Course Correction:
Using Analytics to Predict Course Success

Rebecca Barber, Ph.D
University of Phoenix
4035 Riverpoint Parkway
Phoenix, AZ 85040
1-602-557-7842
Rebecca.Barber@phoenix.edu

Mike Sharkey
University of Phoenix
4035 Riverpoint Parkway
Phoenix, AZ 85040
1-602-557-3532
Mike.Sharkey@phoenix.edu

ABSTRACT
Predictive analytics can support timely intervention to prevent students from failing a course and to provide additional services. This paper discusses just such a model that was created and is being implemented within University of Phoenix. The goal of the model is to identify students that are in danger of failing the course in which they are currently enrolled. Data from the learning management system (LMS) 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

1. INTRODUCTION
While the holy grail of predictive models in higher education would likely be one that can predict graduation at the time a student applies for admission, the reality of the situation is that the delay between the start and end of college is years long, leaving 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 far beyond the scope of what is available in either a student information system (SIS) or a learning management system (LMS).

Nonetheless, we know some of the reasons students fail to graduate. Work schedules, health problems, child care challenges, transportation and financial issues are reasons why students drop out that are nominally outside of the control of the institution [1]. Other students drop out due to a lack of preparation or effort as reflected in their course grades.

Regardless of the reason, a student currently enrolled in a course will often display signs of course failure before either withdrawing or disappearing. 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 terms of the students 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 and the institutions best interest that students remain enrolled or, if they must leave, withdraw gracefully. This allows financial aid issues to be resolved and the student to leave with their GPA undamaged. Yet at University of Phoenix many students just disappear. This is in part a function of our concentrated course calendar; courses last as little as 5 weeks, resulting in the course being over before the student realizes the situation has gotten bad enough that they should withdraw.

It is this last problem that prompted the development of a predictive model. While the student may be wrestling with problems that range from personal tragedy to time management or academic under-preparedness, the University can evaluate the student’s behavior in a course for warning signals and increase that student’s priority for a call from their academic adviser. The adviser can help the student by pointing them to necessary resources, coaching them on time management or even advising withdrawal early, before the student has incurred more fees or damaged their GPA.

This paper discusses the rationale for the model, the process through which it was developed, revised and refined, and the validation and implementation plan for incorporating the model into the operational process. The next section will provide the context within University of Phoenix and further exposition of the problem being addressed. Immediately after that, there will be a brief discussion of the extent literature available that guided some of the decisions made about variables to include in the model.

This will be followed by a discussion of the methods used to develop the model, including discussion of the data elements found to be predictive and those found to be irrelevant to prediction. Next will be a section to discuss the process used to validate the model within our operational environment. Finally, a brief discussion of next steps and plans for broader implementation will be explained.
2. INSTITUTIONAL CONTEXT

University of Phoenix is a regionally accredited degree-granting institution founded in 1976. Based in Phoenix, Arizona, the University has over 200 campuses throughout the United States and the largest online campus in North America. In addition to holding regional accreditation, University of Phoenix holds program accreditations 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, with the student population being more diverse and made up of more minority and first-generation college students than a traditional institution [2]. Nearly every student works, most full time, while pursuing their education. In order to help students complete their degrees, University of Phoenix adopted a focused academic model in which each course lasts between 5 and 9 weeks.

As an open-admissions institution, we accept any student who indicates that they are motivated to complete their degree and holds a high school diploma or equivalent. Motivation, however, does not always make up for a lack of preparation or the challenges that come from having been out of school for many years before pursuing a degree. We struggle to provide our students with the services they need in order to progress academically, including tutoring and coaching.

The problem was that often these services did not arrive in time to prevent a student from dropping out. Students might be unaware of the added services available and academic advisers are only able to contact a particular student every few weeks. Although a number of static triggers exist (such as those that monitor attendance), there is continued interest in improving the process.

3. THEORETICAL FOUNDATION

Garman used logistic regression to predict student success in an online 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. While 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 current course and prior courses [4]. He found 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 way of 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 be 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 clarity 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 will be included.

Since the goal of this project was to work from data we had, studies addressing variables that are unavailable (such as self-discipline, motivation, locus of control and self-efficacy) are not included.

4. METHODS

The overall model design chosen was logistic regression. Logistic regression uses a dichotomous outcome variable, in this case whether the student passed the course or either failed or did not complete the course. Predictive discriminant analysis was also considered, but has been found to perform comparably to logistic regression [11]. All models were developed by degree level and modality (online vs. ground). College will be added for model 3.

Based upon 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 backing for those fields the decision was made to exclude them from the analysis at this time.

