Course Correction:
Using Analytics to Predict Course Success

Rebecca Barber, Ph.D
Apollo Group
4035 Riverpoint Parkway
Phoenix, AZ 85040
1-602-557-7842
rebecca.barber@apollogrp.edu

Mike Sharkey
Apollo Group
4035 Riverpoint Parkway
Phoenix, AZ 85040
1-602-557-3532
mike.sharkey@apollogrp.edu

ABSTRACT

Predictive analytics techniques applied to a broad swath of student data can aid in timely intervention strategies to help prevent students from failing a course. This paper discusses a predictive analytic model that was created for the University of Phoenix. The purpose of the model is to identify students who are in danger of failing the course in which they are currently enrolled. Within the model's architecture, data from the learning management system (LMS), financial aid system, and student system are combined to calculate a likelihood of any given student failing the current course. The output can be used to prioritize students for intervention and referral to additional resources. The paper includes a discussion of the predictor and statistical tests used, validation procedures, and plans for implementation.

Categories and Subject Descriptors

G.3 SPSS; H.2.3 SQL; H.2.4 Oracle; **J.1 [Administrative Data Processing]** Education; **K.3.1 [Computer Uses in Education]** Collaborative learning, Computer-assisted instruction (CAI), Computer-managed instruction (CMI), Distance learning

General Terms

Management, Measurement, Experimentation

Keywords

Learning Analytics, Predictive Analytics, Predictive Modeling, Higher Education, Retention.

1. INTRODUCTION

Whereas the holy grail of predictive models in higher education would likely be one that could predict graduation at the time a student applies for admission, the reality is that the elapsed time between the start and end of college is years long, creating the opportunity for a multitude of factors to interfere with a student's progress. A model to make such a prediction will take years to develop and require data beyond the scope of what is available in an institution's Student Information System (SIS) and LMS.

Nonetheless, there are known reasons students fail to graduate. Work schedules, health problems, child care challenges, transportation, and financial issues comprise some of the reasons students drop out that are largely outside of the institution's control [1]. One predictor of students' decision to drop out that is under the institution's purview, however, is a lack of preparation or effort as reflected in their course grades.

Regardless of the reason, a student enrolled in a course will often display signs of course failure before either formally withdrawing or disappearing altogether. Failing to attend class (or in the case of online courses, failing to participate in discussion forums), sloppy or incomplete assignments, or a significant change in the student's behavior and academic performance are all warning signs that a student may be on the verge of dropping out.

It is generally in both the student's and the institution's best interest that students remain enrolled or, if they must leave, withdraw via the established process. It is this concern that precipitated the development of a predictive model. Whereas the student may be wrestling with problems that range from personal tragedy to time management or academic under-preparedness, the University can monitor the student's behavior in a course for warning signals and increase that student's priority for a call from his or her academic advisor. The advisor can help the student by pointing them to necessary resources, coaching them on time management, or even advising early withdrawal.

This paper discusses the rationale for the model; the process through which it was developed, revised and refined; and the validation of the model by the operational team. The next section will provide the context within University of Phoenix and further exposition of the problem the model is intended to address. Following this discussion is a brief overview of the extant literature that guided some of the decisions made about variables included in the model. This section is followed by a description of the methods used to develop the model, including the data elements found to be predictive and those found to be irrelevant to prediction. We then discuss the process used to validate the model within our operational environment. Finally, we offer a brief discussion of next steps and plans for broader implementation.

2. INSTITUTIONAL CONTEXT

Founded in 1976, University of Phoenix is a regionally accredited, degree-granting institution. Based in Phoenix, Arizona, the University has over 200 campuses throughout the United States and the largest (measured in student enrollment) online campus in North America. In addition to holding regional accreditation, University of Phoenix holds programmatic accreditation in nursing, counseling, business, and education. As of August, 2011 the university enrolled more than 340,000 students in over 100 degree programs, ranging from associates through doctorates.

The university was founded with a focus on working adults who wished to complete their degree. These non-traditional learners remain the focus of the university, resulting in a more diverse

LAK'12, 29 April – 2 May 2012, Vancouver, BC, Canada.
Copyright 2012 ACM 978-1-4503-1111-3/12/04…$10.00.

student population than found in traditional institutions, in terms of racial and ethnic demographics as well as the proportion of first- generation college students [2]. Also, many non-traditional students are employed while pursuing their education. In order to help students complete their degrees in a time-efficient fashion, University of Phoenix adopted a focused academic model in which courses last between 5 and 9 weeks[1].

