticated learning objects (LOs) that integrate interactive instruction, practice exercises and assessment components that promote active learning. LOs are important resources for both learners and instructors and the quantity of LOs is growing (McGreal 2004; Ochoa and Duval 2009), not only in terms of their numbers and content, but also their usage. The use of LOs has infiltrated college teaching, including the disciplines of engineering and computer science, with research showing that the approach increases achievement and promotes success (Francia 2003; Nugent et al. 2006; Riley et al. 2009; Miller et al. 2011a). Further, these online multimedia resources offer tremendous opportunities for providing instructional support to college students with differing abilities, diverse backgrounds, prior knowledge and attitudes towards the subject matter. They offer another advantage in that their computer-based nature facilitates sophisticated tagging, which can collect and combine information about the learners, their interactions with the LOs, and their learning efficiency from the LOs. Tagging is typically based on learner attributes and the content and pedagogical characteristics that can be identified as associated with or predicting learning, but the capability exists for literally every interaction with the LO (e.g., mouse click) to be tracked and time stamped. What results is a tremendous amount of metadata about individual student usage of LOs as well as data aggregated across courses, learning objects, and other demographic characteristics (e.g. gender). The challenge becomes identifying which of the many variables derived from tracked data ***salient*** for predicting and diagnosing student success or failure and using the results to improve learning.

This paper provides a comparison between two approaches for finding salient variables from separate research areas: hierarchical linear modeling (HLM) from education research and Lasso feature selection (FS) from computer science. We provide a detailed analysis and case study comparing HLM with Lasso. We demonstrate that both approaches ***provide separate, but complimentary insights into identifying the salient learner, content, and instructional variables that impact learning*** that should tracked as part of a computer-based tracking system. The HLM approach is an analytical strategy for finding salient variables, individually, taking into account nested or hierarchical data structures, such as students being nested within classrooms. In contrast, the Lasso approach evaluates the variables together using a shrinkage method which reduces to zero the coefficients for variables which have less impact on learning.

The overall purpose of this paper is to show that these two separate approaches, while effective individually at identifying salient variables, are even more effective when used together in a combined approach. To this end, we start by showing that these two approach produce complementary and synergistic results with both overlap in the variables considered to be salient by both approaches and unique variables considered salient by one approach and not the other. Next, we demonstrate the benefits of a combined approach that considers a variable to be salient when **either** Lasso or HLM considers that variable salient. This combined approach leverages these complementary and synergistic results to (1) improve the prediction accuracy for models based on the salient variables and (2) identify additional variables considered to be salient in previous datasets on student learning from other researchers. Additionally, we use these results to provide recommendations about the unique contributions each can provide in understanding

and predicting student learning in LOs. These results provide insights into salient variables that influence learning from multimedia instruction in undergraduate computer science education.

This work sits squarely in the emerging fields of educational data mining and the related field of learning analytics. (see Romero and Ventura (2010) for a discussion of educational data mining; see Berk (2004) for a discussion of learning analytics.) Educational data mining is concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students and optimize learning. Educational data mining encompasses several research paradigms and analysis tools, including quantitative educational statistics, psychometrics, computational modeling, data mining, and machine learning. More recently the term "learning analytics" has been used to describe techniques for analyzing large longitudinal student databases across a variety of courses with the intent to predict learning performance and improve student success. Such databases can include student information systems, online learning management systems, and social and mobile media, including tweets, blogs, and Facebook postings. Relying on methodologies and analytic tools from the disciplines of education and computer science, our research is aligned with the goals of educational data mining and learning analytics. It represents a critical look at how these analytic tools can help answer critical questions related to online learning such as: "Are static or dynamic variables better predictors of learning?" "What content variables are associated with high and low levels of learning?" and "What content variables are most correlated with student attitudes (self-efficacy, motivation)?".

The rest of this paper is organized as follows. Section 2 of this paper presents related work using the HLM and Lasso approaches as well as background research related to the initial identification of salient variables that impact student learning. Section 3 provides a description of the LO dataset considered in this work including the independent variables (i.e., metadata about individual student usage) and dependent variable (i.e., student learning). Section 4 provides descriptions for HLM and Lasso and discusses the similarities and differences in how both approaches identify salient variables. Section 5 starts with an analysis of the predictive accuracy for Lasso, HLM, and combined approach. This section then provides an extensive analysis to validate the salient variables where salient variables found (in HLM and/or Lasso) are compared against salient variables found in previous datasets on student learning. We conclude by discussing the similarities and differences between the two approaches and the unique contributions of each in both understanding student learning and providing insight into how LOs should be used.

2. Background and Related Work

This section discusses literature related to the two approaches from research in education and computer science. There has been little previous work which compared approaches from both fields on the same dataset. Delen (2009) performed an analysis of cancer data which used logistic regression (from inferential statistics) and data mining techniques including decision trees, neural networks, and support vector machines to develop prediction models. However, we have not found any previous work which specifically compares the HLM and Lasso discussed in this paper. Instead, as discussed below, most research has involved the separate use of these two approaches.

2.1. INFERENTIAL STATISTICS FROM EDUCATION

The statistical approach used in education, and social science research in general, is based on inferential statistics and generalization. Inferential statistics allow researchers to collect and analyze samples of individuals, and then make inferences or generalizations about the entire population (Gravetter and Wallnau 2004). For example, in this study we have a sample of students and we wish to identify salient learner characteristics that can be generalized to the population of students in undergraduate computer science classes. If a characteristic (i.e., a variable) such as the student score on a placement test is found to be associated with assessment score in our sample, we can infer that this variable is salient to the learning outcome in the population.

A key step in inferential statistics is the development of an apriori hypothesis, a conjecture about effects that exist in the population. On the data, the hypothesized relationship is tested using a particular statistic such as a correlation, t-test, or regression, and a probability (p-value) is returned. If the p-value falls below a pre-specified cut-off point then the researcher has determined there is a significant effect in the sample, and thus, in the population. The cut-off p-value, often referred to as alpha (α), that is most often used is 0.05 (Cohen et al. 2003), which is the highest acceptable probability of incorrectly claiming there is an effect when there is not.

Regression is a common statistical modeling technique in inferential statistics that assesses the relationships, or correlations, between one or more predictor variables and an outcome, or dependent variable. The most basic linear regression equation models a continuous dependent variable, Y, using a predictor variable, X, and an error term, or variance component. The basic regression model is as follows:

$$Y = B(X) + e$$

where B is the estimated regression coefficient that minimizes the error term, e, and thus provides the best solution for the regression equation. In this context, the dependent variable is the student's score on the LO assessment. To identify salient predictors we test several different variables. To test whether a variable is a salient predictor of LO scores, the regression model determines whether the regression coefficient is significantly different from zero. If there is no relationship between a predictor and the outcome, B will be near-zero and will be non-significant. Significance is assessed based on the strength of the relationship between X and Y and the amount of error observed. Note that predictor variables may be continuous or categorical.

The basic linear regression model has one variance component, the residual variance (e or σ^2). An important assumption of regression is that residuals, or person-specific prediction errors, are independent across individuals. When data are nested, however, as is the case when multiple LO scores are obtained from each student, the assumption of independent residuals is violated. In order to meet the assumption of independent residuals using basic linear regression, separate analyses for each LO need to be performed. This results in numerous regression models—16, in our case, as there were 16 different LOs—for each predictor, some of which could indicate a predictor is salient, while others do not. Furthermore, because all students do not necessarily complete all LOs, the 16 LO-specific models could involve a slightly different sample each time. A succinct modeling strategy is needed that can utilize all available data and facilitates the overall determination of salient predictors of student performance while accounting for the dependence of residuals within students.

Hierarchical Linear Modeling (HLM; (Raudenbush and Bryk 2002)) accounts for this by adding additional variance components and using the normality of the residuals assumption. In this way, HLM captures variability in outcome scores attributable to nesting effects making the assumption of independence of residuals more tenable (Locker et al. 2007). Some examples of hierarchies are individuals nested within classrooms or repeated measures nested within individuals. Indeed, an important feature of educational data is that they are hierarchical. As such, educational data mining methods must explicitly exploit the multiple levels of meaningful hierarchy in educational data (Baker 2011). HLM is ideally suited to account for this clustering since HLM combines the advantages of (1) mixed model ANOVAs, with modeling of fixed and random effects, and (2) regression, with its ability to deal with variables that are discrete and continuous. Just as students are clustered in classrooms, outcome scores from LOs represent another hierarchical framework because repeated measurements (e.g., up to 16 LO scores) are obtained for each individual. Therefore, LO scores are nested within individuals, forming a hierarchy that causes measurements within an individual to be more alike than between individuals. Thus, cases are only considered independent when individuals have been taken into account.

HLM has become the gold standard statistical method to deal with clustered, hierarchical educational data and is gaining increasing use in many fields. Applications can be found not only in education, but also in counseling psychology and psychotherapy (Kahn 2011), early childhood research (Hindman, Skibbe, Miller, and Zimmerman 2010), gerontology (Terracciano, McCrae, Brant, and Costa 2005), organizational research (Schonfeld and Rindskopf 2007), drug prevention (Seo and Li 2009), medical research (Halkitis, Palamar, and Mukherjee 2008), nutrition (Alvarado, Zunzunegui, Delisle, and Osorno 2005), animal behavior (Hernandez-Lloreda, Colmenares, and Martinez-Arias 2004), small-sample or single-case research (Ferron, Bell, Hess, Rendina-Gobioff, and Hibbard 2009), and large-scale survey research (Stack and Kposowa 2008).

2.2. Feature Selection from Computer Science

There has been a considerable amount of recent work in feature selection (FS) in computer science. First, we discuss FS and summarize the recent work. Then, the remainder of this section summarizes recent work on using FS for educational data mining.

Feature Selection Algorithms. There has been extensive work on FS both in the classification and regression domains. One popular FS approach for the regression domain is Lasso (Tibshirani 1996). There are two main reasons for using Lasso (Hastie et al. 2011). First, selecting the subset of salient variables (i.e., features) improves the prediction accuracy on the dependent variables. Second, filtering out the nonsalient variables also simplifies interpretation of the results by the researchers and domain experts. There are several methods for regression-based FS including the subset and shrinkage methods. The subset method uses a discrete process where each variable is retained or discarded. The shrinkage methods, on the other hand, are more continuous and use a separate coefficient for each variable. Although the subset methods are still quite popular, the Lasso approach in this paper uses the shrinkage method. Again, the Lasso approach allows for a fair comparison with HLM since both maintain a set of regression coefficients for all variables considered. Additionally, shrinkage methods do not suffer as much from high variability in the results (Hastie et al. 2011) from fine-tuning parameters. Interested readers should consult Sayes et al. (2007) for a general survey of FS algorithms.