<table>
<thead>
<tr>
<th>Field</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance /week</td>
<td>Model 1-3</td>
</tr>
<tr>
<td>% cumulative course points /week</td>
<td>Model 1-3</td>
</tr>
<tr>
<td>Prior credits earned</td>
<td>Model 1-3</td>
</tr>
<tr>
<td>Discussion post count /week</td>
<td>Model 1 (replaced w/ratio)</td>
</tr>
<tr>
<td>Late assignments</td>
<td>Bad data quality</td>
</tr>
<tr>
<td>Gender</td>
<td>Model 1-3</td>
</tr>
<tr>
<td>Age at program start</td>
<td>Model 1-3</td>
</tr>
<tr>
<td>Unsubstantive post count</td>
<td>&lt;10% populated</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>&lt;50% populated</td>
</tr>
<tr>
<td>Marital Status/Dependants</td>
<td>&lt;25% populated</td>
</tr>
<tr>
<td>Employment Status/Years</td>
<td>&lt;25% populated</td>
</tr>
<tr>
<td>Household Income/Salary</td>
<td>&lt;25% populated</td>
</tr>
<tr>
<td>High school GPA</td>
<td>&lt;25% populated</td>
</tr>
<tr>
<td>Financial aid / Pell Grant recipient</td>
<td>Model 2-3</td>
</tr>
<tr>
<td>Financial status (current/other)</td>
<td>Model 2-3</td>
</tr>
<tr>
<td>Ratio of credits earned/attempted</td>
<td>Model 2-3</td>
</tr>
<tr>
<td>Military status</td>
<td>Model 2-3</td>
</tr>
<tr>
<td>Weekly attendance</td>
<td>Model 2-3</td>
</tr>
<tr>
<td>Days into the course of 1st activity</td>
<td>Model 2-3</td>
</tr>
<tr>
<td>Cumulative GPA</td>
<td>Model 2-3</td>
</tr>
<tr>
<td>Pct point delta to prior courses</td>
<td>Model 2-3</td>
</tr>
<tr>
<td>Orientation participation</td>
<td>Model 3</td>
</tr>
<tr>
<td>Inactive time since last course</td>
<td>Model 3</td>
</tr>
<tr>
<td>Count of messages to instructor</td>
<td>Model 3</td>
</tr>
</tbody>
</table>

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, academic history within the University, including number of transfer credits, number of courses taken so far, and percentage of points earned in the courses taken so far. 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 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 months¹, 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 being an indicator of whether the student passed the course. The model looked at the student through course week 4.

The results of this model varied by degree. For example, bachelor’s degree students through week 2 were as follows:

<table>
<thead>
<tr>
<th>Table 2: Coefficients for Model 1, Bachelor’s degree students</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Week 0</strong></td>
</tr>
<tr>
<td>&lt;65% points in prior courses</td>
</tr>
<tr>
<td>&gt;85% point in prior courses</td>
</tr>
<tr>
<td>Credits earned at Univ of Phx</td>
</tr>
<tr>
<td>Online Posts</td>
</tr>
<tr>
<td>Cumulative points Earned</td>
</tr>
</tbody>
</table>

* All results significant at p<.05 level.

These coefficients allow us to classify each student into one of three tiers in week zero²: high risk, low risk, or grey zone. Initial percentages in the grey zone ranged from 41% to 54% of students.

Weeks 1 through 4 replaced added discussion post information and percentage of assignment points earned. This immediately (by the end of week 1) trimmed the size of the grey zone to between 35% and 40%. By week 2 all master’s degree students were out of the grey zone. By week 3 all bachelor’s degree students were out of the grey zone. Results for students not in the grey 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 grey zone was quite large initially, and the team felt that a better way to categorize the students in that zone was “score” each student 1-10. That score would provide the prioritization needed to make the output actionable. Also, since some courses are as much as 9 weeks long, the time frame needed to be extended to include the additional course weeks.

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 against a data source that was not accessible to the analytics team directly. Finally there was some concern about the way the data had been interpreted by the consulting company.

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 above, 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 allowed post count by week to be incorporated into the model. This ability to automatically update all data fields will make implementation easier.

One limitation of the new data source is that it is only populated beginning in June, 2011. Therefore data extraction processes were developed to use data from June through October, 2011, to develop 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 and responsible party (whether the student was receiving financial aid, paying through their employer, or paying cash) were not significant or resulted in extremely low coefficients. There was also some variability by degree level. For example military status was not significant for associates and bachelors degree students, but was significant and negative for master’s degree student. Our current interpretation is that masters-level military personnel are more likely to be officers and therefore more likely to have substantial responsibilities that could interrupt their studies, however we are investigating other possibilities.