Non-traditional learners may have been out of school for many years before deciding to pursue a post-secondary degree. Adding school responsibilities to busy schedules and refreshing study skills are challenges that all returning students face. We constantly seek new ways to provide our students with the services they need in order to progress academically, including tutoring and coaching. And, although a number of static triggers (such as those that monitor attendance) exist to monitor students for signs of trouble, there is continued interest in improving the information available to academic advisors in order to direct interventions.

3. THEORETICAL FOUNDATION

Garman used logistic regression to predict student success in an online database course based primarily upon scores on a reading comprehension assessment [3]. The only other input to the model was the semester in which the student took the course. Whereas this approach is interesting in that it supports the proposed methodology for our model, the study found the semester variable insignificant and the assessment score only minimally predictive.

Moore looked explicitly at course participation in both the student's current and prior courses [4]. This research indicated increased participation to correlate highly with higher performance in the course. Some other variables, such as student expectations, high school rank, and entry exam scores (ACT) were not significant predictors of student achievement.

The standard measure for monitoring participation in an online course is student discussion postings, and prior research has found final grades correlated with the number of postings both read and written by students [5]. However, other research has found postings to have an indeterminate relationship with course success [6,7]. Ramos and Yudko found that total page hits were more predictive than discussion board use of online course success [8]. The lack of agreement suggests including post counts in the model until they can be definitively excluded.

Regarding demographic variables, Martinez found high school GPA, age, sex, grade in last math class, highest level of math, ethnicity, definite major choice, and work hours planned to predict success in different levels of English courses [9]. In addition, current credit hours, financial aid usage, and program level were predictive of the likelihood of drop out [10]. Where possible these variables were included.

Because the goal of this project was to work from an existing data set, studies addressing variables that are unavailable (such as self-discipline, motivation, locus of control, and self-efficacy) were not included in our literature review.

4. METHODS

Based on the literature, the variables in Table 1 were identified as potentially useful and worth examining. A number of key variables were populated for less than 50% of the cases. Logistic regression drops any records for which all of the fields are not populated, resulting in too large a loss of data. Therefore, despite theoretical support for those fields' inclusion, the decision was made to exclude them from the analysis at this time.

Table 1: Variable Disposition

Field	*Status*
Attendance /week	Model 1-3
% cumulative course points /week	Model 1-3
Prior credits earned	Model 1 (replaced w/ratio)
Discussion post count /week	Model 1-3
Late assignments	Bad data quality
Gender	Model 1
Age at program start	Model 1
Unsubstantive post count	<10% populated
Ethnicity	<50% populated
Marital Status/Dependents	<25% populated
Employment Status/Years	<25% populated
Household Income/Salary	<25% populated
High school GPA	<25% populated
Financial aid / Pell Grant recipient	Model 2
Total student loans taken	Model 2
Financial status (current/other)	Model 2-3
Ratio of credits earned/attempted	Model 2-3
Military status	Model 2
Aattendance	Model 2-3
Days into the course of 1st activity	Model 2-3
Pct point delta to prior courses	Model 2-3
Orientation participation	Model 3
Inactive time since last course	Model 3
Count of messages to instructor	Model 3

4.1 Model Version 1

A consulting company using a limited data set created the initial model. The data included a unique identifier; basic demographic information from a voluntary survey completed by the student at time of admission; and academic history within the University, including number of transfer credits, number of courses taken, and percentage of points earned in these courses. For each course, information was also provided about discussion board postings, points earned by week within the course, and whether the student submitted assignments late. The consulting company used this data to create a logistic regression model.

The analysis of the initial data exposed missing data and data quality issues that would have compromised the final model. Fields, such as submission timeliness and discussion board post quality, were found to be either inaccurate or missing too many records to contribute to the model. The final data set was reduced to data reflecting transfer credits, prior academic activity at University of Phoenix, and week-by-week activity (points earned and discussion posts made) for each course. These data elements were further recoded to create interpretable indicators.

The data set included all activity for approximately three months[2], organized by degree. The SPSS randomization algorithm selected approximately 50% of the data as a hold-out sample, making the remaining 50% available for model development. These data were analyzed using logistic regression, with the outcome variable

[1] Associates degree students take 2 9-week courses at a time. Bachelor's students take 1 5-week course at a time. Graduate students take 1 6-week course at a time. Additional weekly hours make these courses equivalent to traditional semester-based courses.

[2] The data set included more information for some elements, but October to December 2010 was the most complete.

being an indicator of whether the student passed the course. The model assessed student data through course week 4.

