A Comparison of Empirical Models for Predicting Student Retention

Matt Bogard · Tuesdi Helbig · Gina Huff ·Chris James

Abstract Literature indicates that data mining or algorithmic approaches to prediction can provide superior results vis-à-vis traditional statistical modeling approaches (Delen et al, 2004; Sharda and Delen, 2006; Delen et al, 2007; Kiang 2007; Li et al 2009). However, little research in higher education has focused on the employment of data mining methods for predicting retention. Using three years of data for first time first year degree seeking students we compared results from logistic regression, decision trees, neural networks, and ensemble models implemented using SAS Enterprise Miner. Based on validation misclassification rates and robustness, decision trees were chosen for the final model specification. Building models for three time periods (pre-enrollment or 1st day of term, 5th week of term, and end of term) the predictive accuracy of our decision trees improved with each shift in time. The score code generated from SAS Enterprise Miner will make it possible to create custom reports with color coded risk indicators within the SAS BI decision support system. This will enable administrators, advisors, professors and other higher education professionals within the institution to incorporate advanced analytics into retention efforts.

Keywords Student retention, Data mining, Decision trees, SAS Enterprise Miner

Introduction

This paper consists of four major sections followed by a bibliography. Section one consists of a review of the literature related to empirical methods utilized in modeling student retention. Section two discusses data and variables considered in our study. In section three we discuss the model specifications and preliminary results. Section four considers the possibility of scoring new students and an enterprise wide deployment of the model results. The paper closes with a discussion of further research and possible model improvements.

Modeling Student Retention: A Summary of Empirical Methods from Literature

The vast majority of the literature related to the empirical estimation of retention models includes a discussion of the theoretical retention framework established by Bean, Braxton, Tinto, Pascarella, Terenzini and others (see Bean, 1980; Bean, 2000; Braxton, 2000; Braxton et al, 2004; Chapman and Pascarella, 1983; Pascarell and Ternzini, 1978; St. John and Cabrera, 2000; Tinto, 1975) This body of research provides a starting point for the consideration of which explanatory variables to include in any model specification, as well as identifying possible data sources. The literature separates itself into two major camps including research related to the hypothesis testing and the confirmation or empirical validation of theoretical retention models (Herzog, 2005; Ronco and Cahill, 2006; Stratton et al 2008) vs. research specifically focused on the development of applied predictive models (Miller, 2007; Miller & Herreid, 2008; Herzog, 2006; Dey & Astin, 1993; Delen 2010; Yu et al, 2010). Other areas of research seem to stand apart. While not particularly concerned with making accurate predictions or confirming or challenging the established literature, these researchers seek novel ways to measure student characteristics that may be theoretically important to retention, or provide predictive value. For instance, De Witz, Woosley, and Walsh (2009) investigate the relationship between Frankl's construct of purpose in life and Bandura's theory of self efficacy and the possible impact of these measures on student retention. They claim:

> **Many of the reasons that students leave college are outside Tinto's model: finances, poor academic performance, lack of family or social/ emotional encouragement, difficult personal adjustment. (De Witz, Woosley, and Walsh,2009)**

Their idea was that measures of self efficacy and purpose may be one way to capture this information. Others look at opportunities presented by social network analysis (SNA) (Thomas, 2000; Skahill, 2002; Brewe et al, 2009) According to the International Network for Social Network Analysis,*"social network analysis is focused on uncovering the patterning of people's interaction"* (http://www.insna.org/sna/what.html). Thomas integrates network measures of connectedness and centrality into a path analytic model of student retention (Thomas,2000). Skahil found that network metrics related to connectedness could explain differences in retention rates between commuter and residential students (Skahil, 2002). Brewe et al used SNA to characterize community interactions in terms of network density and connectivity and the assessed the impact of those metrics on retention and persistence for physics majors (Brewe, et al, 2009).