While there has been little previous work on using Lasso for educational data mining, there has been considerable work using other FS algorithms for educational data mining. Riley et al. (2009) used FS subset algorithms as part of a two-stage approach for mining empirical usage metadata from LOs. Ramaswami and Bhaskaran (2009) used FS to improve the ability to predict the performance of students. They use six different FS subset algorithms to identify combinations of salient variables. These combinations are then used to train four different classification models. The results show that applying the FS algorithms improves the predictive accuracy for all models. However, there was little to no agreement on the combinations of variables and considerable variance in predictive accuracy—more reasons for going with a shrinkage method as used in Lasso. Additionally, Ramaswami and Bhaskaran (2009) focused just on demographic-like variables (e.g., number of siblings, eye vision, etc.) rather than the demographic and session data considered here. McLaren et al. (2010) used FS on the ARGUNAUT computer supported collaborative learning system which provides feedback to the teacher based on student discussions. Such feedback is generated by training classification models based on previously collected session data. Only a single FS algorithm (subset method) was used in some of the experiments but it improved the performance for all models except support vector machines.

2.3. Learner, Content, and Instruction Variables that Impact Learning

The initial list of learner, content, and instructional parameters to be tracked in our system was developed based on learning and instructional theories dealing with computer-based and multimedia instruction (Mayer 2001), as well as past research examining factors believed to function as indicators of success and achievement in introductory computer science courses. For students, math background, previous programming experience, and comfort level appear to be the primary factors identified (Wilson and Shrock 2001; Chen 2002; Wiedenbeck et al. 2004; Bergin et al. 2005; Ventura 2005). In particular, self-efficacy, a learner's confidence that he/she can accomplish a task, and learner motivation variables have also been shown to be important to learning. The self-efficacy construct has been correlated with achievement outcomes (Sorge 2007) and motivation to learn (Pintrich et al. 1993). Self-efficacy appears to become more accurate over the course of a semester (Wiedenbeck et al. 2004). Some students underestimate or overestimate their ability to perform. Their perception becomes more accurate as students learn to evaluate their abilities based on the direct interaction with the task. Motivation is found to impact students' performance as well. Students with high intrinsic motivation perform significantly better in a computers science course than students with low intrinsic motivation (Bergin et al. 2005).

In contrast to learner parameters, instructional parameters have included the LO topic and degree of difficulty (as determined by course instructors/developers). Research has shown that students who perceived the course material as not difficult tend to perform better than their peers who consider the course difficult (Rountree et al. 2002).

Our own research has identified specific variables that predict learning from LOs, including GPA, ACT score, number of previous programming courses, scores on a mandatory placement test, and a positive evaluation of the learning object (Riley et al. 2009; Nugent et al. 2011). Increased LO difficulty, increased time spent on the assessment, and reports of confusing LOs are also negatively associated with student learning and potentially negatively impact student learning. Student gender was not found to be a significant predictor. A summary list of

variables tracked in this study is found in Table 1 and these measures are discussed more fully in Section 3.3.

3. RESEARCH METHODOLOGY

This section describes the research methodology we used to deploy the learning objects (LOs) to the undergraduate computer science (CS) courses and collect the iLOG dataset used in the rest of this paper. We first discuss the characteristics for the students that completed the LOs and the courses where the LOs were used. We then provide a description of the LOs used in this research. Finally, we provide a description of the tracked variables which are included in the iLOG dataset concluding with a brief description of the tracking system (interested readers should refer to Miller et al. (2011c) for more details).

3.1. PARTICIPANTS AND COURSES

There were five undergraduate computer science classes who completed the learning objects. These ranged from an honors class with students in a joint business/CS program; an introductory undergraduate CS course for computer science majors, a CS course targeted for engineering students and using the Matlab language, a course for non-majors which fulfilled the general education requirement, and a basic CS course that focused on the C programming language. In total, 403 students participated in this version of the study (see Miller et al. (2011a) for details of previous studies). Most of the students were majoring in engineering (38%) or had other college majors (44%). Only 17% were computer science majors. Most of the students were freshmen (49%), male (84%), and had GPAs above 3.5 (48%). Of these students, 75% were taking the computer science course as a requirement and most (66%) had not had a previous programming course. While there were clear differences in the types of students in each course (i.e. the honors students had higher GPA and had taken more programming courses), considered together the students represent the range of backgrounds expected in an undergraduate computer science course.

In all courses, the LOs were part of a student's course grade. All the LOs together counted for between 3-5% of the total course grade based on instructor preference. This was done to ensure students had some motivation to use the LOs. The deployment schedule of the LOs was designed to make sure students had taken the LOs before the lecture/labs on the same topic. This was done to ensure student assessment scores reflected understanding of the LO content and not other sources (e.g., lecture on same content).

3.2. DESCRIPTION OF THE LEARNING OBJECTS

The LOs used in this research follow the Sharable Content Object Reference Model (SCORM) standard for web-based e-learning so they are usable on any SCORM-compliant learning management system (LMS) including Blackboard, Moodle, etc. Each of the LOs contains (1) a tutorial, (2) a set of flash/applet interactive exercises, and (3) an assessment. The tutorial starts with a page that lists the objective of the LO, followed by a set of pages that explains the LO content using text and graphics. The information in each tutorial component is succinct on a particular topic—about several pages of a traditional textbook. The tutorial concludes with a summary and hints for reflections. Each LO also has a set of 1-4 exercises based on the content covered in the tutorial. Exercises generally require several steps to arrive at the correct answer.

Students can repeat exercises as many times as desired. The assessment consists of a set of 7-15 questions depending on the length of the tutorial. All the questions are either multiple choice or true/false. These questions are used to measure whether students have learned the content presented in the tutorial and exercises.

We have developed 16 LOs to cover most topics of the ABET[1] approved syllabus for the introductory computer science core course. The LOs cover a comprehensive range of content ranging from basic concepts as arrays, numeric data, and logic to advanced concepts such as searching, sorting, and recursion. The LOs also cover a range of difficulty as verified by results on the assessments and the subjective view of content experts. Furthermore, because of the different programming languages used (e.g., MATLAB, C, and Java) some revision of the LO content was required to accommodate underlying differences in the languages (e.g., Arrays in MATLAB are 1-indexed instead of 0-indexed). In addition to covering different topics, the LOs also represent varied use of multimedia elements such as text, graphics, and animation. Also, the students in these courses included non-majors, CS majors, CS honors students as well as honors students in a special CS-business program. This required carefully balancing the difficulty of the assessment questions to accommodate students with varying aptitude. To this end, these LOs had undergone a rigorous revision process using proven techniques from educational research including Bloom's taxonomy levels, item-total correlation, and Cronbach's Alpha (Miller et al. 2011b).

3.3. Description of the Variables Being Tracked

A major step in the development of the tracking system was determining what parameters would be tracked and the decision was guided by the research and learning theory discussed in Section 2.3. A total of 81 variables were collected online and included in the dataset. A summary list of variables is shown in Table 1. These were the independent variables (i.e., predictors) used in our research.

The outcome variable (dependent variable) for this research was students' learning, measured by scores on the assessment questions for each LO. As part of the development process, the assessments underwent a thorough psychometric analysis, with the goal of insuring that all had Cronbach Alpha's of 0.7 or above and item-total correlations of above 0.3 (Miller et al. 2011b). Further description of the predictor variables and the instruments used to assess these variables is found below.

Static Student Data.

Motivated Strategies for Learning Questionnaire (MSLQ). To measure learner motivation and the use of specific strategies to learn CS content, we used the Motivated Strategies for Learning Questionnaire (Pintrich et al. 1999), which has been widely used in undergraduate education, including computer science (Bergin et al. 2005), both as a research tool, as well as a means to provide feedback to students regarding their study habits, learning skills and motivation. There were five motivation scales: (a) intrinsic goal orientation, which refers to the student's perception of participation in a class for reasons such as challenge, curiosity, and mastery; (b)

[1] ABET accreditation assures that a college or university program meets the quality standards established by the profession (http://www.abet.org/).

extrinsic goal orientation, which concerns the degree to which the student perceives herself to be participating in a class for reasons such as grades, rewards, performance, and evaluation by others; (c) control of learning beliefs, which refers to students' beliefs that their efforts to learn will result in positive outcomes which are contingent on their own effort; (d) self-efficacy; and (e) task value, students' evaluation of how interest, importance and usefulness of tasks. The learning strategies scales included (a) elaboration, which involves building internal connections between material to be learned; (b) organization, which involves selecting appropriate information; (c) metacognitive self-regulation, which refers to the use of such strategies such as self-testing and questioning; (d) effort regulation, which refers to students' ability control their attention in the face of distracting tasks; and (e) help seeking from peers and instructors. In addition to these standard MSLQ scales, a scale specifically dealing with the use of problem solving strategies was added given its relevance to the computer science content area.

Table 1: Variables collected using iLOG. MSLQ Responses are discussed in Section 3.3. Note that (*) indicates the data are not collected until after the LO session.

Static Student Data	**Static LO Data**	**Interaction Data**
MSLQ Responses (12)	Topic	Exercise Feedback (3)
Gender	Difficulty	Tutorial Time and Clicks (10)
Major and Grade Level (2)		Exercise Time and Clicks (18)
GPA and ACT (2)		Assessment Time and Clicks (10)
Previous Courses (4)		LO Time and Navigation (4)
CS Placement		Evaluation Responses (10)*
Courses (2)		
Total: 24	**Total: 2**	**Total: 55**

Demographics. In the beginning of the study, in each course, students were asked to fill out a questionnaire on demographics: gender, major, GPA, grade level (Freshman, Sophomore, Junior, Senior), SAT/ACT score, the highest math course taken so far, whether they have taken a programming course or IT course before. We also include their scores on the departmental computer science placement exam, which evaluates prior knowledge on CS topics.

Static LO Data. The static LO data consists of the LO topics taken from ABET approved syllabus and LO difficulty scores. The LO difficulty scores were computed subjectively by combining the vote of five content experts on a scale from 1-7 with 7 being the most difficult (Miller et al. 2011a).

Interaction Data.

