J. COLLEGE STUDENT RETENTION, Vol. 13(1) 17-35, 2011-2012

PREDICTING STUDENT ATTRITION WITH DATA MINING METHODS

DURSUN DELEN, PH.D
Oklahoma State University

ABSTRACT

Affecting university rankings, school reputation, and financial well-being, student retention has become one of the most important measures of success for higher education institutions. From the institutional perspective, improving student retention starts with a thorough understanding of the causes behind the attrition. Such an understanding is the basis for accurately predicting at-risk students and appropriately intervening to retain them. In this study, using 8 years of institutional data along with three popular data mining techniques, we developed analytical models to predict freshmen student attrition. Of the three model types (artificial neural networks, decision trees, and logistic regression), artificial neural networks performed the best, with an 81% overall prediction accuracy on the holdout sample. The variable importance analysis of the models revealed that the educational and financial variables are the most important among the predictors used in this study.

INTRODUCTION AND MOTIVATION FOR THE PROBLEM

Student attrition has become one of the most challenging problems for academic institutions. The loss of students usually results in overall financial loss, lower graduation rates, and inferior school reputation in the eyes of stakeholders (Gansemer-Topf & Schuh, 2006). The legislators and policymakers who oversee higher education and allocate funds, the parents who pay for their children's

doi: 10.2190/CS.13.1.b
http://baywood.com

education in order to prepare them for a better future, and the students who make college choices look for evidence of institutional quality and reputation to guide their decision-making processes.

The principal motivations for improving student retention are the economic and the social benefits of attaining a higher education degree (Thomas & Galambos, 2004), both for individuals and for the public. In general terms, the economic and social attributes that motivate individuals to enter higher education are:

1. *public economic benefits*: increased tax revenues, greater productivity, increased consumption, increased workforce flexibility, and decreased reliance on government financial support;
2. *individual economic benefits*: higher salaries and benefits, employment, higher savings levels, improved working conditions, and personal/professional mobility;
3. *public social benefits*: reduced crime rates, increased charitable giving/community service, increased quality of civic life, social cohesion/appreciation of diversity, and improved ability to adapt to and use technology; and
4. *individual social* benefits: improved health/life expectancy; improved quality of life for offspring; better consumer decision making; increased personal status; and more hobbies, leisure activities (Hermanowicz, 2003).

Traditionally, student attrition at a university has been defined as the number of students who do not complete a degree in that institution. Studies have shown that more students withdraw during their first year of college than during the rest of their higher education (Deberard, Julka, & Deana, 2004; Hermanowicz, 2003; Pascarella, Terenzini, & Wolfle, 1986). Since most of the student dropouts occur at the end of the first year (the freshmen year), the majority of student attrition studies (including this study) have focused on first year dropouts or the number of students not returning for the second year. This definition of attrition does not differentiate between the students who may have transferred to other universities and obtained their degrees there. It only considers the students dropping out at the end of the first year voluntarily and not by academic dismissal.

Research on student retention has traditionally been survey driven (e.g., surveying a student cohort and following them for a specified period of time to determine whether they continue their education) (Caison, 2007). Using such a design, researchers worked on developing and validating theoretical models including the famous student integration model developed by Tinto (1993). Elaborating on Tinto's theory, others have also developed student attrition models using survey-based research studies (Berger & Braxton, 1998; Berger & Milem, 1999). Even though they have laid the foundation for the field, these survey-based research studies have been criticized for their lack of generalized

applicability to other institutions and the difficulty and costliness of administering such large-scale survey instruments (Cabrera, Nora, & Castaneda, 1993). An alternative (and/or a complementary) approach to the traditional survey-based retention research is an analytic approach where the data commonly found in institutional databases is used. Educational institutions routinely collect a broad range of information about their students, including demographics, educational background, social involvement, socioeconomic status, and academic progress. A comparison between the data-driven and survey-based retention research showed that they are comparable at best, and to develop a parsimonious logistic regression model, data-driven research was found to be superior to its survey-based counterpart (Caison, 2007). But in reality, these two research techniques (one driven by the surveys and theories, the other driven by institutional data and analytic methods) complement and help each other. That is, the theoretical research may help identify important predictor variables to be used in analytical studies while analytical studies may reveal novel relationships among the variables which may lead to development of new and betterment of the existing theories.

Data mining is the process of extracting valuable knowledge (i.e., non-trivial, logical, previously unknown, and potentially useful patterns) from a large amount of data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Even though it is a relatively new concept, it has been successfully applied to complex problems in areas such as medicine, healthcare, homeland security, transportation, finance, marketing, and entertainment. In this project, we apply data mining to higher education, specifically to the problem of student attrition.

In order to improve student retention, one should understand the non-trivial reasons behind the attrition. To be successful, one should also be able to accurately identify those students that are at risk of dropping out. So far, the vast majority of student attrition research has been devoted to understanding this complex, yet crucial, social phenomenon. Even though, these qualitative, behavioral, and survey-based studies revealed invaluable insight by developing and testing a wide range of theories, they do not provide the much needed instrument to accurately predict (and potentially improve) the student attrition (Caison, 2007). In this project we propose a quantitative research approach where the historical institutional data from student databases are used to develop models that are capable of predicting as well as explaining the institution-specific nature of the problem of attrition. Though the concept is relatively new to higher education, for almost a decade now, similar problems in the field of marketing have been studied using predictive data mining techniques under the name of "churn analysis," where the purpose is to identify the customers who are most likely to leave the company so that some kind of intervention can be done to retain them. Retaining existing customers is crucial because the related research shows that acquiring a new customer costs roughly 10 times more than keeping the one that you already have (Berry & Linoff, 2004).

