2

Focusing on student retention and time to degree completion, this study illustrates how institutional researchers may benefit from the power of predictive analyses associated with data-mining tools.

Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis-à-Vis Regression

Serge Herzog

Understanding student enrollment behavior is a central focus of institutional research in higher education. However, in the eyes of an enrollment management professional, the capacity to explain why students drop out, why they transfer out, or why some graduate quickly while others take their time may be less critical than the ability to accurately predict such events. Being able to identify who is at risk of dropping out or who is likely to take a long time to graduate helps target intervention programs to where they are needed most and offers ways to improve enrollment, graduation rate, and precision of tuition revenue forecasts.

Explanatory models by regression and path analysis have contributed substantially to our understanding of student retention (Adam and Gaither, 2005; Pascarella and Terenzini, 2005; Braxton, 2000), although the cumulative research on time to degree (TTD) completion is less impressive. A likely explanation for this is the more complex nature of the path to graduation, which has lengthened considerably over the past thirty years for a typical student (Knight, 2002, 2004; Noxel and Katunich, 1998; Council for Education Policy, Research and Improvement, 2002). Thus, whereas

The author thanks SPSS, Inc., for the demonstration version of their Clementine software used in this study.

NEW DIRECTIONS FOR INSTITUTIONAL RESEARCH, no. 131, Fall 2006 © Wiley Periodicals, Inc.
Published online in Wiley InterScience (www.interscience.wiley.com) • DOI: 10.1002/ir.185

prediction models for student retention have benefited greatly from more extensive analyses of factors associated with enrollment outcomes, factors influencing TTD completion are less understood. Therefore, developing models to estimate TTD completion is more difficult because research in this area has not matured to the same level.

Yet, comparing the prediction accuracies of these events (that is, retention and TTD completion) actually provides a useful framework to evaluate the relative potentials of different data analysis approaches. Namely, are there significant differences in prediction accuracy between data-mining tools and more traditional techniques when estimating outcomes of varying levels of complexity? Complexity in the data is typically associated with quality, quantity, and interaction of predictor variables and the number of possible outcomes in the dependent variable. To test for such differences, this study compares the prediction accuracy of three decision tree and three artificial neural network approaches with that of multinomial logistic regression. Findings are translated into operationally meaningful indicators in the context of enhanced institutional research on student retention, enrollment forecasting, and graduation rate analysis. Although the selection of predictor variables is guided by the research on retention and degree completion, the discussion focuses on which approach promises greatest prediction accuracy, not on how well either event is explained based on model fit and variable selection.

Prediction Accuracy of Data-Mining Techniques

Published studies on the use and prediction accuracy of data-mining approaches in institutional research are few. Luan (2002) illustrated the application of neural network and decision tree analysis in predicting the transfer of community college students to four-year institutions, concluding that a classification and regression tree (C&RT) algorithm yielded overall better accuracy than a neural network. Estimating the application behavior of potential freshmen who submitted admission test scores to a large research university, Byers González and DesJardins (2002) showed neural network accuracy to be superior over a binary logistic regression approach. A four-percentage point improvement in the overall correct classification rate with the neural network is extended by an additional two percentage points when continuous variables are used in lieu of dummy types. Maximizing accuracy in this way is possible with neural networks because they accommodate nonlinearity and missing values among variables. Van Nelson and Neff (1990) conducted a similar comparative study of two neural networks and a linear regression function that yielded comparable results.

Failure to produce better results with the neural network solutions may be due to the small sample (fewer than five hundred) employed because neural networks typically work best with larger data sets. Using both cognitive and noncognitive variables to predict algebra proficiency—to improve course placement or admission selection—Everson, Chance, and Lykins (1994) contrasted three variations of neural networks with both a linear regression model and discriminant function analysis. Relying on a relatively small sample of six hundred cases, the study employed a cross-validation strategy based on ten training and test subsets of randomly selected cases to infer a population measure for all approaches. The neural networks outperformed the two traditional techniques, achieving higher average coefficients of determination (R^2) and number of correctly classified cases. Although not focused on the issue of prediction accuracy per se, Thomas and Galambos (2004) demonstrated how decision tree analysis enriched the understanding of students' satisfaction with their college experience. Employing a chi-squared automatic interaction detector (CHAID) algorithm, they showed how displayed tree structures revealed patterns in the data that reflected heterogeneity among students that was elusive to regression analysis alone.

Examples outside the higher education domain further illuminate the potential utility of data-mining approaches. Concerned about the sustainability of runaway public school spending—with a 540 percent increase in constant-dollar-per-pupil expenditure from 1940 to 1990—Baker and Richards (1999) applied neural network solutions, using both standard and log-transformed data, to forecast future spending levels. Comparing the prediction accuracy of these models with one that replicated a multiple linear regression model used by the National Center for Education Statistics (NCES) for its annual projections, the study confirmed higher prediction accuracy with the neural networks. Surprisingly, those limited to linear predictors attained the highest accuracy, about twice the level of the NCES regression model, a finding the authors suggested was due to the relatively high linearity of nationally aggregated annual data employed by the study. Using thirty-five different data sets ranging from one thousand to two million records, Perlich, Provost, and Simonoff (2003) contrasted logistic regression with decision tree analysis to predict a variety of binary outcomes (such as customer type at a bookstore, contraceptive methods for women, bank credit approval, presence of diabetes, online shopping behavior, and so on). They found that tree induction yielded better classification results on larger data sets than on smaller ones.

