Predictive Modeling to Improve Retention of Online Students

Rajeev Bukralia

Dakota State University, bukraliar@pluto.dsu.edu

Follow this and additional works at: http://aisel.aisnet.org/mwais2009

Recommended Citation

http://aisel.aisnet.org/mwais2009/12

This material is brought to you by the Midwest (MWAIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in MWAIS 2009 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Predictive Modeling to Improve Retention of Online Students

Rajeev Bukralia
Dakota State University
bukraliar@pluto.dsu.edu

ABSTRACT
Predictive modeling can help identify students who may be “at-risk” to drop out from an online course. An early detection of academically at-risk students will allow instructors and advisors to proactively use appropriate retention strategies. The purpose of this paper is to perform a study to analyze variables and construct a predictive model uniquely suited to identify students who may be more likely to drop out from an online course.

Keywords
Predictive modeling, student retention, online education, academic analytics.

INTRODUCTION
Student retention is a key issue in higher education. It is important to identify at-risk students – or those students who have a greater likelihood of dropping out of a course or program – as it can allow instructors and advisors to proactively implement appropriate retention strategies. This study focuses on identifying and analyzing variables that can influence retention in online courses and comparing predictive modeling techniques to construct a predictive model uniquely suited to identify students who may be more likely to drop out from an online course. The study explores several questions in order to construct this model. For example, what are the variables that may have a greater influence on student retention in an online course? Are some variables better than others in predicting at-risk online students? What is an optimized combination of variables to maximize the accuracy of the prediction? Which predictive models are more accurate in identifying at-risk students? The scope of this study is to build a predictive model that can help improve retention for online courses, rather than for completion of a program or graduation.

LITERATURE REVIEW
The literature review has two primary goals: researching variables relevant for retention and studying various models and predictive modeling techniques used to identify at-risk students. There are possibly dozens of variables (predictors) that may be relevant to accurately identify at-risk students. There are studies that have investigated the role of academic and non-academic variables in student retention (Lotkowski, Robbins, and Noeth, 2004; Campbell and Oblinger, 2007). Variables such as high school grade point average (HSGPA), ACT scores, and socioeconomic status (SES) had a positive relationship to college retention, the strongest being HSGPA, followed by SES and ACT scores. The overall relationship to college retention was strongest when SES, HSGPA, and ACT scores were combined with institutional commitment, academic goals, social support, academic self-confidence, and social involvement. (Lotkowski et al., 2004). Institutions such as Northern Arizona University (Campbell, 2008) have developed graduation prediction and retention models. Such a system uses a predictive model with many variables from SIS to identify at-risk students and to track students to see whether a student successfully graduated within a certain period of time. However, the focus of this study is to identify at-risk students in online courses, rather than identifying them based on their likelihood of graduating in a specific number of years.

There is a need for a model that is uniquely suited to predict the retention of online students because the demographics of online students are somewhat different than those of face-to-face courses. Some administrators and faculty members attribute the lower retention rates in distance-education courses to demographics, since distance-education students are often older, and thus busier, than traditional college students (Carr, 2000).

1 Amit Deokar and Surendra Sarnikar contributed to this paper as well. They are not listed as authors because they served on the conference committee for the MWAIS 2009 conference.
Many studies that were conducted to identify high-risk students used statistical models based on logistic regression (Willging and Johnson, 2004; Pittman, 2008). A couple of retention studies have used machine learning techniques (such as neural networks and decision trees) in addition to logistic regression, which provided the baseline comparison to determine the relative effectiveness of data mining methods (Pittman, 2008). Both logistic regression and neural networks were similar in performance (Campbell, 2008); however this study was not meant specifically for online students.

METHODOLOGY & DATASET

De-identified information was gathered about students from undergraduate online courses taught at Black Hills State University. Each variable was used separately to predict the outcome (likelihood of a student to drop out of a course). Groups of several variables were compared to determine whether a particular combination of variables improves accuracy of predicting at-risk students. The following variables from the SIS were identified for analysis: ACT score, HSGPA, current college GPA, past withdrawals (previous history of dropping a course), credit hours completed, financial aid status, course registration status (final grade posted or withdrew), degree seeking status, gender, and age.

DATA ANALYSIS

A dataset of 365 student records from online courses was analyzed using binary logistic regression in Minitab. Each record was complete in terms of variables examined for all records analyzed. The following table shows the regression table when all 9 independent variables were analyzed together to predict the binary response variable. The binary response variable is course registration status (whether the student did or did not complete the course).

