Towards Adaptive Educational Assessments: Predicting Student Performance using Temporal Stability and Data Analytics in Learning Management Systems

Gautam S. Thakur
Oak Ridge National Laboratory
Oak Ridge, USA
thakurg@ornl.gov

Mohammed M. Olama
Oak Ridge National Laboratory
Oak Ridge, USA
olamahussemm@ornl.gov

Allen W. McNair
Scripps Networks Interactive
Knoxville, USA
WMcNair@scrippsnetworks.com

Sreenivas R. Sukumar
Oak Ridge National Laboratory
Oak Ridge, USA
sukumarsr@ornl.gov

Scott Studham
University of Minnesota
Minneapolis
studham@umn.edu

ABSTRACT

Data-driven assessments and adaptive feedback are becoming a cornerstone research in educational data analytics and involve developing methods for exploring the unique types of data that come from the educational context. For example, predicting college student performance is crucial for both the students and educational institutions. It can support timely intervention to prevent students from failing a course, increasing efficacy of advising functions, and improving course completion rate. In this paper, we present our efforts in using data analytics that enable educationists to design novel data-driven assessment and feedback mechanisms. In order to achieve this objective, we investigate temporal stability of students grades and perform predictive analytics on academic data collected from 2009 through 2013 in one of the most commonly used learning management systems, called Moodle. First, we have identified the data features useful for assessments and predicting student outcomes such as students' scores in homework assignments, quizzes, exams, in addition to their activities in discussion forums and their total Grade Point Average(GPA) at the same term they enrolled in the course. Second, time series models in both frequency and time domains are applied to characterize the progression as well as overall projections of the grades. In particular, the model analyzed the stability as well as fluctuation of grades among students during the collegiate years (from freshman to senior) and disciplines. Third, Logistic Regression and Neural Network predictive models are used to identify students as early as possible who are in danger of failing the course they are currently enrolled in. These models compute the likelihood of any given student failing (or passing) the current course. The time series analysis indicates that assessments and continuous feedback are critical for freshman and sophomores (even with easy courses) than for seniors, and those assessments may be provided using the predictive models. Numerical results are presented to evaluate and compare the performance of the developed models and their predictive accuracy. Our results show that there are strong ties associated with the first few weeks for coursework and they have an impact on the design and distribution of individual modules.

ACCESS '14 New York City, NY, USA

Categories and Subject Descriptors

H.2 [**Database Management**]: Big data; H.2.8 [**Database Applications**]: Data mining —*Learning, estimation*

1. INTRODUCTION

The use of web-based educational systems has grown exponentially in the last few years, spurred by the fact that neither students nor teachers are bound to a specific location [16]. Furthermore, collaborative and communication tools are also becoming widely used in educational contexts so, as a result, virtual learning environments are installed more and more by universities, community colleges, schools, businesses, and even individual instructors in order to add web technology to their courses and to supplement traditional face-to-face courses [5, 9]. Such e-learning systems are sometimes also known as learning management systems (LMSs) or course management systems (CMSs). These systems offer a great variety of channels and workspaces to facilitate information sharing and communication between participants in a course, to let educators distribute information to students, produce content material, prepare assignments and tests, engage in discussions, manage distance classes and enable collaborative learning with forums, chats, file storage areas, news services, etc. Some examples of such systems are Blackboard [1], WebCT [7], TopClass [6], Moodle [5], Ilias [4], Claroline [2], etc. Nowadays, one of the most commonly used is Moodle (Modular Object Oriented Developmental Learning Environment), which is a free learning management system that enables the creation of powerful, flexible and engaging online courses and experiences [8]. These e-learning systems accumulate a vast amount of information, which is very valuable for analyzing students behavior and could create a gold mine of educational data [24]. Learning management systems accumulate a great deal of log data about students activities. They can record whatever student activities are involved, such as reading, writing, taking tests, performing various tasks, and even communicating with peers. They also provide a database that stores all the system's information: personal information about the users (profile), academic results, user's interaction data, etc. However, due to the vast quantities of data these systems can generate daily, it is very difficult to manage manually and to extract useful information.

Organizations struggle to define the impact that emerging technologies have on learning, the delivery of educational systems, and on student performance. Massive open online courses (MOOCs), online classes, mobile devices, and other technologies hold as much promise as disruption. Higher education administrators and policy-makers continue to push for increased graduation rates, while students and families continue to value classroom size and course flexibility. Simultaneously, universities strive to raise completion rates while offering improved course enrollment, schedule flexibility, and hire and retain top-tier faculty. All of these forces converge and complicate the ability to make intelligent investments with a justifiable return on investment (ROI). The ability to innovate and lead among this environment requires finding approaches that improve assessments, feedback and student success while addressing the challenges of an increasingly virtual, dynamic student population.

In the last few years, researchers have begun to investigate various data analytics to help teachers improve e-learning systems [3]. Data analytics have also been applied to explore, visualize and analyze e-learning data in order to identify useful patterns [28, 31], and to evaluate web activity to get more objective feedback for teachers instruction and to find out more about how the students learn [23]. These methods allow us to discover new, interesting and useful knowledge based on students usage data. Furthermore, universities started applying data mining and predictive analytics to data to identify various measures of performance. Barber and Sharkey [13] analyzed data from the University of Phoenix, the largest online campus in the U.S. to develop and validate the utility of a logistic model to provide timely, valuable information to academic advisers.