The influence of data size on prediction performance was also reflected in a study by Long, Griffith, Selker, and D'Agostino (1993), which arrived at comparable accuracy levels for both the decision tree and logistic regression. The lack of superior performance with tree induction may have been due to the use of the older C4 algorithm. Estimating the level of aerobic fitness in adults, Tamminen, Laurinen, and Röning (1999) tested the accuracy of a C&RT versus neural networks and concluded that the latter provided better results. The advantage of using ensembles of resampled data for each prediction or classification in a decision tree (such as the C5.0), referred to as "bagging" and "boosting," was discussed in Opitz and Maclin (1999) and Roe and others (2005). Results from their studies showed that a boosting algorithm may improve prediction by 80 percent over the standard algorithm of a neural network, depending on the characteristics of the data set, the number of predictors, and the size of the tree structure. Boosting is a function of the greater weight attached to misclassified records in the course of sequentially built models from the resampled data, with the final prediction based on aggregate results from all the models.

This brief review of studies shows that data-mining methods offer distinct advantages over traditional statistics. Particularly when working with large data sets to estimate outcomes with many predictor variables, data-mining methods often yield greater prediction accuracy, classification accuracy, or both. However, higher education research provides little insight into which specific data-mining method to use when predicting key outcomes such as retention or degree completion.

Data Sources, Samples, and Variables

Based on students at a moderately selective Carnegie doctoral degree and research university, I accessed three sources to generate the data file: the institutional student information system for student demographic, academic, residential, and financial aid information; the American College Test (ACT)'s Student Profile Section for parent income data; and the National Student Clearinghouse for identifying transfer-out students. Retention predictions are based on the second-year enrollment of 8,018 new full-time freshmen who started in the fall semesters of 2000 through 2003 (or 96 percent of the total cohort, excluding varsity athletes, non-degree-seeking, and foreign students). Another data set used for the TTD-completion analysis captured end-of-fourth-year information of 15,457 undergraduate degree recipients from spring 1995 through summer 2005 (or 99 percent of all recipients after listwise deletion of incomplete records; the first degree received was counted for 85 multiple-degree holders). Forty predictors were used to estimate retention, and seventy-nine variables were included in the more complex TTD forecasts.

Variable selection for the retention predictions reflected an established institutional model to identify at-risk freshmen, as I described elsewhere (2005). Given the importance of first-year curricular experience in the established retention model, course experiences in core areas of the undergraduate experience guided the selection of many variables in the model development for TTD estimation. Variables are listed in Appendix A (at the end of this chapter) by cluster area and identified by type, with additional definitions furnished in Appendix B to ensure clarity of what is measured. Missing values for 493 cases were imputed for the variable measuring the ratio of earned-to-attempted credits by multiple regression (R^2 = 0.70); a general linear model was used to impute total campus-based credits for 804 cases (R^2 = 0.68). Mean value substitution was performed on 230 missing ACT scores.

Analytical Approach

The comparison of prediction accuracy is based on contrasting three-rule induction decision trees (C&RT, CHAID-based, and C5.0) and three backpropagation neural networks (simple topology, multitopology, and three-hidden-layer pruned) with a multinomial logistic regression model. Statistical Package for the Social Sciences, Inc.'s Clementine software version 9.0 was employed, and a brief description of each technique in Appendix C reflects algorithms used by the application and run methods and modeling options available. Detailed accounts of decision tree and neural network theory and application can be found in Murthy (1998), Berry and Linoff (2000), and Garson (1998).

Second-year retention was examined at the end of both the first and the second terms. For practical reasons and for the varying data complexity, the TTD analysis was conducted with and without transfer students to account for their typically distinct path to graduation. To validate prediction accuracy of each method used (that is, the statistical algorithm used), data sets were partitioned through a 50:50 randomized split to allow generated models based on training data to be tested on the holdout samples. The following retention outcomes are estimated: returned for second year (fall to fall), transferred out within one year, and dropped or stopped out for at least one year; degree completion time (TTD) outcomes are estimated for three years or less, four years, five years, and six years or more. This categorization generates a more balanced outcome in the dependent variable and ensures convergence in the regression model.

The profile of the cohort used for the retention analysis is summarized as follows: 56 percent were female students, 7 percent were Asian American, 3 percent were African American, 7 percent were Hispanic, and a little more than 1 percent were Native American. Fifty-six percent resided within commuting distance, and 11 percent were out-of-state residents. Twenty-six percent came from a low-income background, and 18 percent were high-income students. Thirteen percent entered with no financial aid, 19 percent took out student loans, 13 percent received Pell grants, and 68 percent used only forms of gift aid (scholarships and grants). Twenty percent took summer courses before their initial fall enrollment, and 14 percent entered with advanced placement credits. On average, students entered with an ACT score of 23.