<table>
<thead>
<tr>
<th>Table 3: Coefficients for Model 2, Bachelor’s Online students</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Week 0</strong></td>
</tr>
<tr>
<td>&lt;65% points in prior courses</td>
</tr>
<tr>
<td>&gt;85% point in prior courses</td>
</tr>
<tr>
<td>Financial Status not current</td>
</tr>
<tr>
<td>Ratio credits earned/attempted</td>
</tr>
<tr>
<td>Transfer Credits &gt;18</td>
</tr>
<tr>
<td>Days until 1st activity (iday)</td>
</tr>
<tr>
<td>Online Posts (/post)</td>
</tr>
<tr>
<td>Point delta to prior courses (%)</td>
</tr>
<tr>
<td>Cumulative points earned (%)</td>
</tr>
</tbody>
</table>

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

The new variables added to model 2 increased the predictive accuracy. 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.

---

¹ The data set included more information for some elements, but October to December 2010 was the most complete.
² Week zero is the week prior to starting the course. There is no activity in the course yet to use for the model.
To put the coefficients from table 3 into perspective, each
additional 1% of points earned in week one makes a student twice
as likely to pass the class, with the effect getting substantially
stronger as the course progresses. Additionally if the student’s
behavior has changed, as measured by the difference between
their prior average percentage of points earned and their current
course percentage of points earned, the impact is large. Students
earning 10% below their prior level are 3 times less likely to pass.
Having a financial status other than current makes a student half
as likely to pass the course as they would be if they were in a
current status.

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
division between posts made in response to discussion
questions, posts made in learning teams and posts made to the
private forum provided for discussion with the instructor. Since
the literature suggests a link between passing and engagement
with the instructor [6], this division of post source is necessary for
the next phase of analysis. Acquisition of this data is underway.
Additionally, per Ramos and Yudko [8], there is value in tracking
student page views within the learning environment. The technical
team is working to capture this data at the individual student level.
Currently the data is available for all students aggregated, but that
is not useful from a predictive analytics perspective. Effort is
ongoing to secure that data.

In addition to these data elements around discussion board
activity, three additional 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.
Major area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major
area of study, research conducted at another institution. Major

Additionally, per Ramos and Yudko [8], there is value in tracking
student page views within the learning environment. The technical
team is working to capture this data at the individual student level.
Currently the data is available for all students aggregated, but that
is not useful from a predictive analytics perspective. Effort is
ongoing to secure that data.

In addition to these data elements around discussion board
activity, three additional 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, have
been identified as meaningful predictors. Addition of these
elements is proceeding while the discussion data is being sourced.

5. MODEL VALIDATION
Validation of the models initially involved using the 50% hold-
out sample. The risk category percentage differences between
estimation and hold-out samples were universally within two
percent, with the vast majority of cases within one percent.

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 was
started to use the model to create scores that could be provided to
the academic counselors. The academic advisors use the scores for
prioritization, calling students with the highest risk of failure first
even if that student would not normally have received a call, while
delaying calls to students who were clearly on track.

The initial pilot of the model is being conducted with only a few
academic advisors. These more experience 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 it’s utility (does it trigger contact with the right students and
are those students then successful). As of November, 2011, the pilot is
in progress4.

4 The intention of this paper is that by the time the final image-ready
paper is due, the pilot will be complete. The paper will be updated at
that time with the results.

6. NEXT STEPS AND IMPLEMENTATION
Providing a score for students that can be used for prioritization is
useful, but too many students remain in the middle during weeks
0 and 1. Refinement will continue with the goal being to move
those students toward one end of the scale or the other.

The goal remains to provide valuable, timely information to the
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 red/yellow/green
 categorization, a numeric score, or a percentage chance of
withdrawal. Nonetheless, the model will continue to be refined
even after initial implementation.

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
http://www.phoenix.edu/about_us/publications/academic-
annual-report/2010.html
Success in Online Database Courses. American Journal of
Business Education. 3(12), 1-5.
Supporting Courses Predict Their Performances and
Behaviors in Primary Courses? Research and Teaching in
Developmental Education. 23(2), 38-48.
students who enroll and succeed in Psychology web-based
classes. J. Educational Psychology. 92(1), 137-143.
perceptions of effective elements in online learning.
Quarterly Review of Distance Education. 5(4), 277-291.
a web-based learning environment: lessons learned.
Quarterly Review of Distance Education. 5(3), 157-171.
Posts”) Predict Student Success in Online Courses: A
Double Cross-Validation Study. Computers & Education.
50(4), 1174-1182.
discriminant function analysis. Paper presented at the 39th
Annual Meeting of the Research and Planning Group, Lake
Arrowhead CA.
retention in online general education courses. American
Journal of Distance Education. 19(1), 23-26
with linear discriminant analysis in their classification
accuracy. J. Experimental Education. 67, 265 – 286.