Separate models were developed for each degree level. For example, coefficients for bachelor's degree students through week 2 were as follows:

Table 2: Coefficients for Model 1, Bachelor's degree students

	Week 0	*Week 1*	*Week 2*
<65% points in prior courses	-1.44	-0.76	-0.52
>85% point in prior courses	+0.46	+0.58	+0.82
Credits earned at Univ of Phx	+0.07	+0.02	+0.02
Online Posts		+0.15	+0.18
Cumulative points Earned		+3.15	+4.22

* All results significant at p<.05 level.

These coefficients allow us to classify each student into one of three tiers in week zero[3]: high risk, low risk, or neutral risk. Initial percentages that comprised the neutral zone ranged from 41% to 54% of students.

Models in weeks 1 through 4 added discussion post information and percentage of assignment points earned. This parsing immediately (by the end of week 1) trimmed the range of the grey zone to between 35% and 40%. By week 2 all master's degree students were out of the neutral zone. By week 3 all bachelor's degree students were out of the neutral zone. Results for students not in the neutral zone were accurately predicted on average 94% of the time, with no week below 85%. In other words the prediction of pass (low risk) or fail (high risk) was accurate more than 90% of the time.

4.2 Concerns with Model 1

The neutral zone was quite large initially, and the team felt that a better way to categorize the students in that zone was to assigned a "score" to each student, ranging from 1 (unlikely to pass) – 10 (nearly sure to pass). That score would provide the prioritization needed to make the output actionable. Also, since some courses are as long as 9 weeks, the time frame needed to be extended.

There was also concern about the reproducibility and actionability of the model as developed. The data used came from a variety of different databases and, as such, required significant manual intervention to compile. At least one of the data sources required a programmer to do an ad-hoc query to generate a data set that was not directly accessible to the analytics team.

Enhancement of the model was brought in-house. The model was replicated for validation purposes, and then, the process of refinement and further development was started.

4.3 Model Version 2

One of the first problems addressed was that of data availability and validity. As mentioned, there were critical data elements around discussion board postings that were not easily available. However between the initial data request and the start of model 2 a partial feed of the data was added to the enterprise data warehouse environment. This advancement allowed post count by week to be incorporated into the model. Further, the ability to automatically update all data fields will make implementation easier.

One limitation of the new data source is that it has only been populated since June, 2011. Therefore data extraction processes were developed to use data from June through October, 2011, to construct the updated model. Additionally, some variables that had not been included in model 1 were included in model 2. These included military status and financial status.

As model 1 showed, some of the variables that the literature suggested were relevant proved not to be when looking across all programs and levels. Gender, age, military status, pell grant receipt and responsible party (whether the student was receiving financial aid, paying through their employer, or paying directly) were not significant or resulted in extremely low weights. There was also some variability by degree level. For example, military status was not significant for associate's and bachelor's degree students, but was significant and negative for master's degree student. One interpretation is that master's-level military personnel are more likely to be officers and therefore more likely to have substantial responsibilities that could interrupt their studies; however, other explanations could be considered. Because of their lack of predictive power, most of these variables were eliminated from the final version of Model 2.

Model 2 was built using a Naïve Bayes algorithm in RapidMiner and validated using 10x cross validation. A sample of information gain weights on a zero to one scale are in Table 3.

Table 3: Weights for Model 2, Bachelor's Online students

	Week 0	*Week 1*	*Week 2*
Cumulative points earned (%)		1.00	1.00
Financial Status not current	0.78	0.33	0.28
Ratio credits earned/attempted	0.95	0.36	0.28
Points earned - prior courses (%)	1.00	0.40	0.34
Online post count		0.25	0.21
Point delta - prior courses >10%		0.47	0.42
Sum of student loans taken	0.30	0.14	0.05
Number of concurrent courses	0.47	0.05	0.02
Attendance		0.12	0.08

* All results significant at p<.05 level. Contact 1st author for full table.

The new variables added to model 2 increased the predictive accuracy considerably. Model 2 accurately predicted 85% of all students at week 0 (compared to 50% in model 1), rising to 95% by week 3.

Specifically, the ratio of credits earned to credits attempted was a substantial indicator of potential problems, as was a financial status other than current. As might be expected, cumulative points earned remained the most powerful predictor. In addition, whether the point delta between prior and the current course had change by more than 10% was treated as a categorical variable and was also influential. This variable is less about the exact point change than an indicator of a substantive change in behavior. Most of the other variables provided less predictive power than these three, although enough to keep in the model.

4.4 Model Version 3

Development of model version 3 is awaiting the availability of higher quality posting data. While version 2 allows direct access to the total number of posts made by a student, it fails to provide any distinction between posts made in response to discussion questions, posts made in collaborative forums, and posts made to the private forum provided for discussion with the instructor. Because the literature suggests a link between passing and engagement with the instructor [6], this differentiation between post sources is necessary for the next phase of analysis. Acquisition of this data is underway.