Of the graduated students used for the TTD analysis, 57 percent were female, 8 percent were minority students (excluding Asian Americans), 70 percent resided within commuting distance of the campus, and 21 percent were out-of-state students. The average age at graduation was 27 years, and 20 percent were continuously enrolled (that is, never stopped out), 40 percent entered as transfer students (68 percent of whom came from in-state institutions), 16 percent started with an undeclared program major, and 19 percent graduated with honor (with any type of distinction).

Before training the models for a given outcome estimation, data miners typically explore the data to identify patterns among variables for guidance during the model-building process. For example, would merit-based student aid go to fast degree completers with strong academic records? As Figure 2.1 shows, more merit aid is indeed allocated to faster completers with higher grades. Less obvious is the impact of taking courses taught by adjunct faculty. As Figure 2.2 reveals, students finishing up within three years or less are more likely exposed to adjunct instructors. Because estimates of the TTD completion are based on a large number of predictors, whose interaction is not well understood, caution should be exercised in interpreting the contribution of a given variable in the model.

Prediction Accuracy of Tested Models

A comparison of the tested models' percentage of correctly predicted cases across all outcomes combined, based on the validation data sets, is depicted in Figure 2.3 for retention and Figures 2.4 and 2.5 for TTD completion. The overall prediction accuracy level with the validation data sets was within a few percentage points of the training sets, except for the C5.0 models, which

Figure 2.1. Merit Aid, Grades, and Degree Completion Time

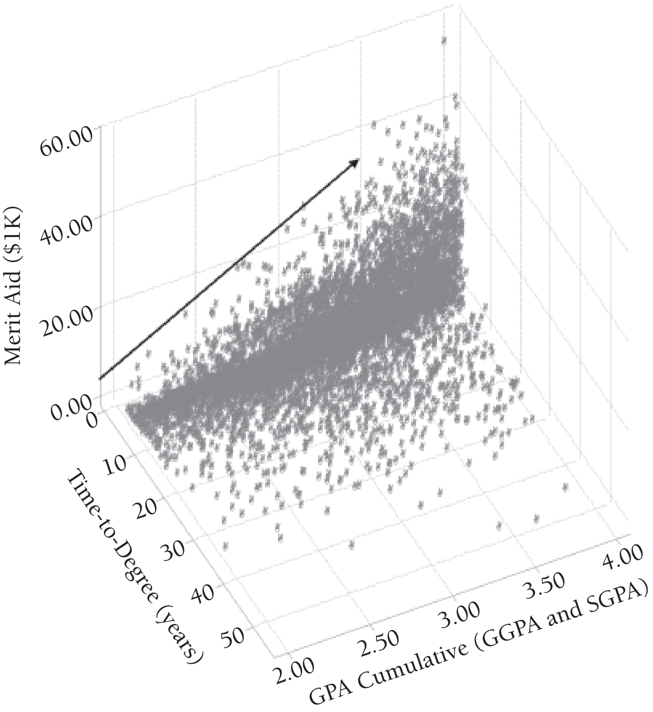
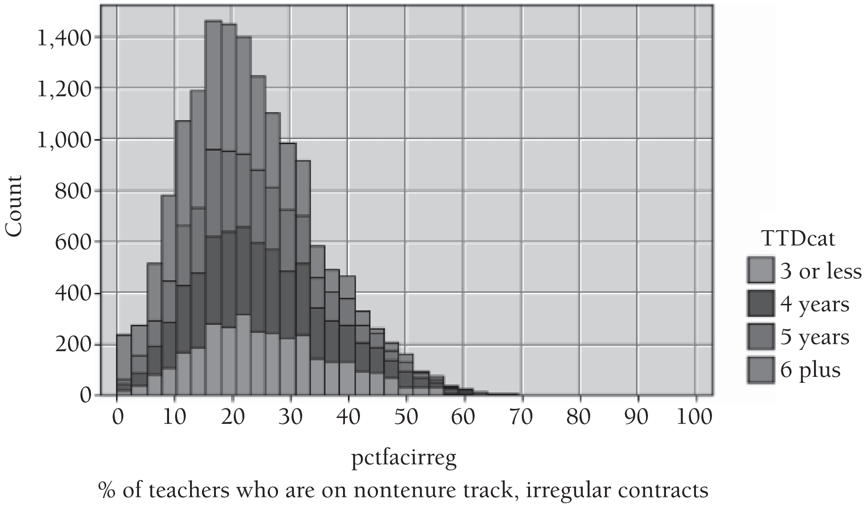