LITERATURE REVIEW

Despite steadily rising enrollment rates in U.S. postsecondary institutions, weak academic performance and high dropout rates remain persistent problems among undergraduates (Caison, 2007; Tinto, 1997). For academic institutions, high attrition rates complicate enrollment planning and place added burdens on efforts to recruit new students. For students, dropping out before earning a terminal degree represents untapped human potential and a low return on their investment in college (Gansemer-Topf & Schuh, 2006; Mannan, 2007). Poor academic performance is often indicative of difficulties in adjusting to college and makes dropping out more likely (Lau, 2003).

A number of academic, socioeconomic, and other related factors are associated with attrition. According to Wetzel et al. (1999), universities which have more open admission policy and where there is no substantial waiting list of applicants and transfers face more serious student attrition problems than universities with surplus applicants. On the other hand, Hermanowicz (2003) found that more selective universities do not necessarily have higher graduation rates, rather other factors not directly associated with selectivity can, in principle, come into play. In addition to the "structural" sides of universities (e.g., admission and prestige of school), the "cultural side" (e.g., norm and values that guide communities) should receive equal attention because a higher rate of retention is often achieved when students find the environment in their university to be highly correlated with their interests (Hermanowicz, 2003).

In related research, Astin (1993) determined that the persistence or the retention rate of students is greatly affected by the level and quality of their interactions with peers as well as faculty and staff. Tinto (1987) indicates that the factors in students' dropping out include academic difficulty, adjustment problems, lack of clear academic and career goals, uncertainty, lack of commitment, poor integration with the college community, incongruence, and isolation. Consequently, retention can be highly affected by enhancing student interaction with campus personnel. Especially for first-generation college students, the two critical factors in students' decisions to remain enrolled until the attainment of their goals are their successfully making the transition to college, aided by initial and extended orientation and advisement programs, and making positive connections with college personnel during their first term of enrollment (Ishitani, 2006).

According to Tinto's (1987) theory of student integration, past and current academic success is a key component in determining attrition. High school GPA and total SAT scores provide insight into potential academic performance of the freshmen and have been shown to have strong positive effect on persistence (Porter, 2008; Tinto, 1993). Similarly, and probably more importantly, first semester GPA has been shown to correlate strongly with retention (Porter, 2008; Vandamme, Meskens, & Superby, 2007). In this study we used these academic success indicators.

Institutional and goal commitment are found to be significant predictors of student retention (Cabrera et al., 1993). Undecided students may not have the same level of mental strength and goal commitment as the students who are more certain of their career path. Therefore, as a pseudo measure of academic commitment, declaration of college major and credit hours carried in the first semester are included in this study. Additionally, students' residency status (classified as either in-state or out-of-state) may be an indicator of social and emotional connectedness as well as better integration with the culture of the institution (Caison, 2007). Students coming from another state may have less familial interaction, which may amplify the feelings of isolation and homesickness.

Several previous studies investigated the effect of financial aid on student retention (Herzog, 2005; Hochstein & Butler, 1983; Stampen & Cabrera, 1986). In these studies, the type of financial aid was found to be a determinant of student attrition behavior. Students receiving aid based on academic achievement have higher retention rates (Stampen & Cabrera, 1986). Hochstein and Butler (1983) found that grants are positively associated with student retention while loans have a negative effect. Similarly, Herzog (2005) found that Millennium Scholarship as well as other scholarships helps students stay enrolled while losing these scholarships because of insufficient grades or credits raises dropout or transfer rates.

In this study, using 8 years of freshmen student data (obtained from the university's existing databases) along with three popular data mining techniques (artificial neural networks, decision trees, and logistic regression), we developed analytical models to predict freshmen attrition. In order to identify the important predictors, we conducted variable importance analyses on these models. Therefore, the main goal of this research was to develop models to correctly identify the freshmen students who are most likely to drop out after their freshmen year. The models that we developed are designed in such a way that the prediction occurs at the end of the first semester (usually at the end of fall semester) in order for the decision makers to properly craft intervention programs (using the explanatory information from the variable importance analyses) during the next semester (the spring semester) in order to retain them.

METHODOLOGY

In this research, we followed a popular data mining methodology called CRISP-DM (Cross Industry Standard Process for Data Mining) (Shearer, 2000), which is a six-step process:

1. understanding the domain and developing the goals for the study;
2. identifying, accessing and understanding the relevant data sources;
3. pre-processing, cleaning, and transforming the relevant data;

4. developing models using comparable analytical techniques;
5. evaluating and assessing the validity and the utility of the models against each other and against the goals of the study, and
6. deploying the models for use in decision- making processes.

This popular methodology provides a systematic and structured way of conducting data mining studies (moving the whole endeavor from an *art* form to a *scientific* experiment), and hence increasing the likelihood of obtaining accurate and reliable results. The sequence of the steps is not strict and moving back and forth between different steps is often inevitable. The attention paid to the earlier steps in CRISP-DM (i.e., understanding the domain of study, understanding data, and preparing the data) sets the stage for a successful data mining study. Roughly 80% of the total project time is usually spent on these first three steps.

