A Data-Mining Approach to Differentiate Predictors of Retention

Chong Ho Yu, Samuel DiGangi, Angel Jannasch-Pennell, Wenjuo Lo, Charles Kaprolet

Applied Learning Technology Institute

Arizona State University

2007 February

Paper presented at Educause Southwest Conference, Austin, TX

RUNNING HEAD: Data mining for retention

Correspondence:

Chong Ho Yu, Ph.D.s

PO Box 612

Tempe AZ 85280

Email: chonghoyu@gmail.com

A Data-Mining Approach to Differentiate Predictors of Retention

Chong Ho Yu, Samuel DiGangi, Angel Jannasch-Pennell, Wenjuo Lo, Charles Kaprolet

Literature Review

Student retention is an important issue for all university administrators and faculty due to the potential negative impact of student attrition. Universities with high attrition rates face the substantial loss of tuition, fees, and potential alumni contributions (DeBerard, Spielmans, & Julka, 2004), while the students themselves also face negative consequences. According to the U.S. Department of Education, National Center for Education Statistics (NCES), students who leave college early are likely to earn less income over their lifetimes when compared to peers who have graduated (NCES, 1989). Despite the identified consequences of college dropout for universities and students, as well as concentrated efforts from all educational institutions on improving student retention, attrition rates remain relatively high across the United States. Data from the National Center for Public Policy and Higher Education (NCPPHE) reveal that only 73.6 percent of first-time, full-time freshmen (enrolled in 2002) returned for their second semester (2007). Looking at college completion data from 2005, only 39.5 percent of undergraduate students enrolled in public institutions completed their degrees within five years (ACT, 2005).

In discussing retention statistics, it is important to explore the definition and methods for calculating persistence and retention. Freshman persistence is commonly defined as returning to regular enrollment status in the first semester of the sophomore year and is strongly associated with the likelihood of eventual graduation from the institution (Mallinckrodt & Sedlacek, 1987). While it is encouraging to see that students who persist following their freshman year alone are

more likely to graduate, it is discouraging because this is where approximately 73.6 percent of students are lost (NCPPHE, 2007). Fortunately, this can guide potential interventions. On the other hand, retention rates are generally calculated based on data from first-time, full-time freshman students who graduate within six years of their initial enrollment date (Hagedorn, 2005).

In this study, retention rates will be studied with data from sophomore students who initially enrolled in the 2002 academic year, following these students through their junior year. Since online courses are a relatively new option for students, freshman year data from this cohort of students could not be analyzed, as there were few online courses available to those students.

Tinto's widely accepted model of student retention (1975) examines factors contributing to a student's decision to continue their higher education. The primary focus of this model is a student's academic and social integration into the university. Tinto argued that from an academic perspective, performance, personal development, academic self-esteem, enjoyment of courses, and identification with academic norms and one's role as a student all contribute to a student's overall sense of integration into the university (1975). The argument is that students who are highly integrated academically are more likely to persist and complete their degrees. The same is true from a social perspective. Students who have more friends at their institution, have more personal contact with academics, and maintain an overall sense of enjoyment with being enrolled at the institution contribute socially to a decision to persist.

Using Tinto's model as a basis for understanding student retention and persistence, this study attempts to determine whether the abundance of online courses and degree programs pose a threat to academic and social integration. Intuitively, students who take a majority of courses online risk being removed from the academic atmosphere and may lack personal interaction with

faculty, two factors that were critical to Tinto's model (1975). This study is unique because of its sole focus on online students, who have largely not been studied thus far. The majority of available retention research on online courses studies retention rates for individual online classes (Nash, 2005; Schrum & Hong, 2002), but as universities are offering an increasing number of online courses, it is important to determine whether these courses impact retention.