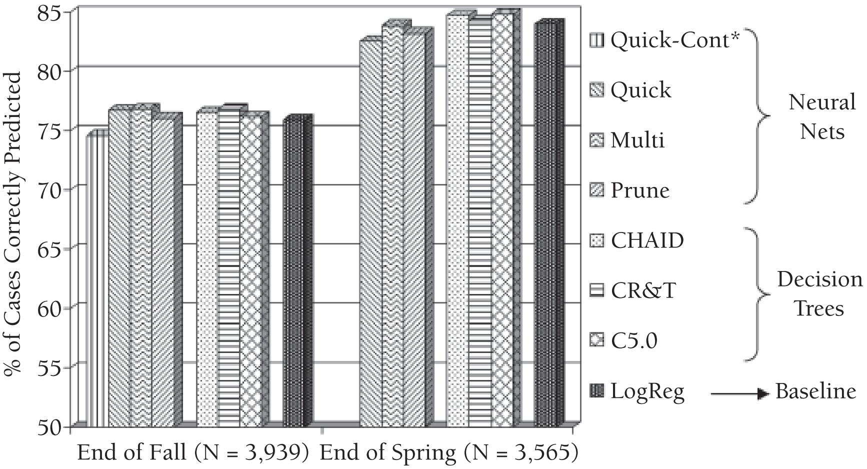
Figure 2.2. Exposure to Adjunct Faculty and Time to Degree (TTD)



Note: The horizontal axis represents the percentage of teachers who are on nontenure track or have irregular contracts.

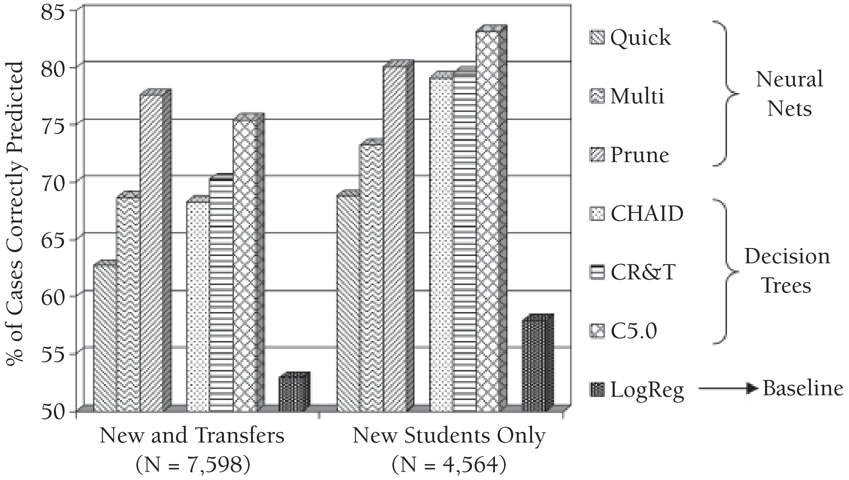
on average dropped by about fifteen percentage points. Although some of the tree and neural net algorithms yielded slightly higher levels of retention prediction over the baseline regression model, that advantage did not appear compelling with either the end-of-fall or end-of-spring data. Accuracy did

Figure 2.3. Model Comparison for Freshmen Retention: Overall Prediction Accuracy with Validation Data



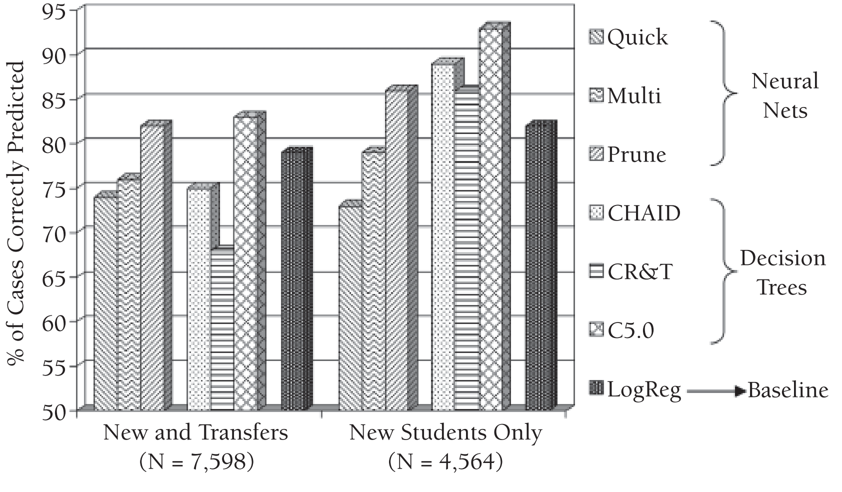
Note: *Continuous variables.

Figure 2.4. Model Comparison for Degree Completion Time: Prediction Accuracy with Validation Data, Three Years or Less



improve by almost ten percentage points as richer information became available at the end of the spring over the fall, but that improvement accrued with the baseline model also. The multitopology neural network performed significantly better in identifying dropout-stop-out students but poorly vis-à-vis the regression model in estimating who is likely to transfer out. The absence of more impressive results with the decision trees and neural net-

Figure 2.5. Model Comparison for Degree Completion Time: Prediction Accuracy with Validation Data, Six Years or More



works was likely due to, among other things, the selection of variables exhibiting little collinearity and interaction effects during the development of the baseline regression model.