In CRISP-DM, the method evaluation step requires comparing the data mining models for their predictive accuracy. Traditionally, in the comparison process the complete dataset is split into two subsets, 2/3 training and 1/3 testing. The models are trained on the training subset and then evaluated on the testing subset. The prediction accuracy on the testing subset is used to report the actual prediction accuracies of all evaluated models. Since the data set is split into two exclusive subsets randomly, there always is a possibility of those two sets not being "equal." In order to minimize this bias associated with the random sampling of the training and testing data samples, we used an experimental design called *k*-fold cross validation. In *k*-fold cross validation, also called rotation estimation, the complete dataset is randomly split into *k* mutually exclusive subsets of approximately equal size. The classification model is trained and tested *k* times. Each time, it is trained on all but one fold and tested on the remaining single fold. The cross validation estimate of the overall accuracy is calculated as simply the average of the *k* individual accuracy measures. A pictorial depiction of this evaluation process is shown in Figure 1. With this experimental design, if the *k* is set to 10 (which is the case in this study and a common practice in most predictive data mining applications), for each of the three model types ten different models are developed and tested, totaling 30 models for this project.

Data Description

The data for this study came from a single institution (a comprehensive public university located in the mid-west region of the United States) with an average enrollment of 23,000 students, of which roughly 80% are the residents of the same state and roughly 19% of the students are listed under some minority classification. There is no significant difference between the two genders in the enrollment numbers. The average freshmen student retention rate for the institution is about 80%, and the average 6 years graduation rate is about 60%.

In this study we used 8 years of institutional data, which entailed 25,224 students enrolled as freshmen between (and including) the years of 1999 and 2006.

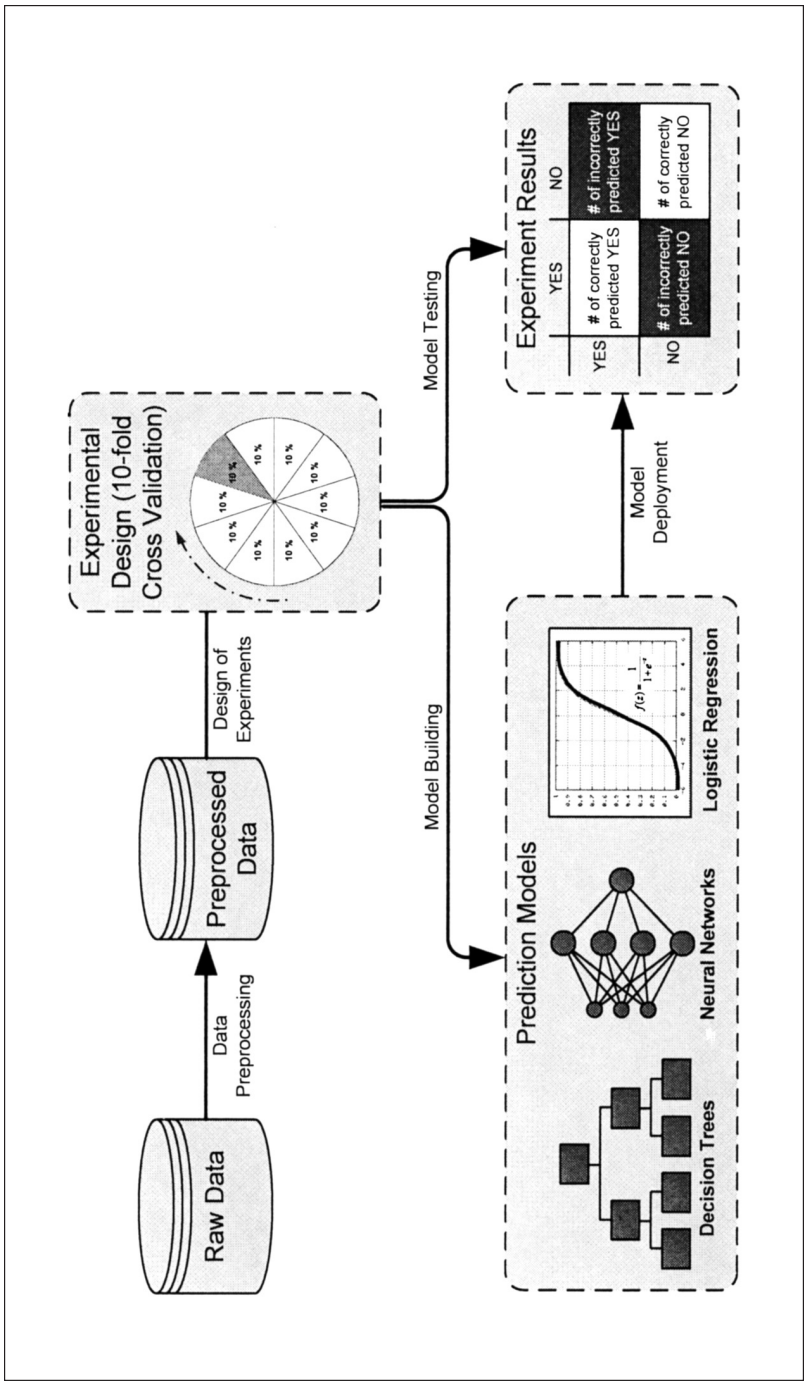

Figure 1. Data mining and cross validation process.