Of the existing research, factors have been identified that are hypothesized to potentially impact the retention of online students. Allen (2006) suggested that online courses potentially distance students from academic integration, social integration, and the overall on-campus experience, which are critical to Tinto's model. Another study identified necessary practices for ensuring high retention rates among online students, citing retention rates of over 80% for their online programs (Schrum & Hong, 2002). Participants in this study reported that students for their online degree programs are very carefully selected, going through interviews aimed at identifying potential success factors in the candidates (2002). Unfortunately, this study did not provide data on retention rates before and after the implementation of such an interview process. However, other studies have confirmed that impact of admissions criteria on retention rates. Devarics and Roach (2000) found that freshman dropout rates for highly selective to less selective universities can vary from eight to 35 percent. While this may be an important strategy to initially ensure the retention of students seeking online courses, larger, public universities may not be able to adopt such selective screening processes.

Data source and methodology

In this study two data sets are compiled by tracking the continuous enrollment or withdrawal of 2003 and 2005 freshmen enrolled at Arizona State University (ASU). The

dependent variable is a dichotomous variable, retention, whereas there are four sets of potential predictors:

1. Demographic: Gender, Ethnic, Residence (in state/out of state), Location (living on campus/off campus)
2. Pre-college academic performance indicators: High school GPA, High school percentile rank, SAT combined scores, ACT combined scores
3. Enrollment: Accumulated earned hours, online class hours as a percentage of total hours during the sophomore year
4. Academic achievement: ASU math placement test scores, Freshman GPA

Stepwise logistic regression is a popular procedure for selecting variables to build a predictive model for retention. Typically, institutional analysis deals with an ocean of data. This study is not an exception. The ASU DataWarehouse provides the research team with over 10,000 subjects spanning across 2002 to 2006. When a sample size is this large, any trivial difference may lead to a seemingly significant result that is not actually significant. In addition, in spite of its popularity, stepwise logistic regression has certain insurmountable problems. For example, while reporting the odds ratio is a common practice to indicate the ratio between the desirable and undesirable events (e.g. retained vs. not retained), the Mantel-Haenszel odds ratio is considered valid only under the assumption that the underlying stratum-specific odds ratios are constant across the strata (Greenland, 1989). Stepwise regression as a tool for variable selection has also been under severe criticism. It was found that stepwise regression tends to yield conclusions that cannot be replicated because this model-building approach capitalizes on sampling error (Thompson, 1995). It is also a well-known fact that the results of stepwise regression are affected by the order of entry of the variables (Glymour, 2001).

As a remedy, classification trees in data mining are used here to examine the variables related to retention. It is important to note that data mining focuses on pattern recognition, hence no probabilistic inferences and Type I error are involved. Also, unlike regression that returns a subset of variables, classification trees can rank order the factors that affect the retention rate.

In this study JMP (SAS Institute, 2006) is employed to construct classification trees based upon Entropy (Quinlan, 1993) as the tree-splitting criterion, which favors balanced or similar splits. Splitting criteria are measures of node "impurity" that determines where to make a split. It is based on the estimated probabilities from the node proportions. Ideally, we would like to partition data in a way that each partition is pure, which means that in a partition data vectors, in which each element represents a variable, should come from a single, homogeneous class. However, this rarely occurs and thus some degree of "impurity" must be expected (Han & Kamber, 2006).

Fitness versus parsimony is pervasive in every type of modeling, but there is a strong rationale for favoring simplicity. To explain an observed phenomenon based upon the data at hand, the best mode is the one that reaches the highest degree of model-data fit for its ample explanatory power. However, the merit of predictive models, such as classification trees, is tied to its accuracy on unseen data. For the same or similar accuracy, smaller numbers of nodes, which means that the tree is less complex, can work better with unseen data. Conversely, a tree uses splitting variables with large numbers of values, thus yielding more nodes that can result in a negative impact on unseen data (Rosella, 2007).

Results

Table 1 and Figure 1-3 show the crucial variables of predicting retention suggested by the classification trees using the two data sets. In both data sets, "cumulated earned hours" is

identified as the most crucial factor contributing to retention. Although there is a slight disagreement between 2004 and 2005 results regarding which variables should be retained and their relative importance, both results show that residence and living location play an important role in retention.

Figure 1. Classification tree of 2004 data set.

