PREDICTIVE MODELING TO FORECAST STUDENT OUTCOMES AND DRIVE EFFECTIVE INTERVENTIONS IN ONLINE COMMUNITY COLLEGE COURSES

Vernon C. Smith, Ph.D.
MyCollege Foundation

Adam Lange
Ellucian

Daniel R. Huston, M.A.
Rio Salado College

ABSTRACT

Community colleges continue to experience growth in online courses. This growth reflects the need to increase the numbers of students who complete certificates or degrees. Retaining online students, not to mention assuring their success, is a challenge that must be addressed through practical institutional responses. By leveraging existing student information, higher education institutions can build statistical models, or learning analytics, to forecast student outcomes. This is a case study from a community college utilizing learning analytics and the development of predictive models to identify at-risk students based on dozens of key variables.

KEYWORDS

online learning, learning analytics, predictive modeling, community colleges, risk levels for online students, faculty

I. INTRODUCTION

The recent demands for increased student success efforts, most notably from the American Graduation Initiative [1], have come as colleges are facing limited funding and resources. The Associated Press [2] and the Center on Budget and Policy Priorities [3] have both recently highlighted the drastic budget cuts implemented by community colleges nationwide. Colleges are, therefore, charged with the task of developing new ways to respond to student needs with personalized, timely outreach efforts while simultaneously being mindful of resource limitations. Higher education has seen a surge of interest in data mining and predictive modeling methods over the past few years. These "learning analytics" have been made possible by capturing and utilizing the large amounts of information collected within campus enterprise systems, most notably the learning management system, to aid in teaching and learning initiatives. Since 2008, Rio Salado College has researched data mining and predictive modeling as viable tools to provide proactive outreach to students using systematic and sustainable methods. There is a growing precedent for this approach, as a handful of institutions have recently implemented similar ideas [4], with limited numbers of community colleges utilizing learning analytics and predictive modeling. The 2011 *Horizon Report* produced by New Media Consortium and the EDUCAUSE Learning Initiative indicated that learning analytics will have a significant impact and forecasts its wide-scale adoption within the next five years [5]. Given the this surge of

interest that is often adopted and replicated across higher education and community colleges as a sector, the publication of research and cases becomes critical to developing theory, practice, and institutional responses to increase student success through learner analytics and predictive modeling.

A. Related Literature

There has been a surge of interest in data mining and predictive modeling methods in higher education. The general trend in this research has been the idea of utilizing the large amounts of information collected within campus systems, most notably the learning management system, to aid in teaching and learning initiatives. Purdue University presented early research on "academic analytics" as a tool to predict at-risk students [6, 7]. Data mining has been applied in the study and analysis of student assessment data [8]. A limited number of institutions, including Rio Salado College, have presented similar applications of these approaches in the higher education sector [9, 10, 4]. Rio Salado has engaged in data mining and predictive modeling research since 2008 [11]. Formal studies have also demonstrated the statistical correlation between LMS activity markers and course outcomes [12].

B. Background

Rio Salado College (Rio) is located in Tempe, Arizona and is one of the ten colleges in the Maricopa County Community College District. During the inception of this study in 2008-2009, Rio had a total unduplicated headcount of 61,340, with over half of those students taking courses in the online setting. Online courses are characterized by an asynchronous, individualized approach to instructional design and course navigation. Most courses at Rio are offered every Monday, for 48 starts annually. Rio creates additional flexibility for the online student population by offering these weekly courses in an accelerated eight-week format. Online courses are delivered to students using RioLearn, a proprietary learning management system (LMS) that was built to serve over 100,000 students in partnership with Microsoft and Dell Computers. This LMS, coupled with an unbundled faculty model where the faculty role has been disaggregated, relying on over 1,300 adjunct faculty, has created a high level of scalability that has been recognized as a national model for productivity [13]. Scalability and increased online enrollments also create the need to follow the behavior of students to identify students at risk, and to build interventions to promote online student success.