LO Session Data. As student work through the LOs, their interactions are tracked and stored using a software tool called Intelligent Learning Object Guide (iLOG) (Nugent et al. 2009; Riley et al. 2009). Each LO contains a Wrapper that tracks all student interactions in the LO session and uploads them in real-time to an external database. The wrapper tracks not only scores on assessment questions, but also the time students spend on specific content, the steps taken on the practice exercises, etc. The specific student interactions collected using iLOG are summarized in Table 1. The data collected also includes the survey and questionnaire responses.

LO Evaluation Survey. At the end of each LO were a series of evaluation questions intended to solicit student impressions of the LO such as ease of use, interest level, learning value, difficulty, comprehensibility, and overall rating. These questions focused on whether the student felt the LOs improved learning compared to a traditional classroom environment. Responses were collected using a five-point Likert scale of strongly agree, agree, indifferent, disagree, and strongly disagree.

Here we provide a brief overview of how the LO Wrapper works. First, the LO Wrapper uses the Easy Shareable Content Object (SCO) Adapter for SCORM 1.2. The SCO adapter provides a direct interface with the SCORM API the LMS uses for displaying the LO. This connection to the SCORM API updates the LO Wrapper when pages are displayed to the user and also provides information about the assessment component. The LO Wrapper also uses existing web technologies including JavaScript and PHP to create a bridge between the LO and an external database. Using this bridge, the wrapper can transmit user interactions to the database and metadata back to the LO. This bridge requires a connection to the Internet, but this is generally not an issue because such a connection is also required for most LMSs. Figure 1 gives examples for each iLOG component (i.e., tutorial, exercise, and assessment) in terms of the statistics collected and the content presented. After the users finishes the assessment, the LO Wrapper automatically uses the JavaScript/PHP bridge to transmit the user interactions in the JavaScript collections to an external database.

4. ANALYTIC APPROACH

The similarities and differences between the HLM and Lasso approaches are summarized here with additional explanation provided in Sections 4.1-4.2. Both approaches similarly examine all the records in the iLOG dataset and decide for each variable whether that variable is salient. Additionally, both approaches provides a regression coefficient for each variable showing how much an increase of one "unit" (e.g., second, click, etc.) for that variable would change the learning outcome (positively or negatively). These approaches differ principally on how the evaluate the variables. HLM uses the L2 penalty through groups of dummy variables. This allows HLM to support nesting effects when multiple records are related such as multiple records for the same student. To determine the salient variables, the HLM approach uses the probability-value based on the correlation—variables which are significant ($p < 0.05$) are considered to be salient. Lasso uses the L1 penalty with shrinkage methods based on a varying lambda parameter. The means used to set the lambda depends on the Lasso implementation (we use internal cross-validation as described in Section 4.2). In this way, Lasso shrinks the coefficients for nonsalient variables towards zero. Any variable with a non-zero coefficient after Lasso has converged is thus considered to be salient (Hastie et al. 2011). Overall, although

HLM and Lasso are both rooted in regression models, there are key differences between these two approaches.

We chose to use HLM and Lasso for several reasons. First, we considered approaches for finding salient variables favored in different research areas. HLM is much more commonly used in education research, while Lasso is more commonly used in computer science research. Second, we desired approaches with a balance of similarities and differences. The approaches needed to be similar enough to allow for a legitimate comparison of salient variables, while different enough to allow for complementary and synergistic results. Lastly, by considering approaches that are effectively one step removed (e.g., L1 vs. L2 penalty), we can evaluate the effectiveness of a combined approach on the same results. This would not be the case had we used to completely unrelated approaches leading to an "apples and oranges" comparison.

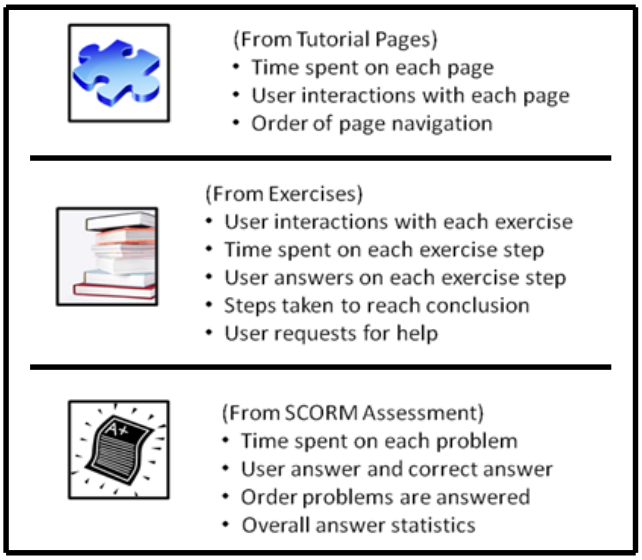

Figure 1: User Interactions Captured Using the LO Wrapper from LO Content.

4.1. HIERARCHICAL LINEAR MODELING APPROACH

The first approach used is Hierarchical Linear Modeling (HLM) which is similar to simple linear regression except that one extra variance component is included in the model to account for hierarchical nesting effects in datasets such as when multiple records are related. The iLOG dataset has such nesting effects since a single student goes through up to 16 different LOs. Thus, in the results reported below, HLM has LOs nested within students.

An example of the HLM equation in our analyses for predicting the LO score using the ACT variable is as follows:

$$LO\ Score_{ij} = B * ACT_i + u_i + e_{ij} \quad (1)$$

where the LO score for individual i and session j is predicted by each student's ACT score and there are two error terms. The u_i term is the average error in prediction for individual i, and e_{ij} is the residual variance for the LO scores after accounting for nesting and the effect of ACT scores. Statistical tests are provided by the software package (SAS and R are used in this study) which assesses the significance of the regression coefficients (B) by providing a probability (p-

value) that the regression coefficient is equal to zero. Small p-values (less than 0.05) indicate that the regression coefficient is not likely to be zero. Thus, it is significantly different from zero, and the predictor variable (ACT) has a significant effect on the outcome. The value of the regression coefficient is interpreted as the change expected in the outcome based on a one-unit increase in the predictor. For example, the regression coefficient for ACT was found to be 1.35, which means that a 1.35% increase in percent correct on the LO can be expected for each one unit increase in the ACT score. In other words, a student with an ACT score of 20 is expected to score 1.35% higher than a student with an ACT score of 19. Although it is not the case with ACT scores, a significant, negative regression coefficient would indicate that as the predictor increases, LO scores tend to decrease.

4.2. LASSO FEATURE SELECTION APPROACH

The second approach used is Lasso which uses regression-based FS to decide which of the variables are most commonly associated with the dependent variable. As previously discussed (Section 2), many of the FS algorithms use a subset method for FS which chooses a subset of the variables as salient and discards the rest. The downside to using a subset method is that it often exhibits high variability in which variables are considered salient (Hastie et al. 2011). Lasso, instead, uses a shrinkage method for FS which maintains a coefficient for each of the variables. Similar to the HLM previously described, these coefficients can be used to measure the "impact" that the variable has on the learning outcome. As alluded to earlier, variables with coefficients that have been shrunk to zero are consider nonsalient, while variables with non-zero coefficients are considered salient.

The actual Lasso approach operates by minimizing the following equation where y_i is the dependent variable (i.e., the learning outcome), x_{ij} is the variable, β_j is the regression coefficient and λ is the tuning parameter (Hastie et al. 2011):

$$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2)$$

Note that the above equation is equivalent to minimizing the sum of squares (as in simple linear regression) subject to the constraint:

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

Lasso has some nice theoretical properties such as the interactions between the error and constraint functions resulting in regression coefficients which converge to zero as opposed to other shrinkage methods (e.g., ridge regression) which may converge to nonzero values. Interested readers should consult Hastie et al. (2011) for further details on Lasso properties.

For this paper, the Lasso approach chooses which variables are salient while minimizing the above equation. Obviously, the tuning parameter has an impact on which variables are considered to be salient. In a sense, Lasso results in multiple set of regression coefficients for

the variables based on different lambda values. We use the Lasso cross-validation[2] technique to choose the optimal lambda value (Simon et al. 2011). In summary, this technique randomly divides all the records into 10 sets. Then, it runs Lasso with 100 different lambda values on the combined records from sets 2-10 and measures the goodness of fit for the salient variables on set 1 used as the validation set. This process is repeated with each set, in turn, as the validation set. The optimal lambda value is the one with the highest goodness of fit averaged across all 10 validation sets.

As a final consideration, the equation for the Lasso approach given above is non-convex and computing the solution originally required solving a (time-intensive) quadratic programming problem. However, recent work has developed efficient algorithms for solving this equation using pairwise coordinate descent (Friedman et al. 2007). In a nutshell, this allows the coefficient for each variable to be updated separately using the following equation where S is the soft-threshold function and $y_i - \tilde{y}_i^j$ is the partial residue on the jth variable.

$$\widetilde{\beta}_j(\lambda) \leftarrow S\left(\sum_{i=1}^{n} x_{ij}\left(y_i - \tilde{y}_i^j\right), \lambda\right) \quad (3)$$

Repeatedly cycling through each variable in turn (until convergence) provides the solution to the Lasso equation.

In the results presented below, for the Lasso approach, we use pairwise coordinate descent algorithm with cross-validation to choose the optimal lambda value.

4.3. Differences Between Approaches

There are two main differences between the HLM and Lasso approaches described above. (We provide further discussion on the similarities and differences between these approaches in Section 6.2.) First, the HLM approach has an additional component (e_{ij} in Eq. 1) to account for nesting effects between similar records in the dataset, while the Lasso approach evaluates each record separately. This allows HLM to evaluate variables with nested data more effectively. Second, the Lasso approach has an additional component ($\lambda \sum_{j=1}^{p} |\beta_j|$ in Eq. 2) that considers regression coefficients for multiple variables at the same time, while the HLM approach evaluates each variable separately. This allows Lasso to evaluate interdependencies between multiple variables more effectively. As will be shown in Section 5, neither approach dominates the other in terms of the salient variables found. Rather, the differences between these approaches allow each approach to find a separate subset of salient variables supporting our notion of using a combined approach to find all the salient variables.

5. Results

As alluded to earlier, the goal of this paper is to demonstrate that a combined approach improves the identification of salient variables in big data compared to using Lasso or HLM alone. Meeting this goal requires two separate steps. First, we demonstrate that the salient variables chosen by our combined approach provide improved predictive accuracy compared to those

[2] This internal cross-validation for Lasso should not be confused with the cross-validation used later in the analysis of predictive performance.

chosen by Lasso or HLM alone. The combined approach uses a more comprehensive set of salient variables since it considers a variable salient when **either** Lasso or HLM consider that variable salient. We show that this more comprehensive set improves model accuracy on new data and not simply the fit on existing data. Second, we validate the salient variables found in our dataset. Obviously, there is no single "ground truth" for salient variables available for our iLOG learning object dataset. However, there is considerable overlap with many of the variables in our dataset also included in previous datasets on student learning from other researchers. We show that our combined approach finds the same salient variables as those found in previous datasets. This supports our claim that the combined approach produces variables that are salient and the results are not simply artifacts of our dataset.