In this paper, we present our efforts to identify critical parameters for designing a data-driven assessment and feedback system by conducting predictive analytics and time series analysis to academic data available in the LMS of a certain university through Moodle. The selected university offers baccalaureate, master, and doctoral degrees in virtually every field from medicine to business, law to liberal arts, and science and engineering to architecture. With \$3.5 billion in revenue and over 64 thousand enrolled students, this university is one of the largest public universities in the country. As the flagship of the state university system, it provides numerous methods for students and faculty to interact. Access to education, the ability to deliver high-quality outcomes, and optimizing student success are consistent themes among higher education administrations.

The intent of this research is also to validate the LMS data elements yielding signals of student performance, which will help in designing better assessments and feedback . We start by first identifying the data features useful for predicting student outcomes such as students scores in homework assignments, quizzes, exams, in addition to their activities in discussion forums and their total GPA at the same term they enrolled in the course. Then, we focus on developing autoregressive and moving average methods to estimate the relative stability of students grades in the course. This macro-approach helps to classify students that are in need of special assessments and require personalized feedback to improve their performance. The results show there is a greater disparity in learning from freshman to senior years in the college and educationist needs to focus on creating assessments that cater to the particular needs of the students at a desired level. Later, we use this understanding in Logistic regression [18] and feed-forward neural network [19] predictive models to identify students as early as possible that are in danger of failing the course they are currently enrolled in. These models compute the likelihood of any given student failing (or passing) the current course. The outcome is to prioritize students for intervention and referral to academic advisers. Numerical results are presented to evaluate and compare the performance of the developed models and their predictive accuracy.

The paper is structured as follows: Section 2 presents related work to understand past and current approaches in educational data mining. In Section 3, we present the data preparation process that selects and refines the LMS raw data before feeding them to the prediction models. In Section 4, we introduce the developed logistic regression and feed-forward neural network prediction models that identify students in danger of failing a course. In addition, we discuss the mathematical model used to assess the temporal stability of students' grades. Numerical results demonstrating the prediction accuracy of the models together with our findings and observations are presented in Section 5. Finally, Section 6 provides concluding remarks.

2. RELATED WORK

Educational data mining is becoming a main stream research for identifying approaches to improve the current standards of our educational systems. There are two main categories of educational data mining currently pursued; one focusing on school education and the other on university education. A learning management system, called Moodle [29], is a learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalized learning environments. It helps to create courses and store educational data of students on a longitudinal scale. This data can then be used for studying student learning abilities, creating adaptive models, and time series analysis for designing better forecasting, which can be used for example, in identifying early success and failure and related intervention models. The dataset that we have used in this work also came from Moodle learning management system.

Primarily, the focus of education data mining has been to create personalized lesson plans using adaptive learning techniques, early prediction of learning effects, and increasing engagements. In [15], the authors described an approach for developing adaptive electronic textbooks and present InterBook, an authoring tool based on this approach which simplifies the development of adaptive electronic textbooks on the Web. This work is one of the early stages of design and development of adaptive learning techniques. Several others [14, 17, 27] shown that adaptive learning provides important step towards enabling feedback based system for improving the educational standards. In [25], the authors developed a method to recommend courses that are suitable for a student. This has been becoming more pervasive in the university degrees, where students have many options, however, no many knows exact nature of the courses. Some studies [12, 32, 36], have used personal information such as demographic information, enrollment, previously obtained grades, number of previously taken courses and their classification, and grade pointer average to serve accurate recommendations. In [33], the authors have used a similar indicator to design a recommendation system that expects to predict correctly student results in approximately 80% of the cases. However, most of these works have been applied to datasets that are either small or work on the aggregate levels.

Several data mining applications are focused toward educators, where the object is to help create accurate feedback, categorization of learners based on their abilities, course creation, and instructional plans [10, 34]. Next, we discuss the Moodle dataset and the data preparation process.

3. DATA PREPARATION

To facilitate the design and delivery of online education, various LMSs and supporting tools emerged. Specifically, the selected university uses "Moodle" to administer, track, and deliver educational content to students, staff, and faculty. Dating back to 2009, the university records student classes, progress grades, test/quiz scores, and more in this system providing faculty and administrators a wealth of data to better understand the university performance.

Based on data provided by this university and literature about data mining in Moodle [29], useful courses for analysis were selected. Courses use a variety of scoring rubrics and values that make direct comparisons difficult. In order to establish a consistent data set, quality criteria were used to evaluate the LMS data. The data quality criteria honed the variety of courses contained in Moodle to a selection with the following criteria:

- Must have a reasonable number of modules (attributes)
- Must have a reasonable number of enrolled students
- Must have a reasonable distribution of grades
- Must be offered consecutively in Fall/Spring semesters

Also students with incomplete or missing data are excluded from further analysis. This data cleansing and transformation process triage the complete data to a set of courses ready for analysis.

Many Moodle data elements (attributes) were examined and several of them were selected for analysis such as students final course grade, most recent GPA, forum activities, and final grades for tests, assignments, quizzes, discussion questions, etc. The model used in [13] contained significant demographic features of students (gender, age, financial aid status, military status). These data features are not present in Moodle and are not considered as sources or features for this research. Now, after selecting the appropriate courses and data features for analysis, we need to categorize the final grades in these courses. The final grades for tests, assignments, quizzes, etc. are transformed into six categories:

1. 90% and above is category 6
2. 80% to 89% is category 5
3. 70% to 79% is category 4
4. 60% to 69% is category 3
5. 50% to 59% is category 2
6. and below 50% is category 1

The process of categorizing the data should improve the prediction accuracy of the models.