Table 1. Freshmen Student Data Used in This Study

Year	Total freshmen students	Returned for 2nd fall	Freshmen attrition (%)
1999	2785	2189	21.40%
2000	2949	2322	21.26%
2001	3088	2473	19.92%
2002	3190	2555	19.91%
2003	3423	2657	22.38%
2004	3249	2541	21.79%
2005	3306	2604	21.23%
2006	3234	2576	20.35%
	Total: 25224	Total: 19917	Average: 21.03%

The data was collected and consolidated from various university student databases. A brief summary of the number of the records (i.e., freshmen students) by year is given in Table 1.

The data contained variables related to student's academic, financial, and demographic characteristics. A complete list of variables obtained from the student databases is given in Table 2. After converting the multi-dimensional student data into a flat file (a single file with columns representing the variables and rows representing the student records), the file was assessed and preprocessed to identify and remove anomalies and unusable records. For instance, we removed all international student records from the dataset because they did not contain some of the presumed important predictors (e.g., high school GPA, SAT scores). In the data transformation phase, some of the variables were aggregated (e.g., "Major" and "Concentration" variables aggregated to binary variables *MajorDeclared* and *ConcentrationSpecified*) for better interpretation for the predictive modeling. Additionally, some of the variables were used to derive new variables (e.g., *Earned/Registered* (Equation 1) and *YearsAfterHighSchool* (Equation 2)).

$$\text{Earned/Registered} = \frac{\text{EarnedHours}}{\text{RegisteredHours}} \tag{1}$$

$$\text{YearsAfterHighSchool} = \text{FreshmenEnrollmentYear} - \text{HighSchoolGraduationYear} \tag{2}$$

The *Earned/Registered* hours was created to have a better representation of the students' resiliency and determination in their first semester of the freshmen

Table 2. Variables Obtained from Student Records

No.	Variables	Data type
1	College	Multi Nominal
2	Degree	Multi Nominal
3	Major	Multi Nominal
4	Concentration	Multi Nominal
5	Fall hours registered	Number
6	Fall earned hours	Number
7	Fall GPA	Number
8	Fall cumulative GPA	Number
9	Spring hours registered	Number
10	Spring earned hours	Number
11	Spring GPA	Number
12	Spring cumulative GPA	Number
13	Second Fall Registered (Y/N)	Nominal
14	Ethnicity	Nominal
15	Sex	Binary Nominal
16	Residential Code	Binary Nominal
17	Marital Status	Binary Nominal
18	SAT High Score Comprehensive	Number
19	SAT High Score English	Number
20	SAT High Score Reading	Number
21	SAT High Score Math	Number
22	SAT High Score Science	Number
23	Age	Number
24	High School GPA	Number
25	High School Graduation Year and Month	Date
26	Starting Term as New Freshmen	Multi Nominal
27	TOEFL Score	Number
28	Transfer Hours	Number
29	CLEP earned hours	Number
30	Admission Type	Multi Nominal
31	Permanent Address State	Multi Nominal
32	Received Fall Financial Aid	Binary Nominal
33	Received Spring Financial Aid	Binary Nominal
34	Fall Student Loan	Binary Nominal
35	Fall Grant/Tuition Waiver/Scholarship	Binary Nominal
36	Fall Federal Work Study	Binary Nominal
37	Spring Student Loan	Binary Nominal
38	Spring Grant/Tuition Waiver/Scholarship	Binary Nominal
39	Spring Federal Work Study	Binary Nominal

year. Intuitively, one would expect greater values for this variable to have a positive impact on retention. The *YearsAfterHighSchool* was created to measure the impact of the time taken between high school graduation and initial college enrollment. Intuitively, one would expect this variable to be a contributor to the prediction of attrition. These aggregations and derived variables are determined based on a number of experiments conducted for a number of logical hypotheses. The ones that made more common sense and the ones that led to better prediction accuracy were kept in the final variable set.

Reflecting the population, the dependent variable (i.e., "Second Fall Registered") contained many more yes records (80%) than no records (20%). Based on our preliminary experimental results and the machine learning heuristics (Wilson & Sharda, 1994), the data set was balanced to include an equal proportion of yes and no records, for a final data set size of 6,454.

Prediction Models

In this study, three popular classification techniques are used: artificial neural networks, decision trees, and logistic regression. These prediction techniques were selected because of their popularity in the recently published literature. A large number of studies compare data mining methods in different settings. Most of these previous studies found machine-learning methods (e.g., artificial neural networks and decision trees) to be superior to their statistical counterparts (e.g., logistic regression and discriminant analysis) in terms of both being less constrained by assumptions and producing better prediction results (Kiang, 2003; Law, 2000; Lim, Loh, & Shih, 2000; Sharda & Delen 2006). Our findings in this study confirm these results.

Artificial Neural Networks

Artificial neural networks (ANN) are biologically inspired, highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions. Formally defined, neural networks are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data (Haykin, 1998). In this study we used a popular neural network architecture called multi-layer perceptron (MLP) with a back-propagation, supervised learning algorithm. MLP, a strong function approximator for prediction and classification problems, is arguably the most commonly used and well-studied ANN architecture. Hornik et al. (1990) empirically show that given the right size and structure, MLP is capable of learning arbitrarily complex nonlinear functions to an arbitrary accuracy level. MLP is essentially the collection of nonlinear neurons (perceptrons) organized

and connected to each other in a feed-forward multi-layered structure. The MPL type of ANN architecture used in this study is graphically shown in Figure 2.