II. RESEARCH

The purpose of predictive modeling research at Rio Salado College was to identify the factors that demonstrated significant statistical correlations with final course outcomes in online courses, in other words, could we identify the factors that led to online success? We chose to define a successful outcome as a final letter grade of "C" or higher and an unsuccessful outcome as a final letter grade below "C," including withdrawals. Furthermore, we sought to develop practical predictive models capable of forecasting course outcomes using a subset of the significant factors. In an effort to develop models that could serve as early alert systems, we also set out to determine the earliest point in the course at which accurate predictions were possible. We also recognized that many factors, such as "points earned," change frequently throughout the course, meaning that a model capable of delivering real time or near-real time predictions would be ideal. To that end, we sought to create a model that would update weekly, every Monday. This research was driven by the long-term objective to facilitate and strengthen the linkages between instructors and at-risk students within a dynamic online environment.

As noted in the previous section, this research aimed to answer three primary questions.

1. Which factors are effective as early or point-in-time predictors of course outcome in online learning?
2. How can community colleges utilize those factors in predictive models to forecast student outcomes?
3. How can community colleges utilize predictive models to implement proactive, effective, and sustainable student interventions with the goal of improving student success rates within the online learning population?

III. METHODOLOGY

A. Data collection

Every online course offered by Rio Salado College varies in terms of requirements, pace, points possible, difficulty, and student demographics. For this reason, we adopted the approach of building course-specific models capable of reflecting the variation that exists among courses. This approach has also recently been proposed by Macfadyen [12]. For simplicity, we will mostly focus on *one* online freshman-level accounting course. The course under study was offered in a fully online format, meaning that no in-person learning component was required.

1. Population selected for analysis

The data used for this study was collected from a freshman-level accounting course during the summer and fall 2009 semesters, which included students with course start dates between the months of May and December. Only fully online sections were included and special populations, such as high school dual enrollment students, were excluded. Only those who received a grade, including withdrawals, were included in the final data set. Students who dropped their course during the drop/add timeframe were excluded, as they did not receive a final grade. The total sample size was N = 539 students.

2. Data sources

Information pertaining to online student activity and assignment grades was obtained from RioLearn. Information pertaining to enrollments, including final grades, was obtained from PeopleSoft, the district-wide student information system used by the Maricopa County Community College District. Data from PeopleSoft and RioLearn were available from an on-site Microsoft SQL Server database. A common set of unique identifiers existed in both systems, thus making it possible to combine information.

3. Assignment grades

The RioLearn database contains all relevant information related to student assignments. Information is stored showing when assignments were submitted, when assignments were graded, and grades earned.

1	Logged in to RioLearn system
2	Logged in to course section homepage
3	Viewed the course syllabus
4	Opened an assessment
5	Completed an assessment
6	Viewed grade book comments from instructor
7	Viewed assessment feedback from instructor
8	Opened a lesson
9	Requested due date change
10	Selected a custom calendar option

Table 1: Some of the student activities logged in the RioLearn LMS.

4. LMS activity tracking and definitions

RioLearn maintains an activity log to document student activity within the system. A list of the logged student activities relevant to this research are shown in Table 1. We determined that a "log-in" should be defined using Activity 2, that is, logging in to the course section homepage. Activity 2 shows interaction with a specific course, as opposed to Activity 1 which shows interaction with the RioLearn system in general. We also determined that activities 3 through 10 should be referred to as "site engagement" activities as they suggest at least some engagement with a course.

5. Activity weighting

When analyzing LMS activity at multiple points in time throughout the duration of a course, we propose that recent behavior should be viewed with more significance than past behavior. For example, when analyzing student activity on the tenth week of a fourteen week class, information relevant to a student's current performance is more likely to be discovered over the last two or three weeks of logged activity, as opposed to activity that occurred several weeks or months in the past. In keeping with this philosophy, we calculated the frequency of student activities as weighted sums based on when each activity occurred relative to the course start and end dates. We refer to this weight as the 'occurrence point.' The weighted sum of any series of discrete LMS activity events is shown in Equation 1, where Δd_i represents the displacement (in days) of the i^{th} log-in event from the course start date and d_0 represents the length of the course (in days).

$$W(a_1, a_2, \ldots, a_n) = \sum_{i=1}^{n} \min\left(\frac{\Delta d_i}{d_0}, 1.0\right)$$