Within the context of work that related to theoretical validation and empirical modeling, some interesting findings merit discussion. Herzog found that the driving factors related to the propensity to retain involved institutional support and financial aid. Particularly, middle-income students were disproportionately impacted by the magnitude of unmet financial need (Herzog, 2005). Ronco and Cahil

looked at instructor types (full time faculty vs. graduate assistant vs. adjunct part time faculty) and found that the impact of instructor type on retention was not statistically significant (Ronco and Cahill, 2006). Stratton et al find that the type of financial aid received has a differential impact on dropout vs. stopout behavior, and caution that failure to distinguish between the risks of stopout and dropout students in predictive modeling could lead to misguided targeted interventions (Stratton, et al, 2008). As a consequence of the fact that the vast majority of researchers based initial model specifications and variable selection on the common body of research previously mentioned, most included variables related to pre-enrollment characteristics, demographics, socioeconomic status, and enrollment characteristics. The number of variables included in the models typically ranged from 10-30 or more. While the studies were redundant in what effects they were attempting to capture, some researchers presented novel ways of measuring these effects. Herzog (2005) presents two such interesting constructs. He utilizes a 'high school preparation index' influenced by Adelman's (1999) 'Academic Resources' composite variable as well as well as a 'peer challenge' variable that "*groups students into three approximately equal-size categories based on the difference between their first-semester GPA and the average grade awarded in classes attended. A weak challenge indicates a student on average received higher grades than his/her classmates, the opposite being the case for a strong challenge."* While the variables chosen for empirical modeling of retention outcomes were common among most of the researchers, with the exception of the few novel innovations previously mentioned, there were some distinguishing characteristics in relation to functional form. Many of the researchers utilized some form of logistic regression to estimate their models (Herzog, 2005; Miller, 2007; Miller and Herried,2008; Ronco and Cahill, 2006; Stratton et al 2008). Within the context of logistic regression, Stratton included the specification of a random utility model (Stratton et al, 2008). Astin and Dey (1995) examined discriminant analysis, linear, logistic, and probit models. Admitting violations of classical regression assumptions (particularly randomly distributed error terms and homoskedasticity of error terms) they found little practical difference between these methods in terms of co-efficient estimates, standard errors, and predicted probabilities (Astin and Dey, 1995). This has also been corroborated by Angrist and Pischke in their work comparing probit models with ordinary least squares:

> **While a nonlinear model may fit the CEF (population conditional expectation function) for LDVs (limited dependent variables) more closely than a linear model, when it comes to marginal effects, this probably matters little (Angrist and Pischke, 2008).**

When looking outside of the literature published in journals primarily focused on education (such as *College and University, Economics of Education Review, Research in Higher Education*) you will find a sharp contrast in methodology. These differences are palpably described by Leo Breiman:

> **There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.(Breiman, 2001)**

Breiman goes on to distinguish between these methods. Classical stochastic methods, or the 'data modeling' paradigm includes techniques such as linear regression, logistic regression, and analysis of variance. The 'algorithmic' or 'data mining' paradigm includes methods such as neural networks and decision trees. Another distinguishing characteristic between the two 'cultures' includes the concern with predictive accuracy. The ability to make accurate predictions across multiple data sets is described as the generalization performance of a model (Hastie, et al, 2009). Researchers engaged in algorithmic approaches look beyond the sample at hand to validate model results (Yu, et al 2010). Generalization

error is a function of the bias variance tradeoff related to model complexity and generalization performance across multiple data sets (Hastie, et al, 2009). None of the previously mentioned authors that utilized of logistic regression models addressed these issues. As suggested by Hastie et al, model selection techniques and partitioning the data into training, validation, and test subsets are possible strategies for addressing generalization error (Hastie, et al, 2009). Other approaches include the use of ensemble models. *The generalization performance of an ensemble of models (which is a collection or combination of predictive models) is typically improved over that of a single predictor (Krogh et al, 1997).*

As a result of this difference in cultures or modeling paradigms, research in higher education related to predicting attrition may be improved if algorithmic approaches are considered. As Breiman notes:

> **Approaching problems by looking for a data model imposes an apriori straight jacket that restricts the ability of statisticians to deal with a wide range of statistical problems (Breiman, 2001).**

Literature indicates that data mining or algorithmic approaches to prediction can provide superior results vis-à-vis traditional statistical modeling approaches (Delen et al, 2004; Sharda and Delen, 2006; Delen et al, 2007; Kiang 2007; Li et al 2009). However, little research in higher education has focused on the employment of data mining methods for predicting retention (Herzog, 2006). In a comparison of logistic regression, decision trees, and neural networks, Herzog finds that data mining algorithms worked better when dealing with larger sets of variables associated with degree completion (Herzog,2006). When oversampling the population of non-retaining students to create a balanced data set, Delen found that machine learning algorithms outperformed logistic regression and ensemble models outperformed individual models in predicting retention outcomes. Specifically the order, from most predictive to least predictive specification, was 1-support vector machines 2- decision trees, 3-neural networks 4-logistic regression (Delen,2010). Yu provides additional examples of the implementation of decision trees, neural networks, and multivariate-adaptive-regression-splines (MARS) in predicting retention (Yu et al, 2010).