Overall prediction results in the TTD models, both with and without the inclusion of transfer students, showed substantial differences. When new and transfer students were examined together, the neural network with three hidden layers achieved a twenty-five-percentage point improvement in correctly predicted cases over the regression model, an overall accuracy of almost 50 percent greater. A similar improvement was achieved with the three decision trees and pruned neural network when the analysis was limited to new students. Further data reduction by excluding transfer students led to more improvement in accuracy for all models (see Figure 2.5). However, the "overall accuracy" discussed herein may not be as meaningful to the institution if its operational interest is focused on identifying those who are predicted to be taking longer to graduate. Hence, in this context, attention is placed on correctly identifying those expected to take six years or more, given that the six-year graduation rate has become a national benchmark for moderately selective institutions. The advantage of using any of the tested decision trees and the pruned neural network is even more pronounced when transfer students were excluded from the data set, with the best model, the C5.0, yielding an accuracy of 93 percent, or eleven percentage points higher than the baseline model. Thus, the potential advantage gained with any one model must be weighed in terms of which outcome is most effective in assisting the development of a successful student intervention program.

The decision trees and neural networks performed especially well in correctly predicting the relatively few students who graduated within three years or less when the data set was restricted to new students; conversely, the regression model performed comparably on the six-years-or-more outcome, given the level of prediction difficulty associated with that outcome. Hence, performance evaluation scores were almost identical for all tested models when predicting students who take six years or more but were notably lower for the regression model when estimating who graduated quickly. A confidence level report, available through the analysis node, further assisted the analyst in interpreting the prediction accuracy for *each* record in the data set. Mean confidence levels for all records confirmed better results for the decision trees and the regression model than for the neural networks, except in predicting TTD for all students, where the pruned neural network scored the highest. Comparing threshold scores for a specified level of accuracy—for example, the level achieved with the baseline approach—may further illuminate which technique is most preferred.

Output from the sensitivity analysis produced by the neural networks showed which predictors weighed in the most when estimating retention and degree completion time. In the latter case, neural networks identified credit hour–related predictors, student age, residency, and stop-out time among the most valuable. Similarly, credit hours, residency, and stop-out time had the largest beta coefficients in the regression model. However, unlike the baseline model, the neural networks also were able to use a student's program major, the trend in grades over time (defined in Appendix B), and exposure to adjunct faculty as effective correlates of TTD. In turn, the regression approach benefited more from knowing a student's English experience (including English 102 grade and English transfer credits) in the estimation process. A look at the hierarchy of generated rules in the decision trees showed that a student's age, average credit load, and grade-related predictors result in early splits in the induction process. These results confirm that enrollment intensity—as it relates to credit load per term, stop-out behavior, and dropped courses—is a key indicator in estimating graduation time. But as the neural network findings suggested, the first program major selected, the number of times a student changes his or her major, grades over time (the GPA trend), and even the type of faculty by whom the students are taught all contribute to a more accurate prediction.

Implications and Conclusions

Having examined the prediction accuracy of several data-mining methods, all relatively new to institutional research, and compared it with that of logistic regression, a well-established approach, the study found that the level of complexity of the data used and the outcome predicted may largely guide the selection of a particular analytical tool. Predicting freshmen retention based on inferential statistics with little collinearity among variables may not be the best use of decision tree or neural net methods. On average, the decision trees and neural networks performed at least as good as the regression model, but the number and type of variables used as predictors—and repeatedly tested and adjusted in previous regression models—did not confer a substantial advantage to the data-mining methods.

However, the data-mining algorithms worked notably better with the larger set of exploratory predictors used to estimate the degree completion time, an outcome not as well studied as student retention. In the most complex analysis, where time to graduation is estimated for new and transfer students simultaneously, the pruned neural network with three hidden layers and the C5.0 decision tree performed best. The different results from the three neural networks confirm the importance of exploring available setup options in the application's neural network node. For example, adding multiple hidden layers and removing underused neurons (pruning) greatly raised the accuracy level compared with the less impressive results with the simple-topology model. Similarly, applying an error-weighted boosting factor to misclassifications in the C5.0 model may yield further improvements, as demonstrated in other studies discussed in the literature review section.

Decision tree and neural network approaches may expand the analyst's understanding of what variables contribute to prediction accuracy when working with large data sets. Sensitivity analysis helped identify important predictors beyond those highlighted in the regression model. Intuitively, one may expect credit load and stop-out time to be related to degree completion time. But what about the type of teaching faculty? That variable, and other less obvious ones, attained fairly high sensitivity scores in the neural network results; in contrast, the baseline regression model showed little statistical significance associated with the type of faculty. Thus, getting the most out of working with large sets of predictor variables is easier with data-mining tools that rank predictors on a single metric. Sensitivity scores from this study are instrumental in guiding the development of simplified models and may eliminate variables over which the institution has little control. For example, among the many financial aid measures used in the TTD analysis, only three scored relatively high on the sensitivity scale, namely, whether a student received a state-funded merit scholarship, the total amount of loans received, and the total amount of aid received from any source. Identifying all sources of aid separately does not appear to contribute much to prediction accuracy. Similarly, knowing whether a student took a remedial math course added little in the presence of indicators measuring math performance (grades) across all introductory math classes.