Figure 1. Weighted sum of n discrete LMS activities

6. Standardized course length

Rio Salado College offers online courses in many different calendar formats. Some courses may have traditional 14 to 16 week offerings during the fall and spring semesters with equivalent, but shortened versions offered in the summer semesters. Additionally, some courses offer an eight week calendar option that students can select within RioLearn. Since, for example, the fourth week of an eight week course is not the same as the fourth week of a fourteen week course, all course lengths (measured in weeks) were converted to a standardized 16-unit scale to support homogenous comparisons within variable-length courses. A 16-unit scale was selected because most all online course offerings at Rio Salado College are sixteen weeks or less. The equation used to map course lengths to the standardized scale is shown in Equation 2, where Δw represents the current displacement (in weeks) from the course start date and w_0 represents the length of the course (in weeks). The $\lceil \ldots \rceil$ brackets represent the Ceiling function [14].

$$Week_{Scaled} = \left\lceil \frac{\min\left(\Delta w, w_0\right)/_{w_0}}{16^{-1}} \right\rceil$$

Figure 2. 16 unit scaled course length

B. Naïve Bayes classification model

The naïve Bayes classification method was selected to predict at-risk students. Naïve Bayes is a probabilistic classification method with a long history of research and application. The method is commonly cited for its accuracy, robustness, and efficiency [15, 16]. A naïve Bayes model was employed to generate estimated probabilities of course success, which were then mapped to one of three warning levels.

Other classification methods, including decision tree and nearest neighbor, were investigated and tested. Ultimately, the naïve Bayes classification model was chosen because it offered significant advantages in several key areas compared to other methods. For instance, the naïve Bayes algorithm is computationally inexpensive, which is an important parameter due to the large student population at Rio Salado College. Also, Naïve Bayes is very scalable, meaning that the addition of more students or input variables will not cause a dramatic reduction in performance. Finally, as mentioned previously, naïve Bayes has demonstrated a strong record of accuracy in a variety of domains over many years of academic research [15, 16].

1. Warning levels

A simple three-level warning indicator was used to provide a user-friendly abstraction of the estimated probability of course success—meaning a grade of "C" or better. The warning levels were labeled as Low, Moderate, and High. A High warning level indicated that a student had a low probability of success if his/her current trajectory did not change. A Low warning level indicated that the student had a high probability of success and Moderate indicated that the student was borderline. Students with an estimated probability of success below 30% were determined to have a High warning level. Students above 70% had a Low warning level and students between 30% and 70% had a Moderate warning level.

2. Model validation

The random sub-sampling cross-validation method [17] was used to test the accuracy of the predictive models. This form of validation runs a model multiple times in randomized simulation environments and averages the results. The primary objective of validation was to confirm that the predicted warning levels were correlated with course success rate. In theory, an accurate model would assign warning levels such that a high success rate would be observed in the Low group, a low success rate would be observed in the High group, and the Moderate group would fall in the middle. This was the desired result because it would indicate that the warning levels provided a truly accurate depiction of where students fell on the at-risk spectrum.

3. Activity performance metrics

In order to provide additional, more specific diagnostic information outside of the derived warning levels, we proposed a simple method of assessing student LMS activity. These activity performance metrics were calculated independently of the warning levels, but were intended to accompany the warning levels when displayed to an end-user. We theorized that instructors might be able to launch more customized interventions for at-risk students if they had information showing student performance within specific LMS activities.

In assessing LMS activity, we adopted a "more is better" philosophy. That is, the more a student performs a certain activity, the better they will be scored in that particular area. We will justify this approach further on in the discussion of the positive correlation between LMS activity and course success.

Our proposed method functioned by comparing a student's weighted sum of LMS activity to the historical mean and standard deviation calculated for that particular activity. Means and standard deviations were calculated separately for every (scaled) week in every course using only the successful completers. This methodology was in keeping with our approach of course-specific analysis as mentioned in our data collection definition to obtain our objective of weekly updates. If a student's weighted sum of activity was within one standard deviation of the historical mean, a "Good" label was assigned. "Below average" was assigned if the student was below one standard deviation and "Above average" was assigned if the student was above one standard deviation. The actual label name varied, depending on the activity and the context.