Data and Variables

Three models were developed using data available upon the first day of the 1st term (pre-enrollment), the 5th week of enrollment, and full semester enrollment. The full semester model included all of the variables in the previous models, including some variables that were not available until the end of the first term. Similarly, the 5th week model included variables that were in the pre-enrollment model, in addition to data not available until the 5th week.

Pre-Enrollment Variables	Description
ACT_MATH_ENG	Average of ACT Math and English
AGE	Age
APPLICATION_TIME	# weeks applied before start of term
ATP_CODE	Indicator based on date of student summer orientation date
AVG_UNEMPLOYMENT_07_09	Student's home county unemployment rate
CONDITIONAL_ADMIT	Students admission status
DW_SEX	Gender
DEV_NEED_SCORE	Measure of # of Developmental Course Needs
EMP_07_09	% of students from home county that choose to work vs. attend

	college
FIRST_GENERATION	Y/N first generation student indicator
HS_GPA	High school GPA
HS_RETENTION	High school's historical retention rate
PCC	Pre-College Curriculum Indicator
PCT_CNTY_COLLEGE	% of students from home county that attend college
PUBLIC_CODE	Indicator for type of high school
REP_RACE	Students reproted race/ethniciyt
STARBUCKS_N	# of Starbucks locations in student's home county
VOC_07_09	% of studnets from home county that attend voactional/tech school
WSCH_07_09	% of students from home county that work and attend school
Distance	Categorical distance of student's home from campus

Variables added for the 5th Week Model	**Description**
ATTEMPTED_HOURS	# of hours attempted for the fall semester
COLL_CODE	College of student's 1st major
FATHER_EDUC	Father's education
FIFTH_WK_ABSENT	# of absences at 5th week
FIFTH_WK_FAIL	# of failing grades at 5th week
F_FAMILY_CONTRIBUTION	Estimated family contribution
F_INST_NEED	Institutional based financial aid amount
F_INST_NEED_LOANS	Institutional based financial aid amount
F_INST_NEED_NONNEED	Institutional based financial aid amount
F_INST_NEED_NONNEED_LOANS	Institutional based financial aid amount
F_INST_NEED_NONNEED_WORKSTUDY	Institutional based financial aid amount
F_INST_NONNEED	Institutional based financial aid amount
F_INST_NONNEED_LOANS	Institutional based financial aid amount
F_INST_NONNEED_WORK_STUDY	Institutional based financial aid amount
F_INST_WORK_STUDY	Institutional based financial aid amount
F_PELL_GRANTS	Pell grant amount
F_PERKINS	Perkins loan amount
F_PLUS	Plus loan amount
F_SUB_STAFFORD	Subsidized stafford loan amount
F_UNSUB_STAFFORD	Unsubsidized stafford loan amount
F_WORK_STUDY	Work study amount
HONORS_FLAG	Indicators for Honors College participant
MAIN_PCT	% of courses taken on main campus
MARITAL_STATUS	Student's marital status
MASTER_PLAN	Indicator for Master Plan (summer program) participation
MOTHER_EDUC	Mother's education
N_ADVISEES	# of advisees associated with student's advisor
ONCAMP	Indicator for on campus living
PASS_RATE	Measure of difficulty of student's schedule

PELL ELLIGIBLE	Flag for reciept of Pell Grant
UNIV_EXP	Type of university experience course

Variables Added for the Full Semester Model	**Description**
ADJ_FALL_GPA	First semester GPA including developmental course work
GREEK	Indicator of participation in greek organizations
N_WITHDRAWS	# courses dropped throughout semester
PCT _ATTEMPTED_EARNED	% of attempted hours earned for the fall semester
UNIV_EXP_GRADE	Grade in university experience course

Methods and Preliminary Results

Three years of data for first time first year degree seeking students was partitioned into training and validation subsets and used to implement logistic regression, decision trees, neural networks, and ensemble models using SAS Enterprise Miner. Variables included varied based on three time periods: pre-enrollment, 5th week enrollment, and the completion of the first semester. The target for prediction was full year fall to fall retention. For logit and neural network models, missing data was imputed utilizing SAS Enterprise Miner's distribution imputation function for categorical variables and Andrew's Wave m-estimation for missing quantitative variables. More advanced techniques such as expectation maximization or multiple imputation is not available in the SAS Enterprise Miner environment. Forward selection based on validation misclassification was utilized for variable selection for the logit models.