A success (recorded as "S") is defined as a student receiving a final course grade that exceeds his/her GPA at the time he/she enrolled in the course. While a Failure (recorded as "F") is defined as a final course grade lower than the GPA at the time of enrollment. Table 1 shows a random sample of the data preparation conducted for all students.

4. BUILDING THE MODELS

We start by discussing the theory to understand the temporal stability of students grades in various math courses from freshman to senior years. Later, two prediction models are selected to train and test against the data. Both logistic (logit) regression [18] and feed-forward neural network (NN) [19] models are used to identify students in danger of failing a course.

4.1 Time Series Analysis Model

In this section, we discuss the theory used to model and characterize fluctuation and temporal stability in students grades. A stationary time series is one for which the probabilistic behavior P of every collection of values

$$x_{t_1}, x_{t_2}, ...x_{t_k} \tag{1}$$

such that

$$P\{x_{t_1} \leq c_1, ..., x_{t_k} \leq c_k\} = P\{x_{t_{1+h}} \leq c_1, ..., x_{t_{k+h}} \leq c_k\} \tag{2}$$

for all $k = 1, 2, ...,$ with times $t_1, t_2..$ and constants $c_1, c_2, ...$ However, such a definition is too strong in reality to be implemented, and therefore we focus on modeling the time conditions for first few moments of the series. In this work, we investigate the nature of the dependency that is influenced by the past values of the grades to understand the relative stability of fluctuation of grades during the rest of the semester. This kind of study will greatly benefit in identifying the average vulnerability among students in different years of college. In order to achieve that, we use the Autoregressive model that is based on the idea that current value of the series is based on the past values, which in this case is based on the previous grades. This model takes the following form in our case,

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + ...\phi_p x_{t-p} + w_t \tag{3}$$

where x_t is stationary, and $\phi_1...$ are constants. We assume that w_t is a Gaussian white noise series with mean zero and variance σ_w^2, unless otherwise stated. However, sometimes the stability of grades is difficult to model since the regressor $x_{t-1}, ...x_{t-p}$ have random components (such as missing an exam or not submitting an assignment) . So using the backshift operator, we define our autoregressive operator model as

$$(1 - \phi_1 * B - \phi_2 * B^2 - ...\phi_p * B^p)x_t = w_t \tag{4}$$

The first order model, will be given by $x_t = \phi x_{t-1} + w_t$. For a set of k different assignments of a course, we get

$$x_t = \phi x_{t-1} + w_t = \phi(\phi x_{t-2} + w_{t-1})w_t = \phi^2 x_{t-2} + \phi w_{t-1} + w_t \tag{5}$$

$$x_t = \phi^k x_{t-k} + \sum_{j=0}^{k-1} \phi^j w_{t-j} \tag{6}$$

The mean of the stationary process can be computed as

$$E(x_t) = \sum_{j=0}^{\infty} \phi^j E(w_{t-j}) = 0 \tag{7}$$

The auto-covariance function that measures the fluctuation will then be represented using the recursion as

Table 1: Random sample of the conducted data preparation process

Week	Item Name	Grade	Grade Max	Grade %	Grade Category	Course Grade	Course Grade (in Points)	GPA (Term)	Success / Failure
Week 1	Chapter 1 Homework	44	50	88	5	C	2	2.551	F
	Chapter 1 Quiz	8	10	80	5				
	Chapter 1 Discussion Question	0	5	0	1				
Week 2	Chapter 2 Homework	50	50	100	6				
	Chapter 2 Quiz	6	10	60	3				
	Chapter 2 Discussion Question	0	5	0	1				
Week 3	Chapter 3 Homework	30	50	60	3				
	Chapter 3 Quiz	6	10	60	3				
	Chapter 3 Discussion Question	0	5	0	1				

$$\gamma(h) = cov(x_{t+h}, x_t)$$

$$= E\left[\left(\sum_{j=0}^{\infty} \phi^j w_{t+h-j}\right)\left(\sum_{k=0}^{\infty} \phi^k w_{t+k}\right)\right]$$

$$= E\left[\left(w_{t+h} + ... + \phi^{h+1} w_{t-1} + ...\right)(w_t + \phi w_{t-1} + ...)\right]$$

$$= \sigma_w^2 \sum_{j=0}^{\infty} \phi^{h+j}\phi^j = \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma_w^2 \phi^h}{1-\phi^2}, h \geq 0. \quad (8)$$

We will apply this theoretical approach to evaluate the stability of grades and model the temporal fluctuation among students enrolled in mathematical courses.