The operational impact of this study on the institution centers on how many additional students who are "at risk" could be correctly identified in due course to allow for effective intervention. Because data-mining techniques scarcely differed from the regression model in overall prediction accuracy and produced inconsistent results in estimating dropout or stop-out versus transfer-out risk in the retention analysis—except, of course, that data-mining application allows automation in scoring future data, which cannot be matched by conventional statistical tools. Compared with the baseline approach, improved prediction accuracy with the multilayer pruned neural network and the C5.0 would yield an additional 105 to 140 correctly identified students at risk of taking at least six years or more to graduate (or 3 to 4 percent of approximately 3,500 undergraduates enrolled after four years). If the prediction is restricted to only students who started as new freshmen, the advantage would range from 170 to 270 (or 7 to 11 percent of approximately 2,450 students).

Because the institution is likely to experience continued strong enrollment growth over the next ten years, the number of additional students who can be identified as at risk will grow commensurately. The benefit of correctly targeting hundreds of additional students each year with appropriate counseling and academic support is difficult to gauge because potential benefits are a function of the effectiveness of the program(s) put in place to promote faster degree completion. But an intervention can proceed with more confidence on knowing that resources are allocated to the right students.

Improvement in the degree completion time by better identification of at-risk students translates not only to higher graduation rates but also to substantial cost savings for students. By way of illustration, timely counseling prompted by a newly identified at-risk student may well obviate the occurrence of a change in program major, which is associated with lengthening time to graduation. Barrow and Rouse (2005) put the total cost of a four-year degree at over $107,000 for the average student who entered college in 2003 and who will graduate in four years. About 72 percent of that total cost is due to the opportunity cost of lost wages (which equal average annual earnings of a high school graduate). Adding the difference in lifetime earnings over forty years between a college graduate and a high school graduate, and assuming an average six-year degree completion time, which approximates conditions at the institution, speeding up time to graduation by one year may save a student around $28,000 in forgone earnings (not counting a likely higher increment of tuition and fees for a six-year graduate compared with a five-year graduate). Faster completion reduces both the net attendance cost and the time to recoup the educational investment cost once gainfully employed. Finally, better estimates of how long students take to graduate may help improve enrollment projections that are used for institutional planning purposes (for example, construction of new classroom buildings).

Current projections are in part based on how quickly students move from one class standing to the next (such as from freshman to sophomore), as expressed in the conversion ratio (that is, the percentage of sophomores in year 2 who were freshmen in year 1). Projected numbers are also disaggregated by demographic attributes, such as student age, gender, or ethnicity. Running data mining–based estimates of degree completion time with data similarly disaggregated may help improve enrollment projections through adjusted conversion ratios in the model. For example, a drop from one year to the next in the estimated proportion of students taking six years or more to graduate may well suggest a need to lower the conversion ratio to project the number of returning seniors in the future (that is, a greater proportion of seniors in the previous year will graduate). Being able to more accurately predict how many students graduate within a given time frame through demographic and academic attributes will help refine future enrollment projections. Such projections may also benefit from information associated with variables that are useful to data mining–based estimates but cannot be used under the more restrictive statistical assumptions that govern regression analyses.

References

Adam, J., and Gaither, G. H. "Retention in Higher Education: A Selective Resource Guide." In G. H. Gaither (ed.), *Minority Retention: What Works?* New Directions for Institutional Research, no. 125. San Francisco: Jossey-Bass, 2005.

Baker, B., and Richards, C. "A Comparison of Conventional Linear Regression Methods and Neural Networks for Forecasting Educational Spending." *Economics of Education Review,* 1999, *18,* 405–415.

Barrow, L., and Rouse, C. "Does College Still Pay?" *Economists' Voice,* 2005, 2(4), 1–8.

Berry, M., and Linoff, G. *Master Data Mining: The Art and Science of Customer Relationship Management.* New York: Wiley Computer, 2000.

Braxton, J. *Reworking the Student Departure Puzzle.* Nashville, Tenn.: Vanderbilt University Press, 2000.

Byers González, J., and DesJardins, S. “Artificial Neural Networks: A New Approach for Predicting Application Behavior.” *Research in Higher Education,* 2002, *43*(2), 235–258.

Council for Education Policy, Research and Improvement [CEPRI]. “Postsecondary Progression of 1993–94 Florida Public High School Graduates: 2002 Update.” *To the Point,* August 2002. http://www.cepri.state.fl.us. Accessed Nov. 18, 2005.

Everson, H., Chance, D., and Lykins, S. “Using Artificial Neural Networks in Educational Research: Some Comparisons with Linear Statistical Models.” Paper presented at the American Educational Research Association Conference, New Orleans, April 5–7, 1994. (ED 372 094)

Garson, G. *Neural Networks: An Introductory Guide for Social Scientists.* London: Sage, 1998.