C. Investigation of Factors

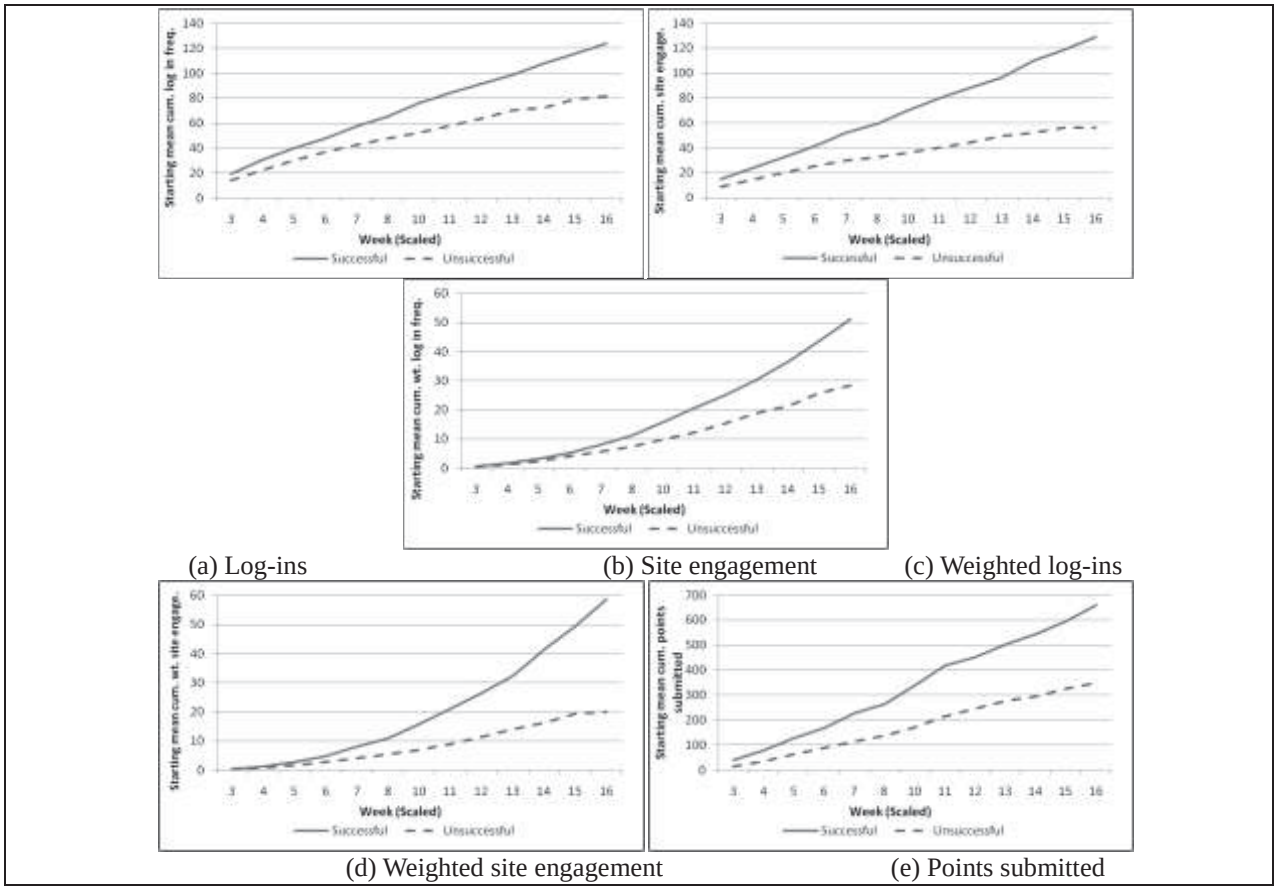
1. Comparison of activity rates by course outcome

Many variables and combinations of variables from RioLearn were investigated as possible input factors for a predictive model. Initial research allowed us to reduce the scope to a smaller set of predictive variables which we will present in this section using the freshman-level accounting course.

a. Visual analysis

Charts 1a-b display the mean log-in and site engagement measurements by scaled week. Definitions for log-in and site engagement activities were given above in Table 1. The means were calculated using the total frequency of activity at the time the week started. For this reason, the horizontal axis starts at "3" which corresponds to the second week for the most common course lengths. Only the active students were included in the calculations for each week. That is, if a student withdrew or was graded prior to the

course end date, they were not included in calculations for any weeks after that point. Charts 1c-d display the mean weighted log-in and site engagement measurements. By visual inspection, it was clear that the successful (solid) and unsuccessful (dashed) groups had different log-in and site engagement rates, even at an early point in the course. Furthermore, it was evident that the differences became increasingly pronounced throughout the duration of the course. Chart 1e displays the mean total of points submitted by scaled week. The conclusions drawn from visual inspection of this chart mirrored the conclusions drawn from the log-in and site engagement charts.