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>Z</th>
<th>P</th>
<th>Odds Ratio</th>
<th>95% CI Lower</th>
<th>95% CI Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.46475</td>
<td>1.73555</td>
<td>-2.57</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSGPA</td>
<td>0.356610</td>
<td>0.333404</td>
<td>1.07</td>
<td>0.285</td>
<td>1.43</td>
<td>0.74</td>
<td>2.75</td>
</tr>
<tr>
<td>ACT score</td>
<td>-0.0487907</td>
<td>0.0436650</td>
<td>-1.12</td>
<td>0.264</td>
<td>0.95</td>
<td>0.87</td>
<td>1.04</td>
</tr>
<tr>
<td>Current college GPA</td>
<td>1.01629</td>
<td>0.224648</td>
<td>4.52</td>
<td>0.000</td>
<td>2.76</td>
<td>1.78</td>
<td>4.29</td>
</tr>
<tr>
<td>Credit hours completed</td>
<td>0.0040243</td>
<td>0.0034536</td>
<td>1.17</td>
<td>0.244</td>
<td>1.00</td>
<td>1.00</td>
<td>1.01</td>
</tr>
<tr>
<td>Past withdrawals</td>
<td>-0.104228</td>
<td>0.0455031</td>
<td>-2.29</td>
<td>0.022</td>
<td>0.90</td>
<td>0.82</td>
<td>0.99</td>
</tr>
<tr>
<td>Financial aid</td>
<td>0.808010</td>
<td>0.331170</td>
<td>2.44</td>
<td>0.015</td>
<td>2.24</td>
<td>1.17</td>
<td>4.29</td>
</tr>
<tr>
<td>Age</td>
<td>0.0136819</td>
<td>0.0389298</td>
<td>0.35</td>
<td>0.725</td>
<td>1.01</td>
<td>0.94</td>
<td>1.09</td>
</tr>
<tr>
<td>Degree seeking status</td>
<td>1.57468</td>
<td>0.628487</td>
<td>2.51</td>
<td>0.012</td>
<td>4.83</td>
<td>1.41</td>
<td>16.55</td>
</tr>
<tr>
<td>Gender</td>
<td>0.410000</td>
<td>0.348407</td>
<td>1.18</td>
<td>0.239</td>
<td>1.51</td>
<td>0.76</td>
<td>2.98</td>
</tr>
</tbody>
</table>

The analysis shows that degree seeking status (P=0.012), financial aid status (P=0.015), and current college GPA (P=0) were strong predictors; while age (P=0.725), gender (P=0.239), and HSGPA (0.285) were found to be weak predictors at 95% CI. Some studies (Lotkowski et al., 2004; Reason 2003) suggested that HSGPA and ACT assessment scores are strong predictors of overall retention in terms of degree completion. The results of binary logistic regression in this study suggest that they are weak predictors in online course completion. The dataset reflected that online courses have a higher percentage of non-traditional, adult learners. The ACT score and HSGPA may not be good predictors for older learners. The first year GPA was found to be a statistically significant predictor of freshman retention (Kiser and Price, 2007). This study suggests that current college GPA is a statistically significant predictor even for online students and not just freshmen. Since many non-traditional students did not have HSGPA and ACT scores in their records, the dataset was also analyzed without those two variables. The results suggest that degree seeking status and current GPA (P<0.05) were still strong predictors, while age and gender...
remained weak predictors. The financial aid status (P=0.270) was no longer a statistically significant predictor, which may be because the dataset had more non-traditional students when HSGPA and ACT score were removed from the analysis. Many non-traditional students do not have financial aid. Based on previous studies (Reason, 2003; Carr 2000), it was hypothesized that age would be a significant factor in online course retention. The findings to this study do not support this hypothesis, as age was the weakest predictor. Another aim of this study was to find an optimized combination of variables to maximize the accuracy of the prediction. The analysis was conducted with various combinations of variables. The analysis showed the prediction accuracy was highest (with 77.4% concordant pairs; Hosmer-Lemeshow goodness of fit = 0.442) when all nine variables were included to predict the binary response variable.

FUTURE WORK

The next step of this research-in-progress study is to compare the logistical regression model with the neural network model in terms of accuracy to predict at-risk students. Some studies show that psychological and technological readiness may also influence student retention in online environments (Liu, 2007). The future scope of this study may include testing this hypothesis by including student surveys and adjusting the predictive model, if necessary. The predictive model can be used to build a computerized system to identify academically at-risk students enrolled in online courses. This study may be extended to include data from other institutions.

REFERENCES