Herzog, S. “Measuring Determinants of Student Return vs. Dropout/Stopout vs. Transfer: A First-to-Second Year Analysis of New Freshmen.” *Research in Higher Education,* 2005, *46*(8), 883–928.

Knight, W. “Toward a Comprehensive Model of Influences upon Time to Bachelor’s Degree Attainment.” *AIR Professional File,* 2002, *85*(Winter).

Knight, W. “Time to Bachelor’s Degree Attainment: An Application of Descriptive, Bivariate, and Multiple Regression Techniques.” *IR Applications,* 2, Sept. 8, 2004. http://airweb.org/page/asp?page=628. Accessed Oct. 15, 2004.

Long, W., Griffith, J., Selker, H., and D’Agostino, R. “A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain.” *Computers in Biomedical Research,* 1993, *26,* 74–97.

Luan, J. “Data Mining and Its Applications in Higher Education.” In A. M. Serban and J. Luan (eds.), *Knowledge Management: Building a Competitive Advantage in Higher Education.* New Directions for Institutional Research, no. 113. San Francisco: Jossey-Bass, 2002.

Murthy, S. “Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey.” *Data Mining and Knowledge Discovery,* 1998, 2, 345–389.

Noxel, S., and Katunich, L. “Navigating for Four Years to the Baccalaureate Degree.” Paper presented at the Association for Institutional Research Conference, Minneapolis, May 17–20, 1998. (ED 422 825)

Opitz, D., and Maclin, R. “Popular Ensemble Methods: An Empirical Study.” *Journal of Artificial Intelligence Research,* 1999, *11,* 169–198.

Pascarella, E., and Terenzini, P. *How College Affects Students.* San Francisco: Jossey-Bass, 2005.

Perlich, C., Provost, F., and Simonoff, J. “Tree Induction vs. Logistic Regression: A Learning-Curve Analysis.” *Journal of Machine Learning Research,* 2003, *4,* 211–255.

Roe, B., and others. “Boosted Decision Trees as an Alternative to Artificial Neural Networks for Particle Identification.” *Nuclear Instruments and Methods in Physics Research,* 2005, *543,* 577–584.

Tamminen, S., Laurinen, P., and Röning, J. “Comparing Regression Trees with Neural Networks in Aerobic Fitness Approximation.” Proceedings of the International Computing Sciences Conference Symposium on Advances in Intelligent Data Analysis, Rochester, N.Y., June 22–25, 1999, pp. 414–419.

Thomas, E., and Galambos, N. “What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analysis.” *Research in Higher Education,* 2004, *45*(3), 251–269.

Van Nelson, C., and Neff, K. “Comparing and Contrasting Neural Network Solutions to Classical Statistical Solutions.” Paper presented at the Midwestern Educational Research Association Conference, Chicago, Oct. 19, 1990. (ED 326 577).

SERGE HERZOG is director of institutional analysis at the University of Nevada, Reno.

Appendix A: Predictors

Retention Analysis

Student Demographics
Gender[a]
Age[b,c]
Ethnicity or race[d]
Residency[d]
Parent income[d]

Precollegiate Experience
High school grade-point average (GPA)[b]
American College Test (ACT) English or math scores[b] (Scholastic Assessment Test [SAT] conversion)
ACT or SAT test date[b,c]
Academic preparation index[b,c]
Prefall summer enrollment[a]
Advanced placement or international baccalaureate (IB) credits[a]
Graduate degree aspiration[a]

Campus Experience
On-campus living[a]
Use of athletic facilities[a]
Dual enrollment with community college (CC)[a]
Fall entry term[d]
Attempted registrations[c]
Average class size[c]

Academic Experience
Academic peer challenge[c]
Fall or first-year GPA[b,c]
Credit load (<15 units)[a]
Major requires calculus 1[a]
Natural or physical science courses[c]
Remedial math taken[a]
Remedial English taken[a]
Math credits earned[c]
All and math transfer credits[a]
Fall or spring math grades[d]
Math D or F grades[a]
Math I or W grades[a]
Passed first-year math[a]
English 101 or 102 grades[d]
Program major type[d]

Financial Aid Need[b]
Fall and spring remaining
First-year total remaining
Fall and spring total need before total aid offered
Fall or spring package by type of aid included[d]
- No aid
- Package with loans, work study, or both
- Grants, scholarships only, or both
- Millennium scholarship only

Second-year package offer by type of aid included as above[d]
Fall or Spring Institutional Aid Amount ($)[b]
Second-Year Institutional Aid Amount Offered ($)[b]
Fall or Spring Pell Grant Aid[a]
Millennium Aid, Fall or Spring[d]
- Never had it
- Received it, maintains eligibility
- Lost eligibility, continues ineligibility
- Lost eligibility, regains eligibility

Time-to-Degree Analysis

Financial Aid
Total aid received[b]
Loans[b]
Grants[b]
Work study[b]
Merit-based aid[b]
Need-based aid[b]
General fund aid[b]

[a] Range or scale.
[b] Flag or binomial.
[c] Ordinal or rank.
[d] Set or multinomial.