(a) Log-ins (b) Site engagement (c) Weighted log-ins

(d) Weighted site engagement (e) Points submitted

Figure 3. Starting mean cumulative LMS activities by scaled week. Freshman-level accounting course, summer & fall 2009.

b. Correlational analysis

The Pearson *r* correlation coefficients were used to quantify the correlation between course outcome and log-in frequency, site engagement, weighted log-in frequency, weighted site engagement, and points submitted. Course outcome was represented by an indicator variable where 1 = successful and 0 = unsuccessful. A total of 13 correlations were calculated for each LMS activity measurement – one for each scaled week. The results observed from this analysis indicated that there were significant correlations ($p < .05$) between course outcome and all five activity markers during every scaled week that was tested.

Factor	Statistic	Week (Scaled)** 3	4	5	6	7	8	9	10	11	12	13	14	15	16
Log in	Pearson r	0.162, p=0.041*	0.136, p=0.089	0.148, p=0.065	0.149, p=0.067	0.176, p=0.031*	0.178, p=0.036*	0.212, p=0.016*	0.218, p=0.014*	0.231, p=0.009*	0.246, p=0.006*	0.247, p=0.006*	0.269, p=0.002*	0.273, p=0.002*	0.288, p=0.001*
Weighted Log in	Pearson r	0.103, p=0.198	0.072, p=0.367	0.109, p=0.177	0.127, p=0.118	0.191, p=0.019*	0.198, p=0.020*	0.258, p=0.003*	0.218, p=0.016*	0.274, p=0.002*	0.295, p=0.001*	0.285, p=0.001*	0.32, p=0.000*	0.273, p=0.004*	0.336, p=0.000*

Table 2. Correlations between "Log in" and "Weighted Log in" variables with course outcomeNon-activity factors

Many other factors outside of LMS activity were also investigated, namely those related to past

enrollment patterns and credit load. Concurrent credit load, total sum of credits successfully completed, and total sum of credits attempted but not successfully completed were found to be effective predictors of course outcome, although they did not exhibit as significant of a correlation as did the LMS activity markers.

D. Predictive Models

1. Eighth day at-risk model

An initial set of predictive models were constructed in 2009 using enrollments from Fall 2008 and Spring 2009. These models were partially intended as a proof of concept, but were also designed to drive an intervention pilot, described in Section 8.1.1. Separate course-specific models were constructed for a total of fifteen courses across six discipline areas. Using the naïve Bayes classification method, the models were designed to run once on the eighth day of class. The models included approximately thirty input factors, including information derived from LMS activity logs, past enrollment patterns, and current enrollment status. The models were intended to produce estimated probabilities of course success which could then be translated to warning levels, of Low, Moderate, and High as described above.

The models were tested cumulatively using random sub-sampling cross-validation with ten repetitions. On average, the highest success rate was found in the Low warning group and the lowest success rate was found in the High warning group. The mean success rate was approximately 70% in the Low warning group, 54% in the Moderate warning group, and 34% in the High warning group. Therefore, the models were successful in accurately assessing the likelihood of successful course completion.