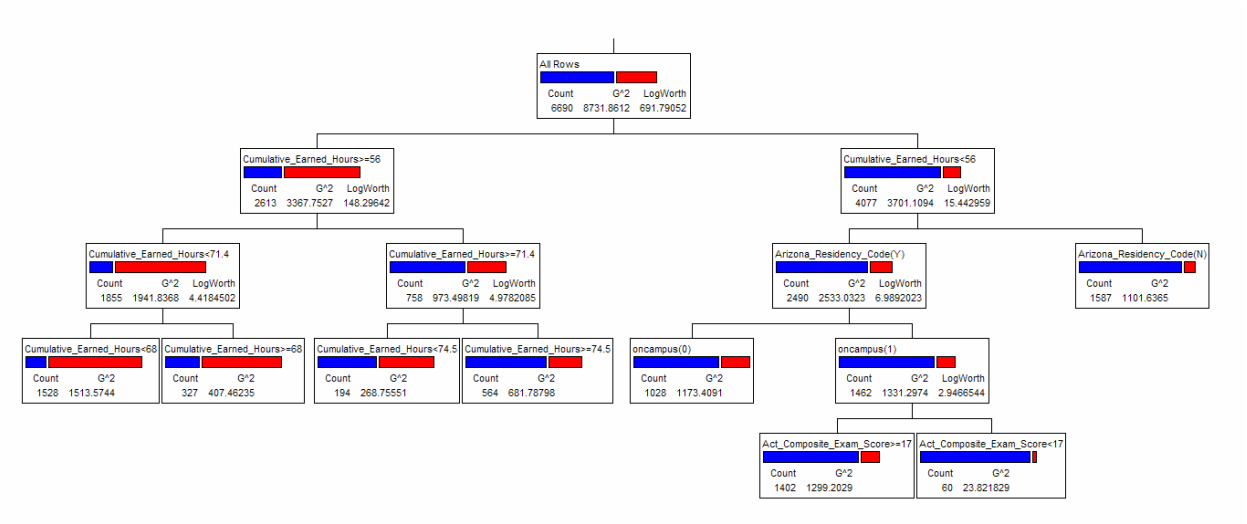

Figure 2. Classification tree of 2005 data set.

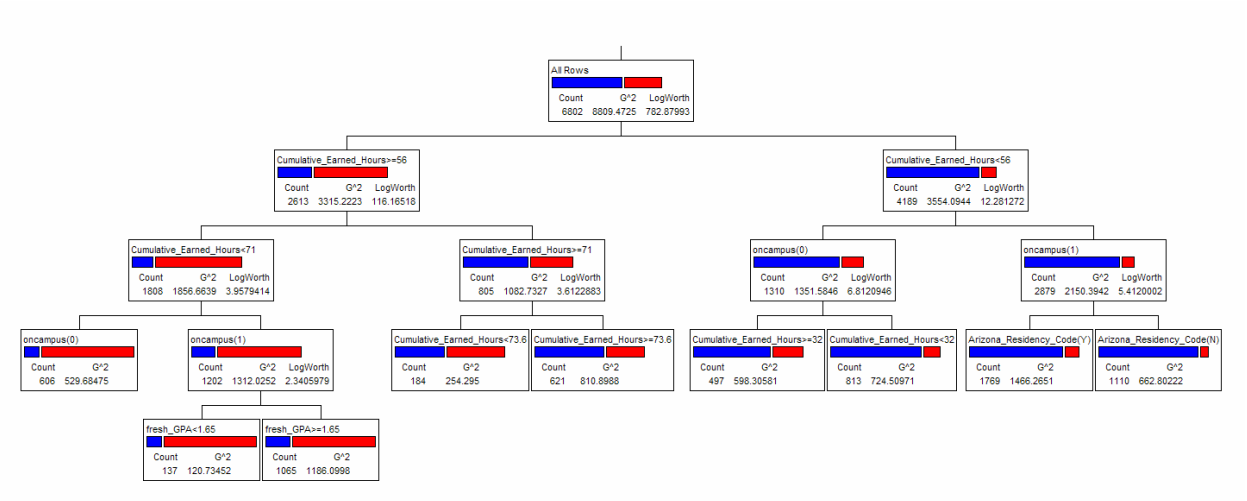

Table1.