5.1. EXPERIMENTAL SETUP

In the results below, we use Lasso, HLM, and a combined approach to determine the salient variables. First, we use the SAS and R software package for HLM. The HLM results include both the regression coefficient and the p-value for each variable. If the p-value for a variable is less than 0.05, HLM considers that variable to be salient. Second, we use the R glmnet software package for Lasso. The Lasso results include multiple sets of regression coefficients for the variables based on different lambda values. We start with Lasso cross-validation method to choose the optimal value for lambda (Simon et al. 2011). Then, we consider the set of regression coefficients (one for each variable) with the optimal lambda value. If the regression coefficient for a variable is nonzero, Lasso considers that variable to be salient. Finally, the combined approach is run manually using the results from Lasso or HLM. As previously discussed, the combined approach considers a variable to be salient when **either** Lasso or HLM (or both) consider that variable salient.

The predictive accuracy for the salient variables is measured using an intermediate model and repeated cross-validation. The intermediate model is created using **only** the salient variables either from the combined approach, Lasso, or HLM. These variables are normalized to the same scale to alleviate potential problems with the model such as level-2 estimation problems due to multicollinearity (Hofmann and Gavin 1998). To allow for the fairest possible comparison, the same type of intermediate model is used to evaluate all three algorithms. We use HLM as the intermediate model. This supports the nesting used to choose the salient variable in the combined approach and HLM. Since Lasso does not use nesting, we want to be certain that nesting is not corrupting the accuracy. Therefore, we also evaluate the Lasso salient variables using a conventional regression model. Repeated cross-validation starts by breaking all the records in our dataset into separate groups. Cross-validation then follows the iterative process of creating a model using the records in all but one fold and evaluating that model using the remaining fold. Since the model is evaluated on effectively new records, this allows a fair estimate of predictive accuracy. To reduce bias in the estimate, cross-validation repeats this process with each fold used as the evaluation fold and new models created on the remaining folds. The final predictive accuracy resulting from cross-validation is averaged over all the folds. We use repeated cross-validation with 5 folds repeated 10 times as suggested in Alfons (2012).

5.2. PREDICTIVE ACCURACY

Table 2 provides the prediction accuracy for the combined approach, Lasso, and HLM measured using repeated cross-validation. The results provided are the root mean squared prediction error (Alfons 2012). Because these results measure the model error, lower values actually indicate improved predictive accuracy. Based on the results, the combined approach ***provides significantly lower prediction error than Lasso or HLM on all ten cross-validation runs*** ($p = 0.001$ level). The combined approach is larger and more inclusive considering a variable salient when either Lasso or HLM consider that variable salient. This allows the combined approach to include "borderline" salient variables such as those salient in combination or when nesting is considered. Starting with more salient variables *provides a more comprehensive view* on our iLOG dataset allowing models to be created that achieve higher predictive accuracy. At the same time, simply increasing the variables available to the model is not an effective workaround. When we tried to use cross-validation using all the iLOG variables (salient and nonsalient), the model often failed to converge at all. This provides more support for the combined approach *that chooses the "middle ground" between using all the variables and only those deemed salient by a single algorithm*. Overall, the results in Table 2 demonstrate that the combined approach provides improved predictive accuracy compared to those chosen by Lasso and HLM. The rest of Section 5 is devoted to validating the salient variables found in our iLOG dataset.

Table 2: Average prediction error using cross-validation repeated 10 times. The combined approach results in significantly lower prediction error compared to the other approaches (t-test, p=0.001 level). Since Lasso does not consider nesting, results are provided for Lasso with and without nesting.

CV Run	Combined	Lasso		HLM
		w/ Nest	w/o Nest	
1	1.66	1.74	1.74	1.79
2	1.65	1.74	1.75	1.79
3	1.66	1.73	1.74	1.78
4	1.66	1.74	1.75	1.79
5	1.66	1.73	1.74	1.79
6	1.66	1.73	1.74	1.78
7	1.66	1.73	1.74	1.78
8	1.66	1.73	1.74	1.79
9	1.66	1.73	1.74	1.78
10	1.66	1.74	1.75	1.79

5.3. SUMMARY OF SALIENT VARIABLES

In general, in the results below, we divide the salient variables into different groups based on their provenance in the iLOG dataset. This includes salient variables related to (a) LO deployment, (b) student demographics, (c) responses on the MSLQ, (d) LO session data, and (e) responses to the evaluation questions. Of the 81 variables, there are a total of 6 for LO deployment, 10 for demographics, 12 for MSLQ, 43 for LO session data, and 10 for evaluation questions.

Table 3 contains a summary on the agreement/disagreement for the salient variables found by the HLM and Lasso approaches. Both approaches agree on the majority of the variables (49 out of 81) as discussed further in Section 5.4. These results support the effectiveness of individual approaches since the variables considered (in particular the MSLQ) have been extensively studied and previously found to be salient to student learning. Nevertheless, the individual approaches disagree on 32 variables. As discussed in Section 5.5 and Section 5.6 neither approach is able to correctly identify all the MSLQ and demographic variables considered salient in previous work. Further, neither approach is able to identify all the salient LO deployment and session data variables previously reported for the iLOG datasets as well (Nugent et al. 2011; Miller et al. 2011a-c). By leveraging the results from HLM and Lasso together, the combined approach identifies the variables considered salient in previous work. This validation reduces concern that the improved predictive accuracy using the combined approach (Section 5.2) is limited to the iLOG dataset. Note that we address concerns that disagreement between approaches may be the result of redundant variables in our correlation analysis in Section 5.7. This analysis shows pairs of variables on which HLM and Lasso disagree can be highly correlated, but are far from redundant.

Table 3: Confusion matrix of variables identified as salient (Yes) or not (No) by HLM and Lasso approaches.

Confusion Matrix		**HLM-Identified**		
		No	Yes	**Total**
Lasso-Identified	No	30	20	**50**
	Yes	12	19	**31**
	Total	**42**	**39**	**81**

5.4. VARIABLES FOUND SALIENT/NONSALIENT BY BOTH APPROACHES

Table 4 contains the variables deemed to be salient by both approaches. Overall, there are several observations which can be drawn based on the groups of data. First, only two of the LO deployment group, course and LO topic, are considered salient by both approaches. On the other hand, these variables contain the highest absolute HLM coefficients and amongst the highest Lasso coefficients. Second, a considerable number of the variables in the demographic group (4/10) are considered salient by both approaches. Additionally, HLM and Lasso agree that the coefficients for all these variables should be positive. Third, both approaches agree that only 4/12 of the MSLQ variables should be salient. Fourth, and perhaps most interesting, only a small minority of (3/43) LO session variables are considered salient by both approaches.

Finally, both approaches agree that a majority of the variables (6/10) in the evaluation questions are salient.

The LO deployment group results are expected because both the Course and LO topic are very important determinants of the assessment scores. First, the different courses contain students with a wide range of backgrounds from students in a programming course for non-majors, to students in an honors course for computer science and business. Student backgrounds have a significant impact on student learning as measured using the student's assessment score, which is consistent with previous research showing that prior achievement and knowledge are predictors of learning (Snow 1994; Schute and Towle 2003). In fact, students in the honors course had significantly higher assessment scores than students in lower-level courses. Second, all the LOs were designed to cover different introductory computer science concepts. These different concepts have very different content which have been shown to be connected to the student's assessment score (Miller et al. 2011a). Interestingly, both approaches do not agree that LO difficulty—which is related to the LO topic—should be salient. We discuss the LO difficulty later in the paper.

The high degree of agreement between the approaches on the demographic and evaluation groups is consistent with previous work. First, prior achievements and knowledge such GPA, ACT, math courses taken, etc., are predictors of learning as previously discussed (Snow 1994; Schute and Towle 2003). Second, the evaluation questions we use are based on those in Nugent et al. 2011). These questions have been previously shown to be significantly correlated with student learning. Furthermore, both the FS and HLM approaches are regression-based and use the same general organization of the data records in the iLOG dataset and the same independent variables. With these commonalities, it can be expected that both approaches will consider many of these well-established predictors of learning to be salient.

The relatively low agreement on variables in the MSLQ group is unexpected. As with the demographic group, these MSLQ variables have previously shown to be useful predictors of learning (Pintrich et al. 1993). However, in our results, there is little agreement on which of these variables is salient. Instead, each approach considers a different subset of the remaining MSLQ variables to be salient. We suspect that the diversity in the student backgrounds, reflected in the MSLQ responses, is making it more difficult for each approach to individually identify all the MSLQ variables as being salient.

Finally, the LO session data group results are expected when the independent variables are compared. The variables in this group are all computed from user interaction data collected without extensive post-processing such as outlier or noise filtering. The number of user interactions collected combined with the granularity (e.g., seconds) results in a wide range of real values compared to the other groups of variables which generally have a much narrower range of fixed values (e.g., each MSLQ question is on a Likert scale with only seven possible values). This wider range of values makes the variable space larger which makes it more difficult to separate the wheat from the chaff and find the salient variables in the LO session data.

Table 5 contains the variables deemed to be nonsalient by both approaches. Few of the variables for MSLQ or the Evaluation Survey groups are nonsalient for both approaches. This is expected because these variables are based on existing education research on learning outcomes. Furthermore, these results validate that using both approaches together can identify more variables known to be salient. The majority of nonsalient variables (80%) are from the LO

session data. Indeed, this high percentage is inflated due to the fact that many of these variables contain overlapping data (e.g., students with high max seconds on a page tend to have high mean seconds on a page). Additionally, students are allowed to skip the tutorial/exercise components. For such students, variables measuring data collected in the tutorial component will obviously be less important to the assessment score.

Table 4: Variables deemed to be salient by both HLM and Lasso. Note that (*) Coefficients for categorical predictors are F-test values.