4.2 Logistic Regression Model

The Binary Logit Model is a model based upon the cumulative logistic probability function that defines a successful event (Binary Response) as a 1 and a failure as a 0 [26]. The successful event that a student improves for the final course grade in any assessment area means that the student will receive a score that exceeds his/her baseline GPA. The successful event will evaluate to a Binary Response as one and the failure to perform above the baseline GPA will evaluate to a Binary Response of zero. The odds that a students outcome will result in a success is $p/(1-p)$, where p is the success probability [11]. The resulting logistic prediction models for the first three weeks of a course are given as:

- Week 1: $\text{Log}(p/(1-p)) = B_0 + B_1 * Homework\_W1 + B_2 * Quiz\_W1 + B_3 * DiscussionQuestion\_W1$
- Week 2: $\text{Log}(p/(1-p)) = B_0 + B_1 * Homework\_W1 + B_2 * Quiz\_W1 + B_3 * DiscussionQuestion\_W1 + B_4 * Homework\_W2 + B_5 * Quiz\_W2 + B_6 * DiscussionQuestion\_W2$
- Week 3: $\text{Log}(p/(1-p)) = B_0 + B_1 * Homework\_W1 + B_2 * Quiz\_W1 + B_3 * DiscussionQuestion\_W1 + B_4 * Homework\_W2 + B_5 * Quiz\_W2 + B_6 * DiscussionQuestion\_W2 + B_7 * Homework\_W3 + B_8 * Quiz\_W3 + B_9 * DiscussionQuestion\_W3$

where the independent variable (predictor) $Homework\_W1$ is the categorized homework grade in week 1, and the other predictors are defined similarly. From Week 2 and Week 3 models, notice that the available information from prior weeks are included in the model. The regression coefficients, $B_0, ..., B_9$, are estimated using Maximum Likelihood estimation [22]. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximizes the likelihood function, so the iteratively Weighted Least Squares method is used instead.

4.3 Neural Network Model

There are numerous different types of neural network paradigms that have been proposed for distinct problem domains [21, 20, 35]. An appropriate neural model that has been previously used for forecasting, prediction, and general decision making, has been the multi-layered feed-forward perceptron model. Multi-layered networks have continuously-valued neurons or processing elements, are trained in a supervised manner, and consist of one or more layers of nodes (called hidden layers) between the input and output nodes [20]. Input nodes are where information is presented to the network, output node(s) provide the decision made by the neural network, and the hidden nodes, in essence, contain the information regarding proper mapping of inputs to proper decisions (outputs). The developed feed-forward NN prediction model consists of 3 input nodes (for Week 1), 10 nodes in the hidden layer, and one output node.

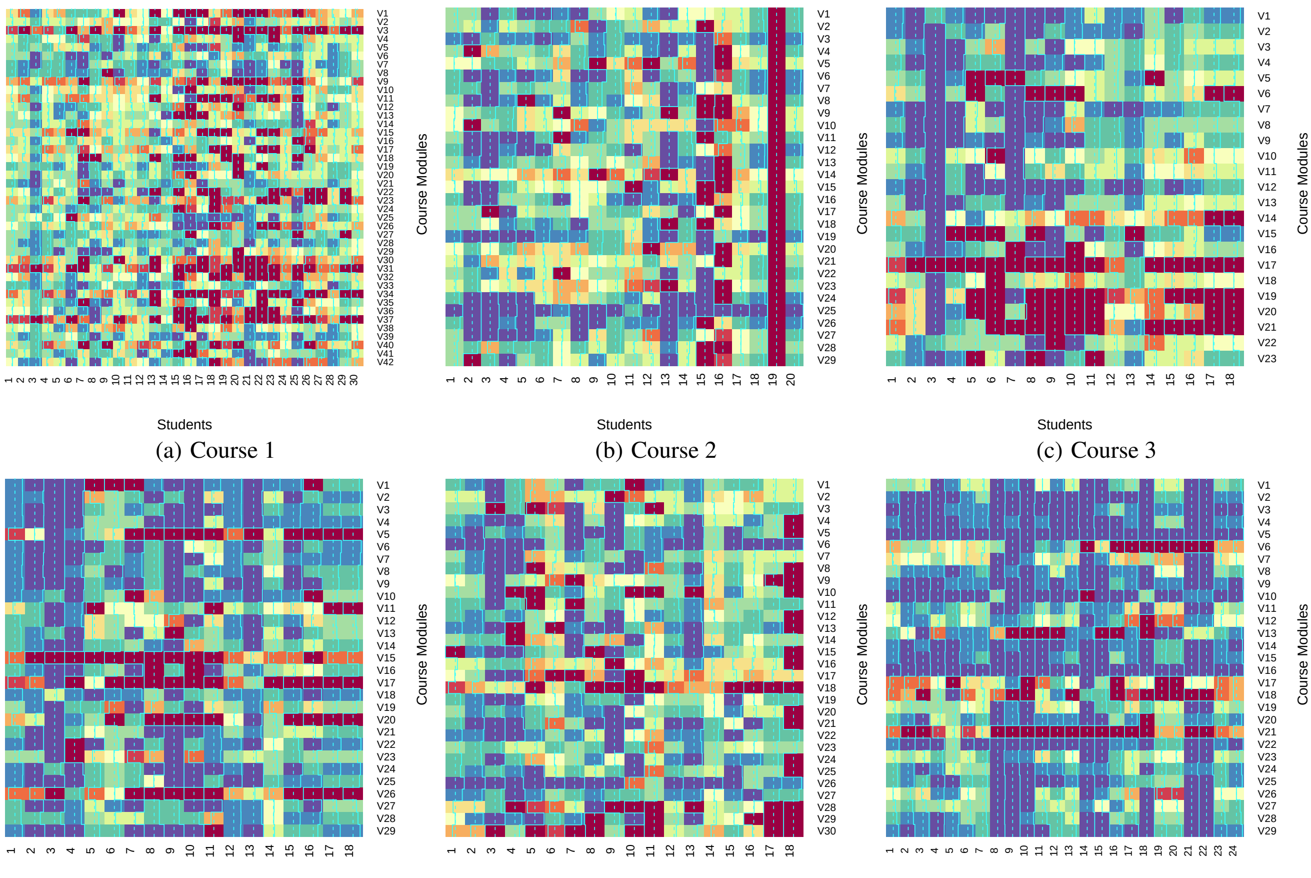
Back-propagation algorithm [30] is used for model development (training set) to estimate the model coefficients. It is an iterative gradient-descent algorithm designed to minimize the mean squared error between the actual output of a node and the desired output as specified in the training set. Weight adjustment starts at the output nodes where the error measure is readily available, and then proceeds by propagating the error measure back through the layers toward the input nodes. More detailed information regarding neural networks can be found in summary papers such as [21, 30, 35].