2. Progress and Course Engagement Model (RioPACE)

Following the successful accuracy tests for the eighth day at-risk models, a new system was developed in 2010 to provide updated information beyond the eighth day of instruction and on a weekly basis. The objective was to provide an automated, systematic early alert system that would allow instructors to launch proactive interventions at any point in the course to assist students who may show signs of struggling (i.e. slipping to a higher warning level). The pilot system was tentatively titled Rio STARS (Rio Student At-Risk **S**ystem), but to avoid confusion with a college sustainability effort also called S.T.A.R.S., the predictive model was renamed RioPACE, an acronym for **P**rogress **A**nd **C**ourse **E**ngagement. RioPACE was designed to determine appropriate warning levels on a weekly basis using updated activity and grade information. It was also intended to provide additional, more specific information to show student performance in the areas of log-in activity, site engagement, and pace. The system was again designed to run within the RioLearn LMS. For the pilot, launched in April 2010, warning levels were shown next to each student on an instructor roster, where one star indicated High warning, two stars indicated Moderate warning, and three stars indicated Low warning. The system allows an instructor to hover their cursor over an indicator, which will generate a small pop-up box showing the performance metrics for three critical areas: log-in frequency, site engagement, and pace. This section will discuss the two primary outputs of RioPACE (i.e. warning levels and activity metrics) and will present the results of accuracy tests for a freshman-level accounting course.

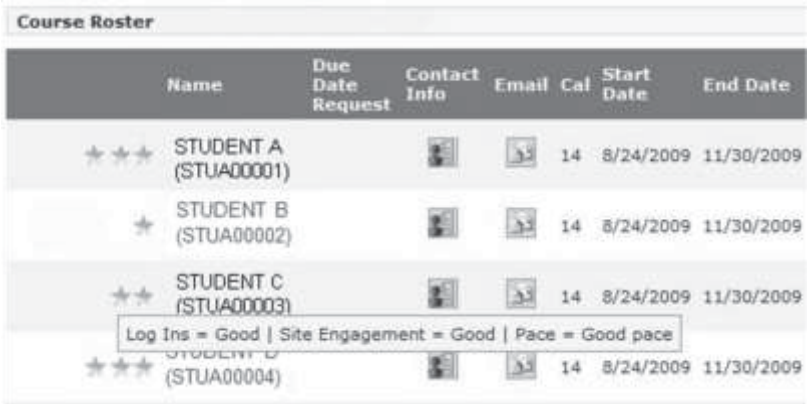


Figure 4. Sample RioPACE output on instructor course roster from the pilot study.

c. RioPACEwarning levels

Warning levels were generated using a naïve Bayes model with five input factors, shown in Table 3. The number of factors was reduced significantly compared to the eighth day at-risk models in an effort to base predictions solely on activity and performance. The reduction also helped to improve model efficiency (i.e. run time).

The model was tested using both weighted and non-weighted activity variables. During testing, it was found that the model yielded higher accuracy rates when using the weighted variables. Therefore, the weighted log-in frequency and site engagement variables were selected for the final model.

1	Weighted log-in frequency
2	Weighted site engagement
3	Points submitted
4	Points earned
5	Current credit load

Table 3. Input factors for Rio STARS naïve Bayes model.

Enrollments from the freshman-level accounting course during the summer and fall 2009 semesters (N = 539) were used to test the Rio STARS model. Random sub-sampling cross-validation with fifty repetitions was performed to validate model accuracy for every scaled week of the course. The results indicated that the distribution of warning levels was approximately uniform at the beginning of class. However, as the course progressed, the frequency of Low and High warning levels increased, while the frequency of Moderate warning levels decreased. By the end of class, the frequency of Moderate warning levels was extremely low as students were mostly seen in the Low and High warning levels. This was clearly the expected result.

Weekly warning levels were also compared to course outcome. The results showed that, for every scaled week, the average success rate in the Low warning level was higher than the Moderate warning level and the success rate in the Moderate warning level was higher than the High warning level. Chart 2 shows the average success rate for each warning level by scaled week. These results demonstrated that the Rio STARS model accurately predicted the likelihood of course success at every point throughout the course. It should also be noted that, as expected, the success rate in the Low warning level increases over time and the success rate in the High warning level decreases over time. This result reflects the increased accuracy with which we can predict course outcome as more activity and grade information accumulates over the duration of a course.