Group	Independent Variable	HLM Coeff (*p*-value)	Lasso Coeff
LO Deployment	Course	27.300 (.000)*	0.152
	LO Topic	53.890 (.000)*	-0.182
Demographic Survey	GPA	5.559 (.000)	0.149
	ACT	1.349 (.000)	0.085
	Highest Math Course	2.051 (.000)	0.064
	Number of Programming Courses Taken	3.997 (.000)	0.135
MSLQ Survey	Self Efficacy for Learning	2.529 (.000)	0.045
	Task Value	1.891 (.000)	0.057
	Problem Solving	2.620 (.000)	0.091
	Effort Regulation	3.071 (.000)	0.139
LO Session Data	assessmentAvgClicksPerPage	-1.452 (.000)	-0.278
	assessmentMaxClicksOnAPage	-0.349 (.000)	-0.055
	exerciseAvgSubmits	0.084 (.038)	-0.001
Evaluation Survey	EvaluationQ3-The LO are a valuable addition to this course	6.046 (.000)	0.059
	EvaluationQ5-The LO helped me understand more about this topic	6.899 (.000)	0.218
	EvaluationQ7-I learned more from this LO than from listening to the professor	4.335 (.000)	0.018
	EvaluationQ8-The material in this LO was difficult for me to understand	-5.280 (.000)	-0.042
	EvaluationQ9-This LO needed to go into greater detail	-4.677 (.000)	-0.132
	EvaluationQ10-Overall how would you rate this LO	5.539 (.000)	0.182

Table 5: Variables deemed to be nonsalient by both HLM and Lasso.

Group	Independent Variable	HLM Coeff (*p*-value)	Lasso Coeff
LO Deployment	Exercise Activity	0.005 (.764)	0
	Exercise Feedback Seconds	-0.019 (.103)	0
	Exercise Feedback Type	0.758 (.208)	0
Demo. Survey	Grade Level	1.540 (.175)*	0
MSLQ Survey	Extrinsic Goal Orientation	0.252 (.700)	0
	Help Seeking	0.542 (.361)	0
LO Session Data	assessmentMaxSecondsOnAPage	-0.001 (.204)	0
	assessmentMinClicksOnAPage	-0.768 (.281)	0
	assessmentMinSecondsOnAPage	0.016 (.71)	0
	assessmentStdDevClicks	-0.005 (.149)	0
	assessmentStdDevSeconds	-0.005 (.145)	0
	exerciseAvgClicksPerPage	-0.067 (.427)	0
	exerciseAvgInterval	0 (.468)	0
	exerciseAvgSecondsPerPage	-0.004 (.066)	0
	exerciseMaxClicksOnAPage	-0.022 (.721)	0
	exerciseMinSecondsOnAPage	-0.002 (.466)	0
	exerciseStdDevClicks	0.064 (.586)	0
	exerciseStdDevInterval	0 (.775)	0
	exerciseStdDevSubmits	0.115 (.16)	0
	exerciseTotalEntries	0.020 (.185)	0
	exerciseTotalInterval	0 (.103)	0
	exerciseTotalSeconds	-0.001 (.21)	0
	totalTime	0 (.876)	0
	totalTutExAssessSeconds	0 (.842)	0
	tutorialAvgSecondsPerPage	-0.002 (.729)	0
	tutorialMaxClicksOnAPage	0.048 (.515)	0
	tutorialMaxSecondsOnAPage	0.001 (.454)	0
	tutorialMinClicksOnAPage	-0.974 (.098)	0
	tutorialStdDevClicks	0.061 (.791)	0
	tutorialStdDevSeconds	0.001 (.798)	0

Overall, these results validate HLM and Lasso both individually and used in a combined approach. Both approaches are able to identify many of the same variables as salient/nonsalient despite differences in the underlying process used. First, both approaches agree that the majority of the variables in the demographic and evaluation groups which is consistent our background knowledge and expectations. This supports the effectiveness of the individual approaches. Second, both approach agree that many variables should be nonsalient. This shows that the combined approach can discriminative and not simply consider all the variables to be salient.

To further establish the effectiveness of the combined approach, we examine the 32/81 variables on which the HLM and Lasso are in disagreement (i.e., salient by one approach, but not the other).

5.5. LASSO-ONLY SALIENT VARIABLES

Table 6 shows the variables found to be salient by Lasso but not by HLM. We observe the advantages of the Lasso approach in identifying some salient variables that are missed by the HLM approach. In particular, Lasso selects as salient the LO difficulty variable, three demographic variables including gender, one MSLQ variables, and seven LO Session variables—none of which are salient according to HLM.

LO Difficulty. The LO difficulty was provided by the content experts to measure the difficulty of the LO including the assessment so we expected this variable to be salient for both approaches and not just Lasso. Upon further investigation, we found that the LO difficulty was salient for HLM on four out of five courses (expected). In one of the four courses, however, difficulty was positively related to assessment scores, which is the opposite of what would be expected. This course contained students who were learning a more limited range of CS topics and were extremely unmotivated to take the less difficult LOs. Because difficulty was not salient in one course and salient in the opposite direction in another, the overall effect of difficulty across all courses did not average out to be significant and positive. On the other hand, Lasso is able to focus more selectively on specific courses since it can automatically fine-tune its shrinkage factor (Hastie et al. 2011).

Demographic and MSLQ Survey. Gender is a demographic variable of particular interest to researchers (Nugent et al. 2009). One possible explanation for why HLM did not find gender to be salient is that HLM assumes residuals, or differences between model-predicted and observed outcome scores, are normally distributed. This assumption may not hold for gender given the relatively small number of female students who took the LOs (16% of the total students). In general, the variables identified as salient by Lasso but not by HLM had residuals that were slightly more skewed, although none were found to be highly problematic. As a result, there is no way to determine whether HLM did not identify them because of violations of assumptions or that there are no direct relationships between the variables and assessment scores. It is also possible that a variable that does not have a direct relationship with outcomes may be moderated by another variable. In other words, the strength of the relationship between, say, extrinsic goal orientation and assessment scores may depend on the individual's level of organization. The HLM analyses performed in this study only investigated direct relationships. In many cases, however, HLM simply may not have found variables to be salient because their overall relationships with assessment scores are near zero or the standard errors for the regression coefficients are too large to find the effect to be significant. Likewise, the MSLQ variables tend to be more similar within each course (since students have more similar backgrounds), and thus the drilling down ability of HLM picks these up as salient. Meanwhile, HLM did not identify these as salient variables because it considers the entire set of values as a whole and is not likely to show significance if the variable is salient between a small combination of variable values. Again, Lasso uses a shrinkage method for FS which allows it to find combinations of predictor variables which are consistently salient to the outcome scores (Zhao and Yu, 2006).

LO Session Data. The LO session indicators, such as the total number of clicks during the exercises or assessment, were not found to be significant predictors by HLM likely because of a large amount of near-zero values. In other words, "floor effects" may be limiting our ability to identify salient session predictors (Pickering 2002). As an example, Figure 2 presents a scatterplot of assessment scores by the total clicks on the exercises (exerciseTotalClicks). The diagonal line shows the relationship that would be expected if there was a significant positive correlation between the predictor and the outcome. Points that overwhelmingly fall on a vertical or horizontal line indicate no relationship is present. Here, the majority of values on the predictor fall very near the lower boundary of zero, making it difficult for HLM to detect a relationship between the two variables. Similar trends are seen in all of the LO session variables presented in Table 6.

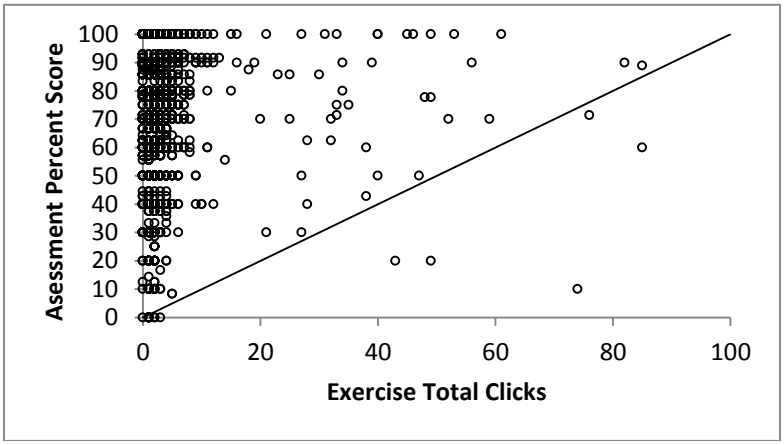

Figure 2: Scatterplot of assessment scores against a nonsalient LO session variable

Overall, HLM is concerned with making generalizations regarding the relationship between the predictor and outcome variables, considering the full range of values of each variable. As a result, for a variable to be considered salient, there must be a monotonically increasing or decreasing trend in the outcome scores as the value of the predictor variable increases. If the average of the LO scores to not appear to change across different levels of the predictor variable, HLM is not likely to find the variable to be salient. On the other hand, Lasso uses a shrinkage method which accommodates combinations of predictor variables. Therefore, a combined approach using Lasso may be the better option on datasets where such variables are relatively common.

Table 6: Variables identified as salient by Lasso but not by HLM.

Group	Independent Variable	HLM Coeff (*p*-value)	Lasso Coeff
LO Deployment	LO Difficulty	-0.227 (.488)	0.217
Demographic Survey	Gender	0.823 (.676)	0.006
	Required Course for Major	-1.966 (.243)	-0.148
	Number of Other Courses Taken	0.210 (.713)	-0.002
MSLQ Survey	Organization	0.243 (.670)	-0.054
LO Session Data	assessmentTotalClicks	-0.022 (.303)	0.034
	exerciseStdDevEntries	0.025 (.749)	-0.016
	exerciseTotalClicks	0.045 (.379)	0.002
	exerciseTotalSubmits	0.020 (.221)	-0.005
	exerciseMinClicksOnAPage	-0.125 (.209)	-0.008
	tutorialAvgClicksPerPage	-0.212 (.483)	-0.357
	totalExToTutClicks	-1.005 (.129)	-0.212

5.6. HLM-ONLY SALIENT VARIABLES

Table 7 shows the variables found to be salient by HLM but not by Lasso. We observe the advantages of the HLM approach in identifying some salient variables that are missed by the Lasso approach. HLM selects as salient two Demographic variables, five MSLQ variables, nine LO Session variables, and four Evaluation variables which are nonsalient according to Lasso.