[3] Week zero is the week prior to starting the course. There is no activity in the course yet to use for the model.

Additionally, per Ramos and Yudko [8], there is value in tracking student page views within the learning environment. Currently, only aggregate data is available, but these data are not useful from a predictive analytics perspective. Accordingly, the technical team is working to capture this data at the individual student level.

To complement these data elements around discussion board activity, three variables will be added based upon research conducted at another institution: major area of study, time since last course, and participation in an orientation program. These elements will be incorporated into the model while the discussion data is being sourced.

One other potential modification will be a review of other types of models. Specifically support vector machines and random forest models will be investigated for improved predictive accuracy.

5. MODEL VALIDATION

Validation of model 1 involved using the 50% hold-out sample. The risk category percentage differences between estimation and hold-out samples were within two percent, with the majority of cases within one percent. For model 2, a 10x cross-validation procedure was used which provide that the model was highly accurate at predicting students who would pass, but with some room for accuracy improvement on those students who were predicted to not pass. These were under predicted in the model.

More important than validation of the model fit, however, is validation of the model's utility. In order to validate that the model was indeed providing actionable information, a pilot has been initiated to create scores that could be provided to academic advisors. The academic advisors then use these scores for prioritization, calling students with the highest risk score first, even if that student would not normally have received a call, while delaying calls to students who scored lower.

The initial pilot of the model is being conducted with only a few academic advisors. These more experienced advisors are looking at the model in terms of both its accuracy (does the information the model provides align with what they learn by talking to the student) and its utility (does it trigger contact with the right students, and are those students then successful?). Statistical validity alone is insufficient; the model must provide actionable information to front-line advisors in a form that can increase student success in order for it to be seen as truly valid.

6. NEXT STEPS AND IMPLEMENTATION

Providing a score for students that can be used for prioritization is helpful, but too many students remain in the neutral zone during weeks 0 and 1. Refinement will continue, with the objective of accurately placing them at one end of the scale or the other.

The overarching goal for this project remains to provide valuable, timely information to academic advisors. Once the pilot completes, the utility will be evaluated and a decision will be made as to whether to implement the model into the production processes, making the results available to all academic advisors. At this point, it is unclear as to whether this will be in the form of a gradient categorization (e.g. red/yellow/green), a numeric score, or a percentage chance of withdrawal. Accordingly, the model will continue to be refined after initial implementation to best suit advisors' needs.

Future refinements will depend on the availability of additional detail data from both the learning management platform and the student information system. Concurrent work is proceeding to substantially improve access to that data, making integration into the model both technologically easier and substantially faster.

7. ACKNOWLEDGMENTS

Our thanks to Andrew Tubley for help with the title of this paper.

8. REFERENCES

[1] Cohen, A. and Brawer, F. 2009. The American Community College. John Wiley and Sons, NY, ISBN: 9780470605486.

[2] University of Phoenix, 2010 Academic Annual Report. http://www.phoenix.edu/about_us/publications/academic-annual-report/2010.html

[3] Garman, G. 2010. A Logistic Approach to Predicting Student Success in Online Database Courses. American Journal of Business Education. 3(12), 1-5.

[4] Moore, R. 2007. Do Students Performances and Behaviors in Supporting Courses Predict Their Performances and Behaviors in Primary Courses? Research and Teaching in Developmental Education. 23(2), 38-48.

[5] Wang, A.Y. & Newlin, M.H. 2000. Characteristics of students who enroll and succeed in Psychology web-based classes. J. Educational Psychology. 92(1), 137-143.

[6] Reisetter, M. & Boris, G. 2004. What works: student perceptions of effective elements in online learning. Quarterly Review of Distance Education. 5(4), 277-291.

[7] Sadik, A. & Reisman, S. 2004. Design and implementation of a web-based learning environment: lessons learned. Quarterly Review of Distance Education. 5(3), 157-171.

[8] Ramos, C. & Yudko, E. 2008. "Hits" (Not "Discussion Posts") Predict Student Success in Online Courses: A Double Cross-Validation Study. Computers & Education. 50(4), 1174-1182.

[9] Martinez, D. 2001. Predicting student outcomes using discriminant function analysis. Paper presented at the 39th Annual Meeting of the Research and Planning Group, Lake Arrowhead CA.

[10] Morris, L., Wu, S. & Finnegan, C., 2005. Predicting retention in online general education courses. American Journal of Distance Education. 19(1), 23-26.