Variables retained by the classification trees

		2004	2005
Demographic	Gender		
	Ethnic		
	Residence	2	3
	Living location	3	2
Pre-college academic performance indicators	High school GPA		
	High school percentile rank		
	SAT combined scores		
	ACT combined scores	4	
Enrollment	Accumulated earned hours	1	1
	Online class hours percentage		
Academic achievement	ASU math placement test scores		
	Freshman GPA		4

Discussion

While it is not surprising to learn that retention is tied to “earned hours,” it is out of our expectation to see that retention is strongly tied to “spatial” factors, including residence (in state/out of state) and living location (on campus/off campus). Possible explanations are that students who are not Arizona residents pay higher tuition; consequently, it drains their financial resources that could have been deployed to support their study. In addition, out of state students might spend more time in traveling back and forth between their hometown and the university, and as a result the burden of traveling time and expenses affect their academic performance. Further, out-of-state students might not have emotional support from their parents and thus they might easily give up their study while facing adversaries. Last, students who live off-campus

might not have immediate access to university resources (e.g. computer labs, library) and therefore their retention rate is lower than those who live on-campus. Since the Internet boom, the US higher education sector has been moving towards increasing numbers of online classes. The rationale is that advanced communication technology enables students to learn anytime, anywhere. Hence, how online courses can be developed to break the "spatial barriers" and thus to improve retention should be further investigated by educational researchers.

References

ACT, Inc. (2005). *National collegiate retention and persistence to degree rates.* Retrieved February 5, 2007, from http://www.act.org/path/policy/pdf/retain_2005.pdf

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth International Group.

DeBerard, M. S., Spielmans, G. I., & Julka, D. C. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. College Student Journal, 38(1), 66-80.

Devarics, C., & Roach, R. (2000). Fortifying the federal presence in retention. *Black Issues in Higher Education, 17,* 20-25.

Glymour, C. (2001). *Mind's arrows: Bayes and Graphical causal models in psychology*. Cambridge, MA: MIT Press.

Greenland, S. (1989). Modeling and variable Selection in epidemiologic analysis. *American Journal of Public Health, 79,* 3, 340-349.

Hagedorn, L. S. (2005). How to define retention. In Alan Seidman (Ed.). *College student retention: Formula for student success.* Westport, CT: Praeger Publishers.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques (2nd ed.)*. Boston, MA: Elsevier.

Insightful, Inc. (2006). *Splus Insightful Miner* [Computer software and manual]. Seattle, WA: The Author.

Mallinckrodt, B., & Sedlacek, W. E. (1987). Student retention and the use of campus facilities by race. *NASPA Journal, 24,* 28-32.

Nash, R. D. (2005). Course completion rates among distance learners: Identifying possible methods to improve retention. *Online Journal of Distance Learning Administration, 8,* 1.

National Center for Public Policy and Higher Education. (2007). *Retention rates - First-time college freshmen returning their second year.* Retrieved February 6, 2007, from http://www.higheredinfo.org/dbrowser/index.php?measure=67

Schrum, L., & Hong, H. (2002). Dimensions and strategies for online success: Voices from experienced educators. *Journal of Asynchronous Learning Networks, 6,* 57-67.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45,* 89-125.

U.S. Department of Education National Center for Educational Statistics. (1989). Digest of Educational Statistics (25th ed.). Washington DC: US Department of Education.

Luan, J. (2002). Data mining and its applications in higher education. In A. Serban & J. Luan (Eds.), *Knowledge management: Building a competitive advantage in higher education* (pp. 17-36). PA: Josey-Bass.

Quinlan, J. R. (1993). *C4.5 programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Rosella, Inc. (2007). Decision tree classification software. Retrieved January 29, 2007 from http://www.roselladb.com/decision-tree-classification.htm

SAS Institute. (2006). *JMP* [Computer software and manual]. Cary, NC: The Author.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*, 4, 525-534.