Demographic and MSLQ Survey. The Lasso approach considers College Major and Placement Test from the Demographic Survey to be nonsalient along with five of the MSLQ Survey group variables: (1) Control of Learning Beliefs, (2) Elaboration, (3) Intrinsic Goal Orientation, (4) Metacognitive Self-Regulation, and (5) MSLQ Average. As alluded to earlier, this is not consistent with our expectations since the Demographic and MSLQ variables have previously been shown to be useful predictors of learning (Snow 1994; Schute and Towle 2003 and Pintrich et al. 1993) and, thus, should probably be considered salient. To provide an explanation, we start by comparing the results for the HLM and Lasso approaches. Based on the results previously discussed, HLM finds 9/12 MSLQ variables as salient whereas Lasso only finds 5/12 as salient. Clearly, the HLM approach finds the vast majority of the MSLQ Survey group variables as salient while Lasso lags behind. Recall that the HLM approach takes into account nested or hierarchical data structures. In the iLOG dataset, MSLQ Survey variables are considered nested because the survey is taken once by each student—the survey responses do not change when that student uses different LOs. These nested variables differ from the LO Session Data and Evaluation Survey variables measured each time (i.e., session) the student uses different LOs. HLM evaluates nested variables separately using the assessment scores from the different LOs. On the other hand, the Lasso approach lacks the capability to evaluate nested variables and, instead, treats them as session variables duplicating the responses for each LO.

When treated as session variables, the values for the Demographic and MSLQ variables are duplicated with the same set of values copied to all sessions involving the same student using

different LOs. Because Lasso does not accommodate nesting, duplicate values across multiple sessions are evaluated separately without any special consideration. As a result, these duplicate values often appear to be nonsalient since the same set of values (from a student) are associated with different assessment scores (from different LOs). This makes it more difficult for the Lasso approach to find nested variables as salient, in particular, the MSLQ variables.

Overall, the Lasso approach finds fewer Demographic and MSLQ variables to be salient. Based on our results, Lasso tends to give less consideration to nested variables because it lacks the capability to use nesting. These variables often appear nonsalient when the same set of values are associated with different assessment scores. This can make it difficult for Lasso approach to identify nested variables on datasets which contain both nested and session variables. Therefore, a combined approach using HLM may be the better option on datasets where nested variables are relatively common.

LO Session Data. For the LO Session Data group, there are nine variables which are not salient according to the Lasso approach: (1) assessmentAvgSecondsPerPage, (2) assessmentTotalSeconds, (3) exerciseAvgEntries, (4) exerciseMaxSecondsOnAPage, (5) exerciseStdDevSeconds, (6) totalSecondsEval, (7) tutorialMinSecondsOnAPage, (8) tutorialTotalClicks, and (9) tutorialTotalSeconds. Again, to provide an explanation, we start by comparing the results for the HLM and Lasso approaches. Based on the results previously discussed, both approaches find nine LO Session Data variables salient which are not considered to be salient by the other approach. Additionally, we observe that there are variables considered salient by one approach, but, while those exact same variables are not considered to be salient by another approach, other variables with similar properties are. For example, Lasso considers exerciseTotalSubmits and exerciseTotalEntries to be salient while HLM, instead, considers the similar variables exerciseAvgSubmits and exerciseAvgEntries to be salient.

Based on further consideration, although both the HLM and Lasso consider variables from all the LO components (tutorial, exercises, and assessment) as salient there is one noticeable trend. The Lasso approach tends to find only variables involving the number of clicks on a page to be salient while the HLM approach tends to prefer variables involving the number of seconds on a page. For the HLM approach, this make sense given that variables involving the number of clicks are more likely to be subject to the "floor effects" described previously in Section 5.5 (and see Figure 2) than variables involving the number of seconds since the number of seconds varies considerably more than the number of clicks.

Now, to explain the results for the Lasso approach, recall that Lasso uses an internal cross-validation to choose the optimal lambda by evaluating multiple sets of coefficients for given lambda values on a separate (validation) set of records. The optimal lambda is the one whose coefficients give the highest accuracy for the assessment scores on the validation set. This cross-validation method for choosing the salient variables is very different from that used by HLM which evaluates whether variables are salient using the same set of records used to compute the coefficients in the first place. In machine learning parlance, Lasso uses an inductive method to determine the salient variables, while HLM uses a transductive method. The downside to using a transductive method is the potential for "fitting the data" thereby choosing additional variables salient only to the current records. Lasso, instead, uses an inductive method to find only the variables actually salient to the underlying data distribution. This helps to explain why Lasso finds slightly fewer variables than HLM (see Table 3) and supports our previous claim that Lasso is more conservative.

Overall, both approaches find a similar number of LO Session Data group variables to be salient. As shown in Table 4, there is very little overlap in the LO Session Data variables chosen by both approaches (only two variables in common). This provides further motivation for using a combined approach to identify salient variables rather than relying on a single approach.

Table 7: Variables identified as salient by HLM but not by FS

Group	Independent Variable	HLM Coeff (*p*-value)	Lasso Coeff
Demographic Survey	College Major	4.830 (.008)*	0
	Placement Test	1.464 (.000)	0
MSLQ Survey	Control of Learning Beliefs	1.360 (.021)	0
	Elaboration	1.492 (.020)	0
	Intrinsic Goal Orientation	1.226 (.034)	0
	Metacognitive Self-Regulation	3.202 (.000)	0
	MSLQ Average	3.595 (.000)	0
LO Session Data	assessmentAvgSecondsPerPage	-0.025 (.002)	0
	assessmentTotalSeconds	-0.002 (.007)	0
	exerciseAvgEntries	0.144 (.000)	0
	exerciseMaxSecondsOnAPage	-0.003 (.044)	0
	exerciseStdDevSeconds	-0.005 (.022)	0
	totalSecondsEval	-0.007 (.007)	0
	tutorialMinSecondsOnAPage	-0.054 (.030)	0
	tutorialTotalClicks	0.143 (.000)	0
	tutorialTotalSeconds	0.002 (.001)	0
Evaluation Survey	EvaluationQ1-The LO (web module) was easy to use	5.992 (.000)	0
	EvaluationQ2-The LO maintained my interest more than listening to the professor	5.002 (.000)	0
	EvaluationQ4-More of the material in this course should be presented via the web	4.802 (.000)	0
	EvaluationQ6-I will use the same LO again in the future if I have questions about this topic	4.225 (.000)	0

Evaluation Survey. There are four variables in the Evaluation Survey group which are not salient: (1) EvaluationQ1, (2) EvaluationQ2, (3) EvaluationQ4, and (4) EvaluationQ6. Summaries of what each question covers can be found in Table 7. When we compare the HLM and Lasso approaches, HLM finds 10/10 evaluation questions as salient while the Lasso only finds 6/10 questions as salient. As with the MSLQ Survey group, these results are unexpected given that the evaluation questions have previously been shown to be useful predictors of learning (Nugent et al. 2011).

To explain this discrepancy in salient variables, recall that the evaluation survey was optional and, thus, many of the students did not fill out the survey after completing the assessment. As a result, approximately 70% of the values for these five variables are missing in the iLOG dataset. Now, the HLM and Lasso approaches operate differently on records containing missing values. The HLM approach has the capability to ignore missing values whereas the Lasso does not, instead, treating missing values as being zeros. The presence of all these zero values tends to skew the regression residuals for these variables and making it more likely that Lasso will treat these variables as being nonsalient. Unfortunately, Lasso operates on all such variables together making it impractical to selectively remove missing values.

Overall, the Lasso approach finds fewer Evaluation Survey group variables to be salient. Many students did not fill out the evaluation survey resulting in considerable missing values for these variables. Unlike the HLM approach, Lasso lacks the capability to ignore missing values. This can make it difficult for Lasso to find salient variables which contain large numbers of missing values. Therefore, a combined approach using HLM may be the better option for evaluating variables on datasets where missing values are relatively common.

5.7. VARIABLE CORRELATION ANALYSIS

As alluded to earlier, the purpose of this correlation analysis is to address concerns that disagreement between approaches may be the result of redundant variables in the dataset. HLM and Lasso are both designed to find *only* salient variables discarding those deemed redundant. As an example, when two variables have 1.0 correlation, these approaches select one from the pair and discard the other. However, these approaches use very different methods as previously discussed. This could well result in each approach picking a different variable from that pair. In this example, even though there appears to be disagreement (because the two variables have different names), both approaches have effectively found the same results.

Table 8 shows pairs of variables with significant correlation where the first variable was identified as salient ***only*** by HLM and second variable ***only*** by Lasso is extremely significant with $n = 1000$ at the $p < 0.0001$ level ($r \geq 0.13$). We picked this small p-value to focus on the pairs of variables with the highest correlation that are most likely to contain redundant variables.

First, the significant correlation between organization and other MSLQ variables is to be expected ($r = 0.69$ with elaboration). The MSLQ is based on extensive research regarding its predictive validity; and the psychometric technical report shows a significant correlation between organization and final course grade based on data from Midwestern college students enrolled in 14 courses, including CS (Pintrich et al. 1993). However, there are studies that say that this particular variable does not predict learning in CS (Bergin et al. 2005) and that organization was not a good predictor (Crede and Phillips 2011). In our study, our HLM finding of non-significance is thus supported in some of the published research; on the other hand, Lasso finding the variable to be salient is also supported by some other published research.

Next, the significant correlation between required course and major ($r = 0.42$) is to be expected. We would anticipate that CS or engineering majors would be taking the targeted course as one of their required degree requirements and that the course would not be required for non-engineering or CS majors. It would also follow that college major would be a significant predictor of CS learning, yet the HLM analyses did not identify major as a salient variable. A similar example is the high correlation between exercise average entries and the total number of

exercise submits; both represent measures of student interactivity within the exercise component of the LO.

Overall, this analysis shows correlations between pairs of variables where the first variable is considered salient only by HLM and the second only by Lasso. Although significant, these correlations are far smaller than 1.0 where we could safely discard one variable as redundant. Although variables in the same pair may be related, each still contains unique data worth including in the model. Exploring further, existing literature has reported at times that these variables are salient and at other times nonsalient. These borderline salient variables are likely to result in disagreement between HLM and Lasso because different methods are required to identify these variables as salient. Lastly, predictive accuracy provides the final indication that disagreement is not the result of redundant variables. As shown in Section 5.2, including all the variables improves the overall predictive accuracy. This would not be the case for redundant variables that provide no additional information.

Table 8: Pairs of variables with significant correlation ($r \geq 0.13$) where the first variable was identified as salient only by HLM and second only by Lasso.