5. FINDINGS AND OBSERVATIONS

Our observations are divided into two parts. First we present results and findings for the time series analysis to discover and characterize the relative fluctuation in students' grades. Later, we illustrate the results and findings for the logit and neural network models that are used for predicting students successes.

5.1 Temporal Analysis and Stability Measurement

In this section, we apply the time series analysis that was discussed in the section 4.1 to discover and characterize the relative fluctuation in students' grades. We characterize this fluctuation to classify the courses and help devise better predictive and assessment mechanisms. We have analyzed 260,698 records of more than 11,000 students who were enrolled in 270 mathematical courses (including the course that was offered more than once during 2009-2013). These math courses were offered from freshmen to senior level at the university. Our main objective would be to identify and characterize grade fluctuations during the freshman to the senior years at the university. A fluctuation in grade of a student is defined as a relative change in the recent grade to the historic one, thereby impacting the relative ranking of students in class. For example, Alex got 93% in first three quizzes but lagged in fourth one and got only 71%; somehow improved his performance in the fifth quiz and got 85%. Such fluctuations make assessments very chal-



(a) Course 1 (b) Course 2 (c) Course 3

(d) Course 4 (e) Course 5 (f) Course 6

Figure 1: A sample of math courses offered at the *freshman* level in the university. Heatmaps (Red spectrum shows higher grades) show relative ranking of the students, which highly changes in the course during the entire semester. This indicates challenges in designing accurate prediction and assessment models.

lenging and hence there is an immediate need to characterize such variations. We choose mathematical courses (e.g. Algebraic and calculus courses) as they are exhaustively quantitative and provide a solid ground for analyzing the changes at a level of higher granularity.

An example hypothesis would be to validate the math courses that are offered at freshman (or sophomore) levels, which show higher grade fluctuation than the math courses offered at the senior level in the university (This assumption is based on the influence of external factors on freshman such as lack of social circle and adjusting to new environment). Our approach of modeling and characterizing temporal stability of grades among university students will aid creating better forecasting mechanism.

The math courses such as PreCalculus I and II enrolls freshmen, while advanced courses such as Real Analysis are offered at the senior level. Each of these courses have at least 18 modules, such as quiz 1, 2, 3, ..,n, midterms, and finals. Our model monitors relative fluctuation and weighs student among one-another at the completion of every module. Then, it calculates the moving grade and relative standing of the students in the class. These progressing grades are used to characterize the fluctuation. Our analysis has shown that most of the freshmen and sophomore level math courses are very unstable and reflect high fluctuation in the students grade over the course of entire semester.

Our model finds that these students' grade vary a lot and therefore their ranking in the course changes remarkably throughout the semester. This result indicates a challenge in forecasting courses which are offered at the freshman/sophomore level in the university. The results of six sample courses are shown in a HeatMap in Figure 1. The x-axis shows the students enrolled in this course and the y-axis shows the course modules (quiz, exams, and finals). The blocks in the Heatmap reflect the relative grade of students for a set of modules in the course. A color band that is relatively constant for students in all modules show their grades' stability and steady

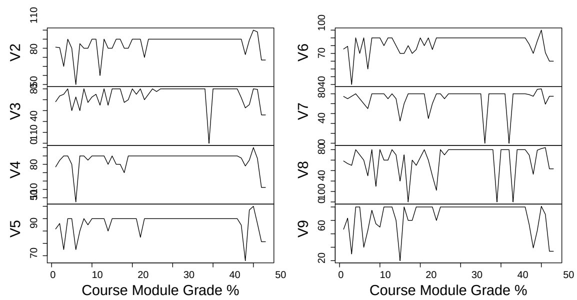


Figure 2: Stability of grade and smooth progression in *senior* level courses.

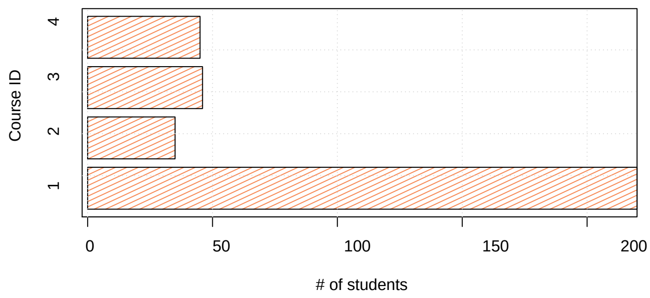


Figure 4: Students sample for modeling and prediction.