Decision Trees

Decision trees are powerful classification algorithms that are becoming increasingly popular with the growth of data mining in the information systems field. Popular decision tree algorithms include Quinlan's (1986, 1993) ID3, C4.5, C5, and Breiman et al.'s (1984) CART (Classification and Regression Trees) and CHAID (CHi-squared Automatic Interaction Detector). As the name implies, all decision tree techniques recursively separate observations into branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms (e.g., information gain, Gini index, and Chi-squared test) to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed.

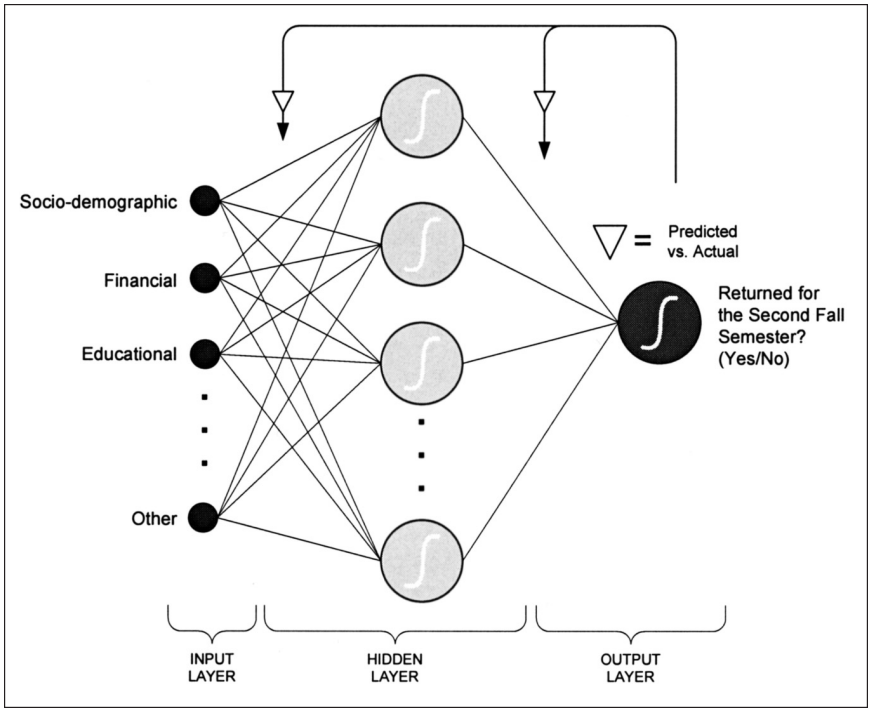

Figure 2. MLP type artificial neural network architecture used in this study.

The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (increasing the discrimination) of the resulting two or more subgroups of samples. The most commonly used mathematical algorithm for splitting includes Entropy-based information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-Squared test (used in CHAID). In this study we used the C5 algorithm (Quinlan, 1993) implementation in SPSS Clementine (SPSS, 2008), which uses an improved version of C4.5 and ID3 algorithms.

Logistic Regression

Logistic regression is a generalization of linear regression. It is used primarily for predicting binary or multi-class dependent variables. Because the response variable is discrete, it cannot be modeled directly by linear regression. Therefore, rather than predicting a point estimate of the event itself, it builds the model to predict the odds of its occurrence. In a two-class problem, odds greater than 50% means that the case is assigned to the class designated as "1," and "0" otherwise. While logistic regression is a very powerful modeling tool, it assumes that the response variable (the log odds, not the event itself) is linear in the coefficients of the predictor variables. Furthermore, the modeler, based on his or her experience with the data and data analysis, must choose the right inputs and specify their functional relationship to the response variable.

RESULTS

Based on the 10-fold cross validation results, the artificial neural network model was able to classify freshmen students with an overall accuracy rate of 81.19%. The model was more accurate in classifying those students who did not return for the sophomore year (i.e., attrition; 93.83%), than those who did return (i.e., retention; 68.55%). The lower accuracy in the false positive rate (incorrectly identifying students as potential attrition) may be preferable to having lower accuracy in the false negative rate (incorrectly identifying students as not potential attrition). The decision tree model was 78.25% accurate overall in classifying students into attrition or retention groups. Similar to the artificial neural network model, the decision tree also had superior performance in classifying students who are likely to drop out (92.53% accuracy) then the ones who are not (63.96% accuracy). Overall, the artificial neural network model outperformed the decision tree model, as the overall accuracy rate and the accuracy rates for each class of the dependent variable for the ANN model exceed those for the decision tree model. Both artificial neural networks and decision tree models have surpassed the prediction accuracy obtained with the logistic regression model. Table 3 shows the overall accuracy for each model, and the accuracy for each class of dependent variable.