HLM Salient	Lasso Salient	Correl
Elaboration	Organization	0.69
Metacognitive Self-Regulation	Organization	0.63
Intrinsic Goal Orientation	Organization	0.48
Control of Learning Beliefs	Organization	0.28
collegeMajor	requiredCourse	0.42
collegeMajor	Number of Other Courses Taken	0.16
EvaluationQ2-The LO maintained my interest more than listening to the professor	Organization	0.13
EvaluationQ4-More of the material in this course should be presented via the web	Organization	0.14
EvaluationQ6-I will use the same LO again in the future if I have questions about this topic	Organization	0.19
exerciseAvgEntries	exerciseTotalSubmits	0.65
exerciseAvgEntries	exerciseStdDevEntries	0.61
tutorialTotalClicks	assessmentTotalClicks	0.32
tutorialTotalClicks	exerciseTotalSubmits	0.18
Placement Test	assessmentTotalClicks	0.15

6. SUMMARY AND CONCLUSIONS

6.1. RESEARCH IMPLICATIONS FOR BIG DATA

Much of the impetus for ***both*** educational data mining and learning analytics has come from the vast amounts of digital data generated by online learning systems, and our research is based on

such data. However, in addition to online learning systems, there is also a large amount of data at local, district, university, state and federal levels that can inform education policy, management, and budgeting decisions. Student information systems at the state and university level, and public educational data repositories such as the National Center for Education Statistics contain a wealth of data that could be mined and analyzed to improve student outcomes and the productivity of our education systems.

Federal agencies, including the U.S. Department of Education, are recognizing the need to manage, analyze and use such educational data (Bienkowski, Feng, and Means 2012). This agency has started an initiative to consolidate collection, analysis, and reporting of K-12 performance data (EDFacts 2010) and has begun funding statewide longitudinal data systems. The goal of this initiative is to aggregate student data across educational careers to help states, districts, and schools make data-informed decisions to improve student learning and outcomes (Statewide Longitudinal Data Systems Grant Program). Additionally, the NSF has recognized the national "big data" challenges in dealing with the large, diverse, complex, longitudinal, and/or distributed data sets generated from digital sources available today and in the future (NSF 2012).

Harnessing the power of big data using educational data mining and learning analytics is becoming a goal of educational sectors, and research on methods and techniques to analyze such data is becoming critical. This paper has shown that two separate approaches, while effective individually at identifying salient variables, are even more effective when used together in a combined approach. By leveraging the complementary and synergistic results, the combined approach was able to improve the prediction accuracy for models using the salient variables and identify additional variables considered salient in previous datasets on student learning. In this way, the combined approach provides a more comprehensive picture of the predictors of student learning on the iLOG dataset. Because these predictors are extremely useful for researchers, we envision our combined approach as the first step in better understanding a wide variety of big data. The application of our combined approach to other datasets is discussed further in the future work.

6.2. Conclusions on Similarities and Differences

In comparing the two approaches, it is important to look at similarities and differences in both the processes and the results generated by these approaches since the differences in results can typically be traced back to variations in these approaches. Table 9 gives a summary of the differences between these approaches, while Table 10 gives a summary of the similarities, as a means of introducing this discussion of the results. Although there are more differences than similarities between these approaches, the approaches are still predominantly in agreement in terms of the salient variables found: of the 81 variables tracked, there was 60% agreement. These results provide clear direction for what variables should be tracked and tagged within an online environment for undergraduate computer science. As alluded to earlier, the 40% of results that showed disagreement provide support for using a combined approach to identify all the salient variables. Furthermore, expanding discussion of differences in Section 4.3, these results can be explained by variations in these two approaches, including (a) handling of missing data, (b) organization of data, (c) use of variable combinations or individual variables.

Missing Data. In order to insure that each approach was working from the same dataset, we did not use typical data imputation methods to account for missing data. As a result there were

sections (such as answers to the evaluation questions) with widespread missing data. The Lasso approach considers variables with considerable missing data as less salient; the HLM approach simply ignores the missing values and calculates results based on available data. Thus, the Lasso approach downgrades variables with missing data, while the HLM approach provides results which may be based on a biased sample. This is evident in looking at the evaluation question results where HLM found more salient features than did Lasso. Given these clear differences it is not surprising that variables with considerable missing data resulted in different results from the two approaches. This could suggest the ***need for data imputation methods*** as results from both approaches are likely to be impacted.

Organization of Data. Both approaches used the same iLOG dataset. However, the HLM approach has the capability to take into account nested or hierarchical data structures such as LOs nested within students. This allowed HLM to give special consideration to nested variables such as those involving the Demographic and MSLQ survey where a single student has only one set of values. The Lasso approach lacked this capability and was forced to evaluate these variables along with session variables which varied from session to session. The resulting duplication of values—and also the assumption that these values are the same from session to session—made it difficult for Lasso to find these variables as salient. Therefore, we suggest including in the combined approach at least one approach that accommodates variables with a nested or hierarchical data structure.

Table 9: Summary of the Process Differences between Lasso and HLM.

Characteristic	Lasso	HLM
Approach	• Nonsalient variables have zero coefficients • Evaluate results using separate set of records (inductive)	• Nonsalient variables identified by statistical test • Evaluate results using current set of records (transductive)
Missing Data	• Variables with widespread missing data are considered less salient	• Missing data are ignored in calculations
Organization of Data	• Evaluates each record separately (i.e., session variables)	• Takes into account hierarchical nature of data (i.e., nested variables)
Variables	• Results provided for individual variables and variable combinations	• Results provided for individual variables

Use of variable combinations versus individual variables. The Lasso approach uses a shrinkage method which allows it to find combinations of predictor variables which are consistently salient to the outcome scores (Zhao and Yu, 2006). This contrasts with the HLM approach which considers individual variables without automatically testing combinations. A variable salient by itself but not in combination with others is more likely identified by HLM. On the other hand, a variable that is a weak predictor of outcomes by itself but is found in combination with other variables is more likely to be identified by Lasso. Therefore, we suggest

including in the combined approach at least one approach that can accommodate variables salient only in combination.

In summary, the HLM approach is most valuable when one wants to draw broad generalizations about a particular population, in this case undergraduate computer science students exposed to multiple learning objects. This approach would be most appropriate for answering questions such as "What can we conclude about learner, content, and instruction attributes that contribute to learning from LOs in undergraduate computer science?" The Lasso approach is more valuable in looking for relationships that involve a particular group of students (i.e. the honors students) or particular LOs (i.e. the more difficult LOs). The Lasso approach can be extremely helpful in providing information relevant to the individualization of instruction, i.e., adapting the instructional presentation for particular types of students or particular type of courses.

Finally, there is one additional important lesson we want to emphasize. Based on our results, these approaches are complementary and synergistic. To gain the greatest insight from student use and learning from LOs—both should be considered—with careful matching of the approach to the research question and the particular information desired.

Table 10: Summary of the Process Similarities between Lasso and HLM.

Characteristic	**Lasso**	**HLM**
Approach	• Minimize the residuals for the sum of squares	
Dependent Variable	• Evaluate salient variables using continuous, dependent variable	
"Noise" in data	• Accept "noise", use all data, including outliers	
Results	• Coefficients provide information about direction of relationship between predictor variable and outcome	

6.3. FUTURE WORK ON EDUCATIONAL DATA MINING

Future work should consider using a combined approach to identify variable salience, relying on the strengths and unique functionalities of each approach and taking into account each one's relevance to the instructional context, the research questions, and the nature of the data collected. In order to further define the strengths of each approach, future research should also consider how the two approaches could directly address the issue of isolating effective types of treatments for particular types of learners, the precursor for the design of adaptive algorithms and instruction. Such research could contribute to the growing body of literature informing the design of educational materials and environments that adapt to the particular characteristics of the learner. To this end, we intend to investigate how a combined approach can improve educational data mining (EDM) in related areas.

Additional Learning Measures and Datasets. In this work, we evaluate all three approaches for choosing salient variables on the iLOG LO dataset. The iLOG dataset uses the student score on the LO assessment as the measure of learning. Although standard for LO datasets, such a measure does not take into account any prior knowledge the student has on the LO content. Now, we do include student placement exam results, which evaluates prior knowledge, as an independent variable. Nevertheless, we still intend to investigate alternative measures of learning such as assessment scores normalized using placement exam results. Additionally, we have shown in Section 5 that the combined approach found the same salient variables as those

found in previous datasets. This lends some support to our claim that the combined approach works beyond the iLOG dataset. However, for previous datasets, the salient variables were found using different approaches. As such, to further demonstrate the effectives of the combined approach, we intend to evaluate the predictive accuracy directly on the previous datasets.

Additional Approaches. The HLM and Lasso approaches considered in this paper that are both based on an underlying regression model. As shown in Section 5, when combined, even such related approaches allow for more salient variables in the ground truth to be discovered than using a single approach. The next step is adding additional approaches to the combined approach to provide an even more comprehensive view of the salient variables in the ground truth. Two additional approaches that look promising (since they are orthogonal, but not completely unrelated) are (1) Group Lasso (Yuan and Lin 2006) that decides whether a subset of variables should be salient together and (2) Bayesian models that use orthogonal transformations on the data to decide which variables should be salient (Davis, Pensky and Crampton 2011). Furthermore, adding these approaches to the combined approach allows for more flexibility when balancing the number of redundant variables found with the ground truth (as discussed in Section 5.2). This balance could be achieved by requiring a specified number of votes (from single approach) that a variable is salient before the combined approach deems it so. With fewer votes, the combined approach is more likely to identify all the salient variables in the ground truth, but is also more likely to find redundant variables.

Predictive Models. EDM has also been used for creating predictive models for student assessment and learning styles. Romero et al. (2008) compares different EDM algorithms for classifying students based on usage metadata from a web-based course. Twenty-five models were used from five different categories including statistical classification, decision trees, rule induction, fuzzy-rule induction, and neural networks. The datasets used were created from Moodle LMS logs on the use of LOs (e.g., assignments, forums, quizzes). The authors assume that all variables are salient and focus more on using records with an equal number of labels. Including such variables will reduce the accuracy for these algorithms. Thus, it is important to carefully determine which variables are salient ***before*** applying such models. Papadimitriou et al. (2009) discuss the MATHEMA system which uses EDM to select LOs for students based on the students' learning style. Data are collected from assessment questions (pre/post) and survey responses and mined using didactic strategies to predict whether the student uses a diverger, assimilator, converger, or accommodator learning style. Such mining assumes that all variables are salient. Again, the hybridized approach could help improve these predictive models by determining which variables are actually salient.

Association Rules. EDM has also been used for mining association rules from existing data. Romero et al. (2010) focuses on mining rare association rules from recorded user interactions. This paper applies common/rare rule mining techniques to the educational data collected from LOs. The results show that confidence and relative support are higher for rare rules compared to common rules. Rare rule mining techniques search more of the dataset because rare rules involve fewer records than common rules. Removing nonsalient variables reduces the overall search space and makes it easier to find high quality rules. Kruger et al. (2010) use a similar approach to automate and alleviate the preprocessing needed to mine the data. Again, the dataset is likely to contain nonsalient variables which make it more difficult to find high quality rules.

7. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0632642 and an NSF GAANN fellowship. The authors would like to previous iLOG team members for their programming and data processing work: Sarah Riley, Erica Lam, WayLoon Tan, Beth Neilsen, and Nate Stender. The authors would also like to thank Shiyuan Wang and Chao Rong for their invaluable advice on HLM.

8. REFERENCES

ALFONS, A. 2012. cvTools: Cross-validation tools for regression models. R package version 0.3.2.

ALVARADO, B., ZUNZUNEGUI, M., DELISLE, H., AND OSORNO, J. 2005. Growth trajectories are influenced by breast-feeding and infant health in an afro-colombian community. *Journal of Nutrition*, 2171–2178.

BAKER, R. 2010. *International Encyclopedia of Education* (3rd edition). Oxford, UK: Elsevier, Chapter Data mining in education.

BERGIN, S., REILLY, R., AND TRAYNOR, D. 2005. Examining the role of self-regulated learning on introductory programming performance. In *Proceedings of the 1st international workshop on Computing education research*. 81–86.

BERK, J. 2004. The state of learning analytics. *T&D*, 34–39.

BIENKOWSKI, M., FENG, M., AND MEANS, B. 2012. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Tech. rep., U.S. Department of Education.

CHEN, C. 2002. Self-regulated learning strategies and achievement in an introduction to information systems course. *Information Technology, Learning, and Performance Journal* 20, 11–23.

COHEN, J., COHEN, P., WEST, S., AND AIKEN, L. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd edition). Mahwah, NJ: Lawrence Earlbaum Associates, Inc.

CREDE, M., PHILLIPS, L. A. 2011. A meta-analytic review of the Motivated Strategies for Learning Questionnaire. *Learning and Individual Differences* 21, 337-346.

DAVIS, J., PENSKY, M., AND CRAMPTON, W. 2011. Bayesian feature selection for classification with possibly large number of classes. *Journal of Statistical Planning and Inference* 141, 3256–3266.

DELEN, D. 2009. Analysis of cancer data: A data mining approach. *Expert Systems* 26, 100–112.

EDFacts. 2014. The edfacts initiative. U.S. Department of Education.

FERRON, J., BELL, B., HESS, M., RENDINA-GOBIOFF, G., AND HIBBARD, S. 2009. Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods* 41, 372–384.

FRANCIA, G. 2003. A tale of two learning objects. *Journal of Educational Technology Systems* 3, 117–190.

FRIEDMAN, J., HASTIE, T., HOFLING, H., AND TIBSHIRANI, R. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1, 302–332.

GRAVETTER, F. AND WALLNAU, L. 2004. *Statistics for the Behavioral Sciences* (6th edition). Belmont: Wadsworth/Thomson Learning.

HALKITIS, P., PALAMAR, J., AND MUKHERJEE, P. 2008. Analysis of HIV medication adherence in relation

to person and treatment characteristics using hierarchical linear modeling. *AIDS Patient Care and STDs* 22, 323–335.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2011. *The Elements of Statistical Learning* (2nd edition). Springer-Verlag.

HERNANDEZ-LLOREDA, M., COLMENARES, F., AND MARTINEZ-ARIAS, R. 2004. Application of piecewise hierarchical linear growth modeling to the study of continuity in behavioral development of baboons (papio hamadryas). *Journal of Comparative Psychology* 118, 316–324.

HINDMAN, A., SKIBBE, L., AND ZIMMERMAN, M. 2010. Ecological contexts and early learning: Contributions of child, family, and classroom factors during head start, to literacy and mathematics growth through first grade. *Early Childhood Research Quarterly* 25, 235–250.

HOFMANN, D. AND GAVIN, M. 1998. Centering Decisions in Hierarchical Linear Models: Implications for Research Organizations. *Journal of Management* 24, 623–641.

KAHN, J. 2011. Multilevel modeling: overview and applications to research in counseling psychology. *Journal of Counseling Psychology* 58, 257–271.

KRUGER, A., MERCERON, A., AND WOLF, B. 2010. A data model to ease analysis and mining of educational data. In *3rd International Conference on Educational Data Mining* (EDM). 131–140.

LOCKER, L., HOFFMAN, L., AND BOVAIRD, J. 2007. On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods* 39, 723–730.

MAYER, R. 2001. *Multimedia Learning*. New York: Cambridge University Press.

MCGREAL, R. 2004. *Online Education Using Learning Objects*. Psychology Press.

MCLAREN, B., SCHEUER, O., AND MIKSATKO, J. 2010. Supporting collaborative learning and e-discussions using artificial intelligence techniques. *International Journal of Artificial Intelligence in Education* 20, 1–46.

MILLER, L., SOH, L.-K., NUGENT, G., KUPZYK, K., MASMALIYEVA, L., AND SAMAL, A. 2011a. Evaluating the use of learning objects in CS1. In *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education*. 57–62.

MILLER, L., SOH, L.-K., NEILSEN, B., LAM, E., SAMAL, A., KUPZYK, K., AND NUGENT, G. 2011b. Revising computer science learning objects from learner interaction data. In *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education*. 45–50.

MILLER, L., SOH, L.-K., NUGENT, G., AND SAMAL, A. 2011c. iLOG: A framework for automatic annotation of learning objects with empirical usage metadata. *International Journal of Artificial Intelligence in Education*, 215–236.

NSF. 2012. Core techniques and technologies for advancing big data science and engineering. National Science Foundation.

NUGENT, G., KUPZYK, K., MILLER, L., MASMALIYEVA, L., SOH, L.-K., AND SAMAL, A. 2011. Learning analytic approach to identify attributes of learners and multimedia instruction that influence learning. In *Proceedings of the World Conference on Educational Multimedia, Hypermedia, and Telecommunications*. 2021–2028.

NUGENT, G., KUPZYK, K., RILEY, S., MILLER, L., HOSTETLER, J., SOH, L.-K., AND SAMAL, A. 2009. Empirical usage metadata in learning objects. In *Proceedings of the Frontiers in Education*. 1–8.

NUGENT, G., SOH, L.-K., AND SAMAL, A. 2006. Design, development, and validation of learning objects. *Journal of Educational Technology Systems* 34, 271–281.

OCHOA, X. AND DUVAL, E. 2009. Relevance ranking metrics for learning objects. *IEEE Transactions on Learning Technologies*, 34–48.

PAPADIMITRIOU, A., GRIGORIADOU, M., AND GYFTODIMOS, G. 2009. Interactive problem solving support in the adaptive educational hypermedia system mathema. *IEEE Transactions on Learning Technologies* 2, 93–106.

PICKERING, R. 2002. Statistical aspects of measurement in palliative care. *Palliative Medicine* 16, 359–364.

PINTRICH, P., SMITH, D., GARCIA, T., AND MCKEACHIE, W. 1993. Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement* 53, 801–813.

PINTRICH, P., SMITH, D., GARCIA, T., AND MCKEACHIE, W. 1999. Ann Arbor, MI: University of Michigan. *A Manual for the Use of the Motivated Strategies for Learning Questionnaire*. Ann Arbor, MI: University of Michigan.

RAMASWAMI, M. AND BHASKARAN, R. 2009. A study on feature selection techniques in educational data mining. *Journal of Computing* 1, 7–11.

RAUDENBUSH, S. AND BRYK, A. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd edition). Newbury Park, CA: Sage.

RILEY, S., MILLER, L., SOH, L.-K., SAMAL, A., AND NUGENT, G. 2009. Intelligent learning object guide (iLOG): A framework for automatic empirically-based metadata generation. In *Proceedings of the International Conference on Artificial Intelligence in Education*. 515–522.

ROMERO, C., ROMERO, J., LUNA, J., AND VENTURA, S. 2010. Mining rare association rules from e-learning data. In *Proceedings of the 3rd International Conference on Educational Data Mining* (EDM). 171–180.

ROMERO, C. AND VENTURA, S. 2010. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics* 40, 601–618.

ROMERO, C., VENTURA, S., ESPEJO, P., AND HERVAS, C. 2008. Data mining algorithms to classify students. In *Proceedings of the 1st International Conference on Educational Data Mining*. 8–17.

ROUNTREE, N., ROUNTREE, J., AND ROBINS, A. 2002. Predictors of success and failure in a CS1 course. In *Proceedings of the 33rd SIGCSE technical symposium on Computer Science Education*. 121–124.

SAYES, Y., INZA, I., AND LARRANGA, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2507–2517.

SCHONFELD, I. AND RINDSKOPF, D. 2007. Hierarchical linear modeling in organizational research longitudinal data outside the context of growth modeling. *Organizational Research Methods* 10, 417–429.

SEO, D. AND LI, K. 2009. Effects of college climate on students' binge drinking: hierarchical generalized linear model. *Annals of Behavioral Medicine* 38, 262–268.

SHUTE, V. AND TOWLE, B. 2003. Adaptive e-learning. *Educational Psychologist* 38, 105–114.

SIMON, N., FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* 39, 1–13.

SNOW, R. 1994. *Mind in context: Interactionist perspectives on human intelligence*. Cambridge: Cambridge University Press, Chapter Abilities in Academic Tasks.

SORGE, C. 2007. What happens? Relationship of age and gender with science attitudes from elementary to middle school. *Science Educator* 16, 33–37.

STACK, S. AND KPOSOWA, A. 2008. The association of suicide rates with individual-level suicide attitudes: A cross-national analysis. *Social Science Quarterly* 89, 39–59.

TERRACCIANO, A., MCCRAE, R., BRANT, L., AND COSTA, P. 2005. Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore longitudinal study of aging. *Psychology and Aging* 20, 493–506.

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society* 58, 267–288.

VENTURA, P. 2005. Identifying predictors of success for an objects-first CS1. *Computer Science Education* 15, 223–243.

WIEDENBECK, S., LABELLE, D., AND KAIN, V. 2004. Factors affecting course outcomes in introductory programming. In *16th Workshop of the Psychology of Programming Interest Group*. 97–110.

WILSON, B. AND SHROCK, S. 2001. Contributing to success in an introductory computer science course: a study of twelve factors. In *Proceedings of the 32nd SIGCSE technical symposium on Computer Science Education*. 184–188.

YUAN, M. AND LIN, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society* 68, 49–67.

ZHAO, P. AND YU, B. 2006. On model selection consistency of lasso. *Journal of Machine Learning Research* 7, 2541–2563.