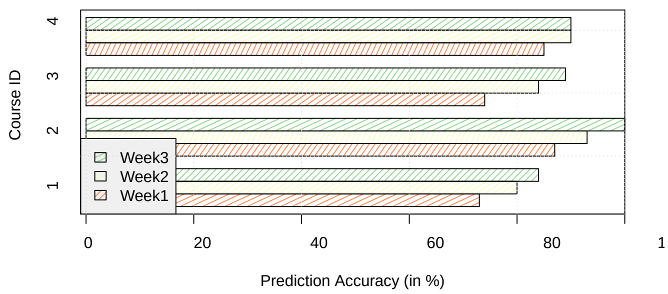


Figure 5: Prediction accuracy of Logistic Regression model.

progression in the course. This knowledge certainly increases the prediction accuracy of student success or failure.

In Figure 3, we show the relative stability of students grades in six senior level math courses. We found the majority of courses have students with high level of predictability in the later years of the university education. The frequency analysis also shows the stability. A set of time-series is plotted in Figure 2 to demonstrate the relative stability (even the relative grades are dropping for all the students) in a senior level mathematics course.

This shows that students require greater intervention plans and close monitoring during their initial years at the university, despite the courses they enrolled in are relatively easier. We also find that students become more and more responsible and focused at the senior level in the university. Their grades are stables and would be more predictable. We believe that our examination of course grade stability will result in creating better assessments and adaptive feedback systems. Next, we use this information and discuss the prediction mechanisms.

5.2 Modeling and Prediction

Approximately half of the dataset is used for model development, and the other half is used for model validation. Figure 4 shows a random sample of students selected for modeling and prediction. We display the prediction performance results of the Logit and NN models in Figures 5 and 6 respectively. For courses fitting the criteria selected, success/failure can be predicted with at least eighty-four percent (84%) accuracy by week 3. For each model, prediction accuracy increased with time, and without a decrease in performance as the sample size (number of students enrolled) grew. Numerically, results indicate that the NN model performance is comparable to the logistic regression. The neural network models do hold some advantages. The NN model performs better with large numbers of variables which enables the addition of future data features to the existing model. As the use, collection, and quality of data in the LMS improves, data features can be examined for fit and impact on successful student outcomes using the existing model.

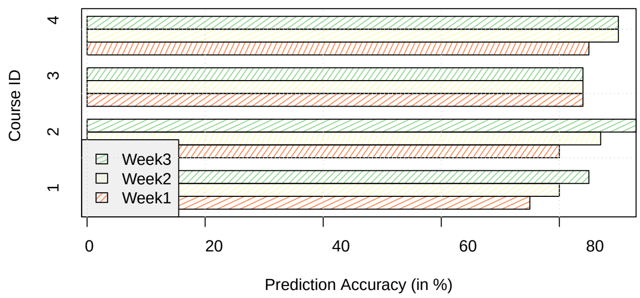


Figure 6: Prediction accuracy of Feed-Forward Neural Network model.

Table 2: Estimated coefficients show that homework has more impact on outcomes

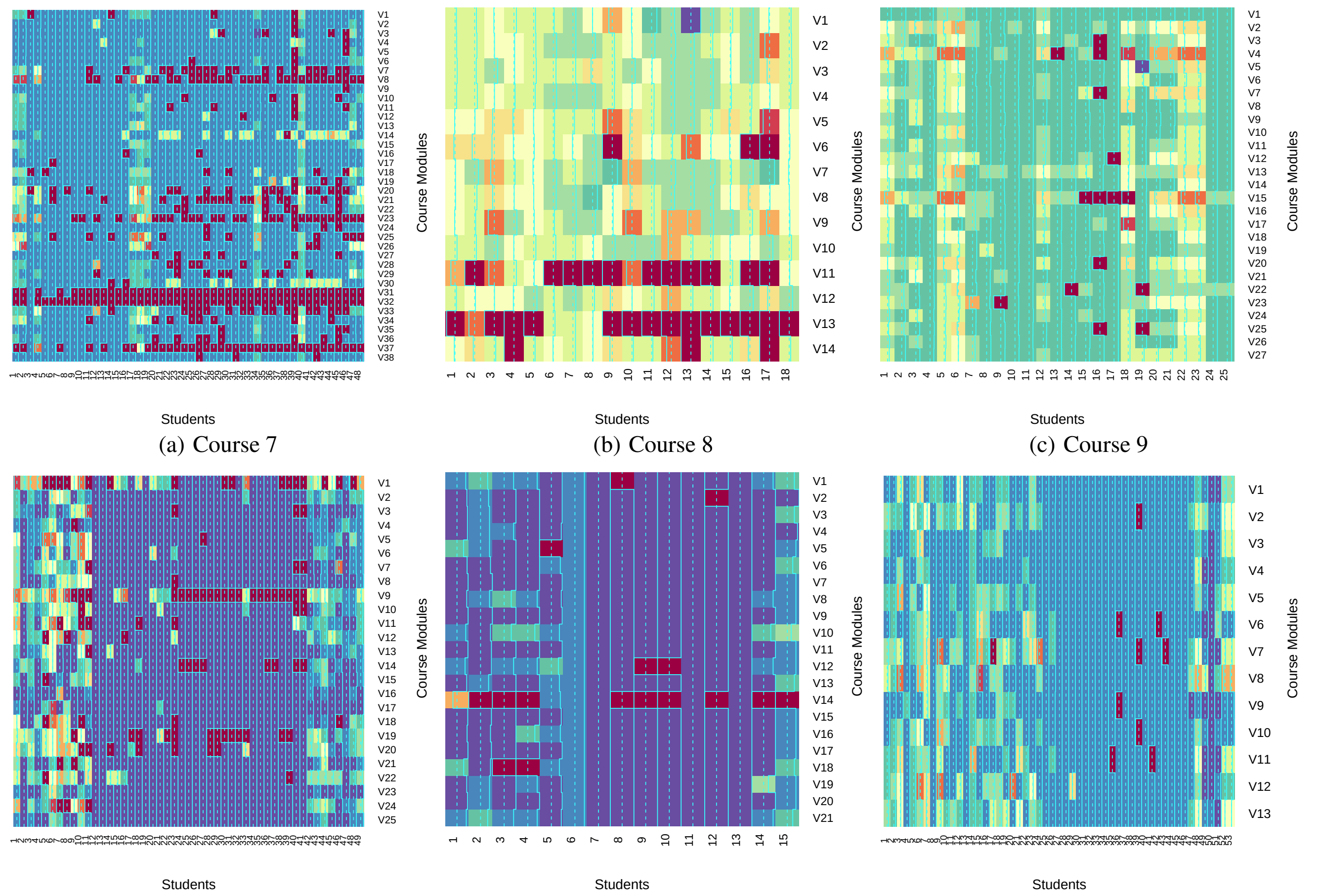
Attribute	Week 1	Week 2	Week 3
HW 1	7.0176	10.5828	5.7513
Quiz 1	0.5291	2.9466	3.3499
Discussion Q 1	0.2120	0.0507	2.9892
HW 2		10.4699	5.9732
Quiz 2		4.7673	4.3514
Discussion Q 2		4.8074	5.1734
HW 3			5.8005
Quiz 3			4.0474
Discussion Q 3			2.1235