Table 3. 10-Fold Cross Validation-Based Prediction Results

	Neural Network		Decision Tree		Logistic Regression	
	Yes	No	Yes	No	Yes	No
Model Predictions						
Yes	2,212	199	2,064	241	2,043	473
No	1,015	3,028	1,163	2,986	1,184	2,754
Per-Class Classification Accuracy (in %)	68.55	93.83	63.96	92.53	63.31	85.34
Overall Classification Accuracy (in %)	81.19		78.25		74.33	

In addition to assessing accuracy for each model, relevant features were examined to determine the important variables in the model. For the ANN model, sensitivity analysis was used; sensitivity analysis rates predictor variables according to the deterioration in modeling performance that occurs if that variable is no longer available to the model. The basic measure of sensitivity of a predictor variable is calculated as the ratio of the error of the model without the inclusion of the variable to the error of the model that included the variable. The more sensitive the network is to a particular input, the greater the deterioration one can expect, and therefore the greater the ratio. The shortcoming of this approach is that it assumes the independent contribution of variables to the outcome of the model, which may not hold true in situations with interdependent variables that are important only if included as a set. The sensitivity results of the ANN model are shown in Figure 3. In Figure 3, the y-axis shows the normalized relative importance of each variable while the x-axis lists the independent variables in the order of importance from left (most important) to right (the least important).

For the decision tree model, variable importance measures can be inferred from the decision tree structure. The decision tree is constructed in a top down fashion using an entropy-based information gain measure at each consecutive node/branch. At the top of the tree, the most discriminative variable is determined based on the information gain, and the tree is split into two mutually exclusive branches. This process is repeated for each branch until the stopping criterion is reached, which is commonly the unbiased prediction accuracy on a holdout sample. The higher on the tree a variable is shown; the more importance is given to that variable. A partial tree generated for this study is shown in Figure 4.

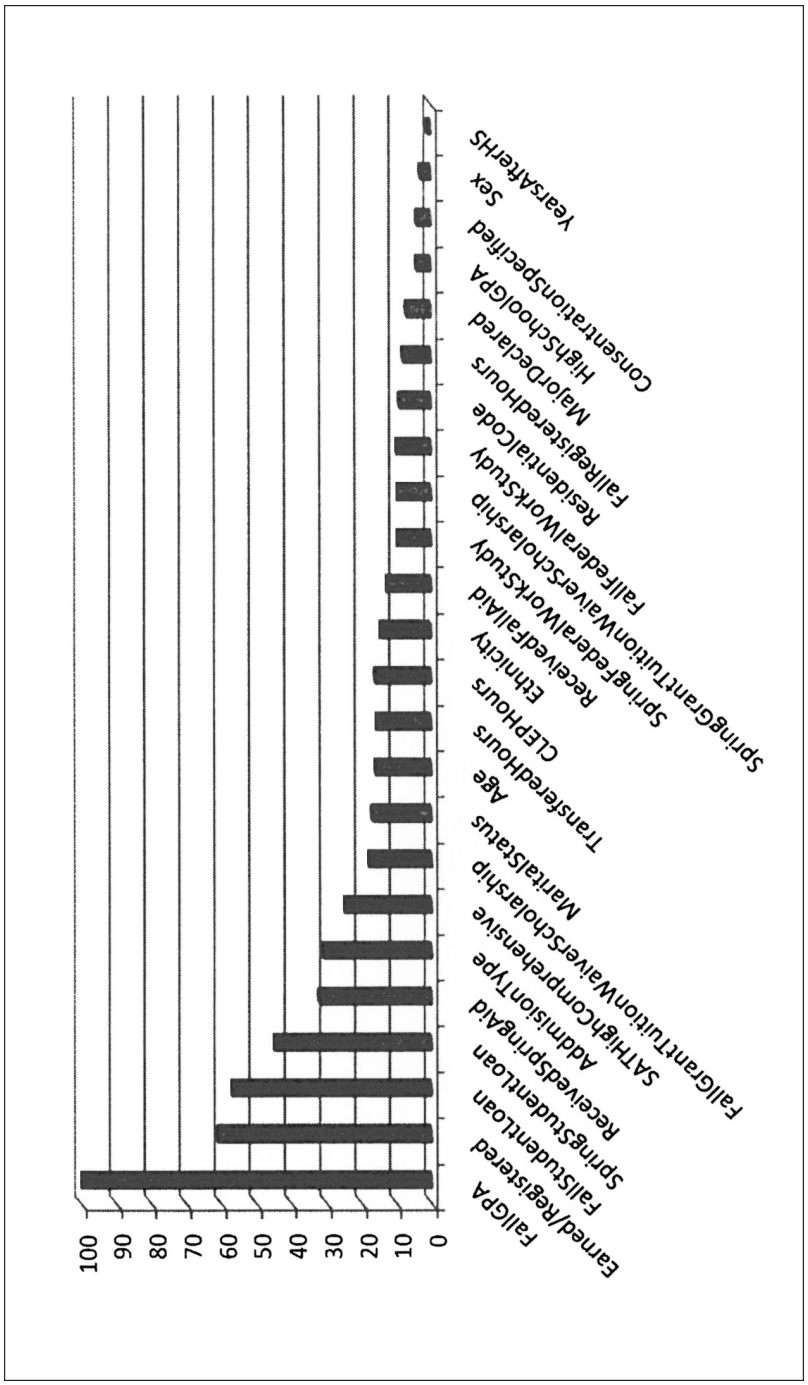

Figure 3. Graphical representation of the ANN sensitivity analysis results.