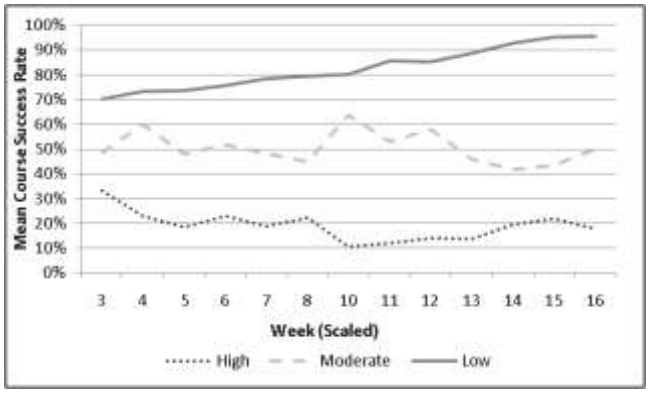


Figure 5. Mean success rate for each warning level by scaled week (from cross-validation results)

d. RioPACE activity metrics

The predictive model, RioPACE, calculated activity metrics for log-in frequency, site engagement, and pace using the method proposed using activity performance metrics. Table 4 shows the three possible levels for each activity metric.

Log-in frequency	Excellent, Good, Below average
Site engagement	Excellent, Good, Below average
Pace	Working ahead, Good, Falling behind

Table 4. Performance levels for RioPACE activity metrics

D. Interventions

Although this is an area that requires considerably more research, we will discuss two initial outreach programs that were conducted as a result of the predictive models described in this paper. Experimental trials were conducted with both of these programs. Both programs were implemented prior to the development of Rio STARS. As of April 2010, planning is underway to formulate intervention strategies for the new STARS system.

1. Eighth day interventions

One intervention strategy developed for this study involved faculty-designed outreach delivered directly to the students predicted to be at-risk. In summer and fall 2009, the faculty chairs from six disciplines were invited to participate in a pilot intervention trial. The chairs were asked to design a student success intervention strategy based on predictions generated by their discipline-specific model on the eighth day of class. In order to make the scope of the initial pilot more manageable, it was determined that the interventions would focus solely on students in the Moderate warning level. Each discipline developed a unique intervention strategy, but most of them involved some form of direct and informal contact via telephone.

In order to conclusively measure the impact of the faculty-designed interventions, a control group was established. Students in the control group still had access to all Rio Salado College services and were still exposed to the traditional forms of success and retention outreach efforts that all students receive. The trial was conducted during summer and fall 2009. Results were collected from May through October 2009 and included a total of N = 2,300 enrollments from the Moderate warning level, half of which were randomly assigned to the control group. The results indicated that the faculty-designed outreach programs did not generate significant improvements in success rates, perhaps due to the fact that it was difficult for instructors and staff to reach students by telephone. In fact, only one-third of phone call attempts made by instructors and staff during the trial resulted in direct student contact. However, there was evidence to indicate that students who received direct contact succeeded more often than students who received non-direct contact (i.e. voicemail) or no contact (i.e. wrong number, no answer, etc.).

2. Automated course welcome emails

During development of the predictive models, it was observed that early log in activity was one of the strongest predictors of long-term student success. In fact, spring 2009 online students in general education courses who logged in on the first day of class succeed 21% more often than students who did not log in on the first day. There was clearly no absolute evidence for a causal relationship between early log in activity and long-term course success. Nonetheless, it seemed logical to encourage students to log in early

and thus adopt the attributes of previously successful students. To that end, an automated welcome email system was piloted for online students in two selected disciplines, starting in October 2009. The pilot was continued for each weekly start date through the end of December 2009. The system distributed welcome emails to students the evening prior to their start date, which encouraged them to log in to the course homepage, contact their instructor, and begin participating in the course.

Results were collected from October and November 2009. A total of N = 364 students were eligible for the trial and half were randomly placed in a control group. The results showed that the emails generated an approximately forty percent decrease in drop rate compared to the control group, which did not receive the email. The difference observed in drop rate between the treatment and control groups was statistically significant. A significant increase in early log in rate was also observed. Subsequent attempts to expand the system have not yielded the same positive impact, so additional research is needed to determine the best method of scaling this particular outreach effort.