The neural network model also performs better when data violates a normal distribution. By plotting the distribution of the final grades in dataset, we found out many course grades do not fit the same pattern as normal distribution. The work in [19] shows that regression analysis and traditional techniques used to predict student success (for admittance purposes in graduate programs) are not particularly effective. Therefore, the Feed Forward NN model was used for comparison to address any drawbacks. The primary constraint of the NN models is that they require more computation than logistic models.

We also examined the contribution of individual predictors by examining the regression coefficients. In logistic regression, the regression coefficients represent the change in the logit for each unit change in the predictor. Table 2 shows the absolute value of estimated coefficients for a course selected at random. It is observed from Table 2 that "homework" is a leading indicator of student performance. The magnitude of the coefficient indicates its contribution on the predicted value (dependent variable).

Since "homework" was shown as a strong indicator of success/failure, we went back and looked at the course rubric and found out that posts to forums were used to discuss, share, and respond to homework. Thus, the data was re-categorized to include the total number of posts for all students in the course discussion forums. This was used as an additional input variable (predictor) in the previously developed prediction models to investigate if "posts" aid in the prediction of success or failure. A similar method was applied to transform the number of posts into three categories:

- High: Number of posts above the 75th percentile of the sample



(a) Course 7 (b) Course 8 (c) Course 9

(d) Course 10 (e) Course 11 (f) Course 12

Figure 3: A sample of math courses offered at *senior* level in the university. Heatmaps (Red spectrum shows higher grades) show relative standing of the students, which are constant in the class during the entire semester, thereby indicating easier to predict and assess.

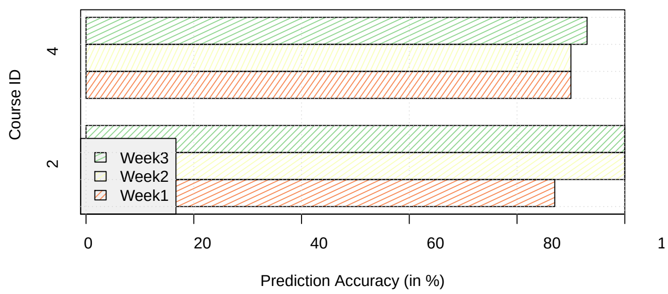


Figure 7: Prediction accuracy of Logistic Regression with "posts"

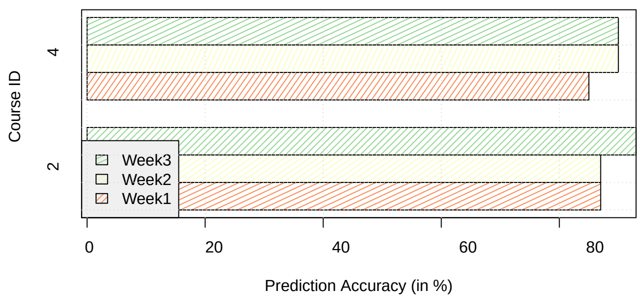


Figure 8: Prediction accuracy of Feed-Forward Neural Network with "posts"

- Medium: Number of posts between the 25th and 75th percentile of the sample
- Low: Number of posts below the 25th percentile of the sample.

In Figures 7 and 8, we show the prediction accuracy results for the Logit and NN models respectively, after including number of posts in discussion forums as an additional input variable. It is observed that the prediction accuracy of both models improve when posts are included in the models.

6. CONCLUSION

Data miming can be used to explore and identify unknown patterns in any data. The knowledge gained is useful when applied to drive improvements in specific processes. Mining the university's learning management system delivers useful insight regarding students, faculty, and administration. Data mining the LMS should not serve to prescribe actions or behaviors to the student or faculty, but can be used to augment the learning management system as a notification that a student needs more interaction.