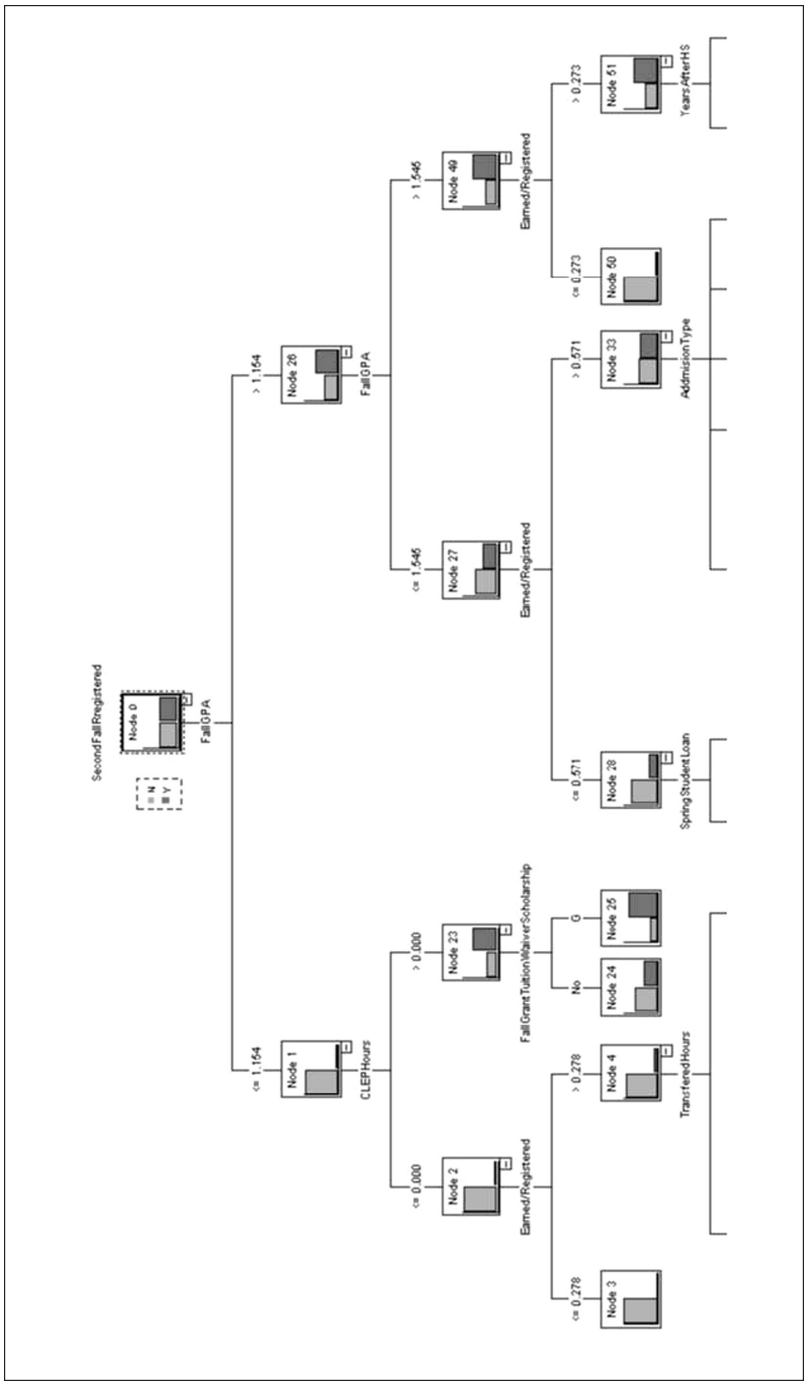

Figure 4. A partial pictorial representation of the C5 algorithm-based decision tree.

DISCUSSION AND CONCLUSION

Our results show that, given sufficient data with the proper variables, data mining methods are capable of predicting freshmen student attrition with roughly 80% accuracy. Among the three prediction methods compared in this study, artificial neural networks performed the best, followed by decision trees and logistic regression. From the usability standpoint, despite the fact that artificial neural networks had better prediction results, one might chose to use decision trees because compared to neural networks, they portray a more transparent model structure. Decision trees explicitly show the reasoning process of different prediction outcomes, providing a justification for a specific prediction, whereas artificial neural networks are mathematical models that do not provide such a transparent view of "how they do what they do." Recent trends in forecasting is leaning toward using a combination of forecasting techniques (as opposed to one that performed the best based on the test dataset) for more accurate and more robust outcome. That is, it is a good idea to use these three models together for predicting the freshmen students who are about to dropout, as they confirm and complement each other.

Successful student retention practices at the institutional level follow a multi-step process, which starts with determining, storing (in a database), and using student characteristics to identify the at-risk students who are more likely to dropout, and ends with developing effective and efficient intervention methods to retain them. In such a process, data mining can play the critical role of reasonably accurately predicting attrition as well as explaining the factors underlying the phenomenon. Because machine learning methods (such as artificial neural networks and decision trees used in data mining) are capable of modeling highly nonlinear relationships, they are more appropriate techniques to predict the complex nature of student attrition with a high level of accuracy.

The success of a data mining project relies heavily on the richness (quantity and quality) of the data representing the phenomenon under consideration. Even though this study used a large sample of data (covering 8 years of freshmen student records) with a rather rich set of features, more data and more variables can potentially help improve the data mining results. These variables, which are mentioned in recent literature as important, include:

1. student's social interaction (being a member of a fraternity or other social groups);
2. student's prior expectation from his educational endeavors; and
3. student's parent's educational and financial background.

Once the initial value of this quantitative analysis is realized by the institution, new and improved data collection mechanisms can be put in place to collect and potentially improve the analysis results.