Outside aid[b]
University of Nevada, Reno (UNR) Foundation aid[b]
Academic department-based aid[b]
Grants-in-aid[b]
Millennium Scholarship[b]
Pell Grant aid[b]

Student Demographics
Gender[a]
Age[b]
Ethnicity or race[d]
Residency[d]

Precollegiate Experience
ACT English[b]
ACT math[b]
ACT composite[b]

General Experience
Initial status (new versus transfer)[a]
Initial program major[d]
Program major in fourth year[d]
Number of program major changes[b]
Declared a minor[a]
Completed a senior thesis[a]
Attempted registrations[b]
Participated in varsity sports[a]
Stop-out time since first enrollment (percent)[b,c]

Course Grades
Remedial math[d]
College algebra[d]
College general math[d]
College trigonometry[d]
Introduction to statistics[d]
Business calculus[d]
Calculus 1[d]
English 101[d]
English 102[d]
Core humanities 201–203[d]
General capstone[d]
Program major capstone[d]
Cumulative GPA[b,c]
GPA trend[b]
Number of D or F grades (percent)[b]
Number of I or W grades (percent)[b]
Number of replacement grades[b]

Outside Course Experiences
Took overseas courses (USAC)[a]
Took continuing education courses[a]
Took courses at TMCC (local CC)[b]
Took courses at WNCC (local CC)[b]
Took internships[c]
English courses transferred in[c]
Math courses transferred in[c]
English distance courses[a]
Math distance courses[a]
Core humanities transferred in[c]

Campus Course Experiences
Took honors courses[a]
Took independent studies[c]
Repeated a course[a]
Took remedial math or English[a]
Capstone courses taken[b]
"Diversity" courses taken[b]
Natural science courses in three core areas (three variables)[b]

Credit Hours
Total credits accumulated[b]
Total transfer credits[b]
Total campus credits[b]
Total math credits[b]
Total upper-division science credits[b]
Total credits transferred in[b]
Earned:attempted credits (percent)[b]
Average credit load per enrolled term[b]

Faculty Teaching Courses Taken
Percentage of women[b]
Percentage of ethnic or racial minority[b]
Percentage of part-time faculty[b]
Percentage of adjunct faculty[b]
Percentage at full-professor rank[b]
Average age of faculty[b]

Appendix B: Variable Definitions

Initial study major Declared or premajor, undeclared, nondegree seeking, intensive English.

Attempted registrations Registration attempt at time of fully subscribed class section during registration period.

Stop-out time Number of fall or spring semesters not in attendance after first campus-based course enrollment.

Number of replacement grades Students may repeat up to twelve lower-division credits to replace original grades at institution.

Grade-point average (GPA) trend Ratio of twenty-four-credit GPA to cumulative GPA after fourth year.

Natural sciences core course offerings geared for three groups of majors Social science, natural science, and engineering.

Parent income Grouped into upper, middle, and bottom thirds with a "missing" category for students without a federal aid application and without data from the American College Test (ACT) Student Profile Section.

ACT English and math scores A Scholastic Assessment Test (SAT)-ACT conversion table was used for students with only SAT scores.

Appendix C: Model Descriptions and Generated Characteristics for Time-to-Degree (TTD) Models

CHAID Chi-squared automatic interaction detector; uses chi-squared statistics to identify optimal splits with two or more subgroups. Starts with most significant predictor, determines multiple-group differences, and collapses groups with no significance; the merging process stops at the preset testing level. Generated tree depth: 4.

C&RT Classification and regression tree; generates splits based on maximizing orthogonality between subgroups (measured by the Gini index of diversity); all splits are binary, and outcome variable can be continuous or categorical. Generated tree depth: 5.

C5.0 Uses the 5.0 algorithm to generate a decision tree or rule set based on the predictor that provides maximum information gain. The split process continues until the sample is exhausted. Lowest-level splits are removed if they fail to contribute to model significance. Generated tree depth: 4, no boosting; rules for each outcome: 38, 23, 40, 53 (default outcome: four years).

Neural net simple topology (quick method) Using the software default settings for learning rates (alpha and eta decay), this is a one-hidden-layer model with 158 nodes at the input, 8 in the hidden layer, and 4 at the output.

Neural net multiple topologies Creates several networks in parallel based on a specified number of hidden layers and nodes in each layer. Used default learning rates, with 158 input nodes, 5 nodes in the hidden layer and 4 at the output for the final solution; initial parallel networks with 2 to 20 nodes in one hidden layer and 2 to 27 and 2 to 22 nodes in two hidden layers, respectively.

Neural net prune method Starts with a large network of layers and nodes as specified and removes (prunes) weakest nodes in input and hidden layers during training. Three hidden layers were specified, with the following number of nodes from input to output: 38, 4, 2, 2, and 4.

Logistic regression Direct variable entry with main effects output only, resulting in a pseudo R^2 of 0.728 (Cox and Snell); likelihood convergence set on default.