Analysis shows that for a given rubric student success or failure can be predicted with a high degree of confidence within the first few weeks of the course. The rubrics define which modules are given weight in student grades (for the courses selected: homeworks, quizzes, discussions, forum activities were graded as part of the rubric). The ability to design courses that support predictive analytics and alert faculty to student struggles earlier in their coursework is an obtainable goal. In addition, there is a great disparity among the freshman/sophomore and senior students' grades, requiring special intervention plans for the former. The interventions will help students in early career irrespective of course rigor to perform significantly well. Pragmatically speaking, the macro environment surrounding students creates too many factors to predict student success and number of assessments. Time and again educational systems produce students able to rise above conditions that seem destined for failure. These potential insights favor questions about why students do things. The use of learning management systems, digital classrooms, and online learning environments increases the amount and type of data available to better understand when, where and how students are doing.

7. REFERENCES

[1] Blackboard.

[2] Claroline.

[3] Data Mining in E-learning (Advances in Management Information): C. Romero (Editor), C. Romero, S. Ventura: 9781845641528: Amazon.com: Books.

[4] Ilias.

[5] Moodle.

[6] TopClass.

[7] WebCT.

[8] *Moodle E-Learning Course Development: A complete guide to successful learning using Moodle*. Packt Publ., 2006.

[9] *Using Moodle: Teaching with the Popular Open Source Course Management System* . O'Reilly Media; Second Edition edition, 2007.

[10] *Handbook of Educational Data Mining*. CRC Press, 2011.

[11] *Principles of Econometrics* . Wiley 4 E, 2011.

[12] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst.*, 23(1):103–145, Jan. 2005.

[13] R. Barber and M. Sharkey. Course correction. In *Proced. of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, page 259, New York, Apr. 2012.

[14] R. I. Brafman, C. Domshlak, and S. E. Shimony. Qualitative decision making in adaptive presentation of structured information. *ACM Trans. Inf. Syst.*, pages 503–539, 2004.

[15] P. Brusilovsky, J. Eklund, and E. Schwarz. Web-based education for all: a tool for development adaptive courseware. *Computer Networks and ISDN Systems*, 1998.

[16] P. Brusilovsky and C. Peylo. Adaptive and Intelligent Web-based Educational Systems. *International Journal of Artificial Intelligence in Education*, 13(2-4):159–172, Apr. 2003.

[17] P. De Bra, P. Brusilovsky, and G.-J. Houben. Adaptive hypermedia: From systems to framework. *ACM Comput. Surv.*, 1999.

[18] G. Garman. A Logistic Approach To Predicting Student Success In Online Database Courses, July 2011.

[19] B. C. Hardgrave, R. L. Wilson, and K. A. Walstrom. Predicting graduate student success: A comparison of neural networks and traditional techniques. *Computers & Operations Research*, 21(3):249–263, Mar. 1994.

[20] R. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4(2):4–22, 1987.

[21] E. Másson and Y.-J. Wang. Introduction to computation and learning in artificial neural networks. *European Journal of Operational Research*, 47(1):1–28, July 1990.

[22] S. Menard. *Applied Logistic Regression Analysis, Volume 106; Volume 2002ă(Google eBook)*. SAGE, 2002.

[23] E. Mor and J. Minguillón. E-learning personalization based on itineraries and long-term navigational behavior. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters - WWW Alt. '04*, page 264, New York, New York, USA, May 2004. ACM.

[24] J. Mostow, J. Beck, H. Cen, A. Cuneo, E. Gouvea, and C. Heiner. An Educational Data Mining Tool to Browse Tutor-Student Interactions:Time Will Tell! 2005.

[25] K. Nishino, Y. Iribe, S. Mizuno, K. Aoki, and Y. Fukumura. The development of a course recommendation system for e-learning students. *Int. J. Knowl. Web Intell.*, 2012.

[26] R. S. Pindyck and D. L. Rubinfeld. *Econometric models and economic forecasts*. McGraw-Hill New York, 1981.

[27] B. Raufi and J. Georgieva. Adaptive web-based systems: From framework to implementation. In *Proceedings of the 11th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing on International Conference on Computer Systems and Technologies*, NY, USA, 2010.

[28] C. M. Riccardo Mazza. Exploring usage analysis in learning systems : Gaining insights from visualisations.

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

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. pages 318–362, Jan. 1986.

[31] L. Talavera and E. Gaudioso. Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *Workshop on AI in CSCL*, pages 17–23, 2004.

[32] A. Umyarov and A. Tuzhilin. Leveraging aggregate ratings for better recommendations. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 161–164, New York, NY, USA, 2007. ACM.

[33] C. Vialardi, J. Bravo, L. Shafti, and A. Ortigosa. Recommendation in Higher Education Using Data Mining Techniques. *International Working Group on Educational Data Mining*, June 2009.

[34] E. S. W. Hämäläinen, T. H. Laine. Data mining in personalizing distance education courses.

[35] J. Wang and B. Malakooti. A feedforward neural network for multiple criteria decision making. *Computers & Operations Research*, 19(2):151–167, Feb. 1992.

[36] Z. Zhang, S. Shang, S. R. Kulkarni, and P. Hui. Improving augmented reality using recommender systems. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, NY, USA, 2013.