As the sensitivity analysis and the decision tree structure indicate, the most important predictors for student attrition are those related to the past and present educational success of the student. In order to improve the retention rates, institutions may choose to enroll more academically successful students. Also, it might be of interest to monitor the academic experience of freshmen students in their first semester through looking at a combination of grade point average and the ratio of completed hours over enrolled hours.

The focus (and perhaps the limitation) of this study is the fact it aims to predict attrition using institutional data. Even though it leverages the findings of the previous theoretical studies, this study is not meant to develop a new theory, rather it is meant to show the viability of data mining methods as a means to provide an alternative way to understand and predict student attrition at higher educations. From the practicality standpoint, an information system encompassing these prediction models can be used as a decision aid to student enrollment management departments at higher-educations who are sensitive to student retention.

Potential future directions of this study include: (i) extending the predictive modeling methods to include more recent techniques such as support vector machines and Rough set analysis; (ii) enhancing the information sources by including the data from survey-based institutional studies (which are intentionally crafted and carefully administered for retention purposes) (in addition to the variables in the institutional databases); and (iii) deployment of the system as a decision aid for administrators to assess its suitability and usability in real-world.

REFERENCES

Astin, A. (1993). *What matters in college? Four critical years revisited*. San Francisco, CA: Jossey-Bass.

Berger, J. B., & Braxton, J. M. (1998). Revising Tinto's interactionalist theory of student departure through theory elaboration: Examining the role of organizational attributes in the persistence process. *Research in Higher Educat*ion, *39*(2), 103-119.

Berger, J. B., & Milem, J. F. (1999). The role of student involvement and perceptions of integration in a causal model of student persistence. *Research in Higher Education, 40*(6), 641-664.

Berry, M. J. A., & Linoff, G. S. (2004). *Data mining techniques for marketing, sales, and customer relationship management* (2nd ed.). New York: Wiley Computer Publishing.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Caison, A. L. (2007). Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. *Research in Higher Education, 48*(4), 435-449.

Cabrera, A. F., Nora, A., & Castaneda, M. A. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education, 64*(2), 123-139.

Deberard, S. M., Julka, G. I., & Deana, L. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal, 38*(1), 66-81.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining.* MIT Press, 1-34.

Gansemer-Topf, A. M. & Schuh, J. H. (2006). Institutional selectivity and institutional expenditures: Examining organizational factors that contribute to retention and graduation. *Research in Higher Education, 47*(6), 613-642.

Haykin, S. (1998). *Neural networks: A comprehensive foundation*. Englewood Cliffs, NJ: Prentice Hall.

Hermanowicz, J. C. (2003). *College attrition at American research universities: Comparitive case studies*. New York: Agathon Press.

Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in Higher Education, 46*(8), 883-928.

Hochstein, S. K., & Butler, R. R. (1983). The effects of the composition of a financial aids package on student retention. *Journal of Student Financial Aid, 13*(1), 21-27.

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feed-forward network. *Neural Networks, 3*, 359-366.

Ishitani, T. T. (2006). Studying attrition and degree completion behavior among first-generation college students in the United States. *The Journal of Higher Education, 77*(5), 861-885.

Kiang, M. Y. (2003). A comparative assessment of classification algorithms. *Decision Support Systems, 35,* 441-454.

Lau, L. K. (2003). Institutional factors affecting student retention. *Education, 124*(1), 126-137.

Law, R. (2000). Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management, 21*(4), 331-340.

Lim, T., Loh, W., & Shih, Y. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning, 40,* 203-229.

Mannan, M. A. (2007). Student attrition and academic and social integration: Application of Tinto's model at the university of Papua New Guinea. *Higher Education, 53*(2), 147-165.

Pascarella, E. T., Terenzini, P. T., & Wolfle, L. M. (1986). Orientation of college and freshman year persistence/withdrawal decisions. *Journal of Higher Education, 57*(2), 155-168.

Porter, K. B. (2008). Current trends in student retention: A literature review. *Teaching and Learning in Nursing, 3*(1), 3-15.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning, 1,* 81-106.

Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Mateo, CA.: Morgan Kaufmann.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications, 30*(2), 243-254.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5,* 13-22.

SPSS (2008). *SPSS Clementine User Manual.* A Comprehensive Data Mining Toolkit (version 12).

Stampen, J. O., & Cabrera, A. F. (1986). Exploring the effects of student aid on attrition. *Journal of Student Financial Aid, 16*(2), 28-37.

Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student opinion data with regression and decision tree analysis. *Research in Higher Education, 45*(3), 251-269.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition*. Chicago, IL: University of Chicago Press.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). Chicago, IL: The University of Chicago Press.

Tinto, V. (1997). Classroom as communities: Exploring the educational character of student persistence. *Journal of Higher Education, 68*(6), 599-623.

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting academic performance by data mining methods. *Education Economics, 15*(4), 405-419.

Wetzel, J. N., O'Toole, D., & Peterson, S. (1999). Factors affecting student retention probabilities: A case study. *Journal of Economics and Finance, 23*(1), 45-55.

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems, 11,* 545-557.

Direct reprint requests to:

Dursun Delen, Ph.D.
Oklahoma State University
William S. Spears School of Business
700 North Greenwood Ave.
Tulsa, OK 74106
e-mail: dursun.delen@okstate.edu