IV. CONCLUSIONS

The promise of "Big Data" to increase effectiveness and student success in higher education continues to generate great expectations at many institutions, yet there are few specific cases that show these steps—especially among community colleges where online learning has the most potential for growth. This study demonstrated the strong correlation that exists between LMS activity markers and course outcome, which has also recently been noted by other researchers [12]. We have shown that log-in frequency, site engagement, pace, assignment grades, and some non-LMS enrollment factors can serve as effective predictors of course outcome, even as early as the eighth day of class. Two predictive models were presented showing that these factors can be used to accurately predict the likelihood of course success at any given point in the semester. We have also demonstrated how additional activity metrics can be coupled with these predictions to provide more detail and context for the instructor when conducting student outreach and intervention. A summary of two initial intervention pilots was provided showing occasional positive results. Clearly this is an area that calls for more research to develop theory, practice and policies to harness the potential of learning analytics and predictive modeling.

V. ABOUT THE AUTHORS

Vernon C. Smith is Provost at MyCollege Foundation, a non-profit (501c3) organization seeking to help low-income youth in America gain high-quality college credentials more affordably. Vernon is former faculty and Vice President of Academic Affairs at Rio Salado College.

Adam Lange is a Reporting Analyst/Developer at Ellucian (formerly Datatel). Previously, Adam worked as a Programmer Analyst III in Institutional Research at Rio Salado College.

Daniel R. Huston is the interim Director of Research, Planning and Development.

VI. ACKNOWLEDGEMENTS

This study was supported by the Maricopa County Community College District and through the student success efforts of Dr. Maria Harper-Marinick, Executive Vice Chancellor and Provost. Additionally, support and encouragement was provided by Dr. Chris Bustamante, President of Rio Salado College.

VII. REFERENCES

1. **Obama, B.** "Remarks by the president on the American Graduation Initiative." Macomb Community College. Warren, MI. (, July 14, 2009).
2. **Terence, C.** Colleges cap enrollment amid budget cuts. Associated Press. (January 14 2010). http://www.pbs.org/nbr/headlines/US_Competing_for_Admission/index.html.
3. **Johnson, N., Oliff, P., and Williams, E.** An update on state budget cuts. Center on Budget Policy and Priorities. (December 18, 2009). http://www.cbpp.org/files/3-13-08sfp.pdf.
4. **Green, K.** LMS 3.0. *Inside Higher Ed* (November 4 2009). http://www.insidehighered.com/views/2009/11/04/green.

5. **Johnson, L., Smith, R., Willis, H., Levine, A., and Haywood, K.**, The 2011 Horizon Report. (2011).
6. **Campbell, J., De Blois, P. B., and Oblinger, D.G.** Academic analytics. A New Tool for a New Era. *EDUCAUSE Review* 42(4): 42–57 (2007).
7. **Baepler, P., and Murdoch, C.** Academic Analytics and Data Mining in Higher Education. *International Journal for the Scholarship of Teaching and Learning* 4(2) (2010).
8. **Figini, S., and Giudici, P** Statistical models for e-learning data. *Statistical Methods and Applications* 18(2): 293-304 (2009).
9. **Smith, V., and Lange, A.** Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions. Research Paper Presentation, Council for the Study of Community Colleges, Seattle, Washington, April 16-17, 2010.
10. **Kolowich, S.** The new diagnostics. *Inside Higher Ed* (October 30 2009). http://www.insidehighered.com/news/2009/10/30/predict.
11. **Lange, A., Corona, S., and Ushveridze, A.** Improving Student Persistence and Success through Predictive Modeling Analytics. WCET Annual Conference, Phoenix, AZ (November 2008).
12. **Macfadyen, L., Dawson, S.,** Mining LMS data to develop an ‘‘early warning system” for educators: A proof of concept. *Computers & Education* 54: 588-599 (2010).
13. **Auguste, B., Cota, A., Jayaram, K., and Laboissiere, M.** Winning by degrees: the strategies of highly productive higher- education institutions. *Education* 66 (2010). http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Winning+by+degrees+:+the+strategies+of+highly+productive+higher-+education+institutions#
14. **Weisstein, E.W.** "Ceiling Function." From MathWorld--A Wolfram WebResource. http://mathworld.wolfram.com/CeilingFunction.html.
15. **Lewis, D.** Naive (Bayes) at forty: The independence assumption in information retrieval. *Lecture Notes in Computer Science* 13: 4-15 (1998).
16. **Domingos, P., and Pazzani, M.** On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* 29: 103-130 (1997).
17. **Kohavi, R.** A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2 (12): 1137–1143 (1995).