Logistic Regression

Logistic regression is an alternative method for modeling discrete choice utilizing maximum likelihood estimation based on the logistic function as opposed to ordinary least squares (linear probability models). A major advantage of logistic regression for dichotomous dependent variables is that it overcomes the inherent heteroskedasticity associated with linear probability models (Hosmer and Lemeshaw, 2000).

Neural Networks

A neural network can be thought of as a nonlinear model of complex relationships composed of multiple 'hidden' layers (similar to composite functions)

> **A neural network is a two-stage regression or classification model, typically represented by a network diagram. The central idea is to extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of these features. (Hastie, et al 2009)**

For the neural networks, a multilayer perception architecture (with 3 hidden units) was specified. This architecture consists of a single layer with hyperbolic tangent activation functions.

The specification of neural network model architecture requires choosing among a number of alternatives including the number of hidden units and hidden layers as well as combination and activation functions that connect the layers and ultimately are used to make predictions. There is no set theory or established methodology for these specifications and the best specification is typically the result of some model selection approach (Arifovica and Ramazan, 2001).

SAS Enterprise Miner provides an automated model selection process through the 'AutoNeural' node which conducts limited searches for optimizing network configurations, based on some initial user settings. In this project, we specified a single layer architecture which adds hidden nodes one at a time and using a variety of activation functions.

Although some attempts have been made to quantify variable importance in the context of neural networks (Gevrey, et al 2003), SAS Enterprise Miner does not directly accommodate variable selection for neural networks. Inputs selected by a logistic regression using stepwise selection or decision trees were used in both the multilayer perceptron and autoneural network implementations.

Decision Trees

The decision tree algorithm we implemented (using SAS Enterprise Miner- similar to CART and CHAID) searches through the input space and finds values of the input variables (split values) that maximize the differences in the target value between groups created by the split. Splits are evaluated based on specified measures of 'worth' (for instance, given the settings we used in SAS Enterprise Miner this is based on the (-)log of the chi-squared p-value associated with the split adjusted for previous splits). The process for determining the best split is two stage. First, for each individual variable in the input space, the best splitting value is obtained based on worth. Then the level of worth is compared across all the inputs, and the input with the best level of worth is chosen. The data is partitioned *at the best splitting value on the best input*. Speaking heuristically, the variable chosen for the initial split in the tree may be analogously interpreted as *'the most important'* or the variable that '*explains the most variation in the dependent variable.'* The tree is 'grown' until all possible statistically significant splits are produced, and then 'pruned' based on cross validation misclassification to produce an optimal tree. The final model is characterized by the split values for each explanatory variable and creates a set of rules for classifying new cases.

Ensemble Models

An ensemble model can be thought of as a collection of a number of predictors (models) (Krogh, 1997).In SAS Enterprise Miner we implemented an ensemble model consisting of the logit, neural network (using the autoneural specification), and decision tree models.

Model Performance

With the primary goal of predicting retention outcomes and minimizing generalization error, we produce metrics such as misclassification and percentage of correct predictions for each model. (as is common in the machine learning literature, and the 'fit' metric preferred by Hastie, et al, 2009; Kennedy, 2003; Studenmund, 2001; Thomas, Dawes, and Reznik, 2001)

For teach time period, the order from best to worst (in terms of percentage of correct predictions for the prediction of 'not retained') for each model was:

Pre-Enrollment 1) autoneural 2) neural network 3) ensemble 4) logistic regression 5) decision tree

5th Week 1) decision tree 2) autoneural 3) ensemble 4) logistic regression 5) neural network network

Full Semester 1) ensemble 2) neural network 3) autoneural 4) decision tree 5) logistic regression

However, the practical differences between the performance of each model in each time period is not that great. The decision tree is only at a clear disadvantage to other models in the pre-enrollment period.

The decision tree algorithm is robust to missing data, allowing us to avoid the necessity of data imputation or the pitfalls of listwise deletion, and despite the relevance of some of the practical results obtained by Angrist and Pischke, 2008; Astin and Dey,1995; Cleary and Angel, 1984; D'Agostino,1971; Lunney 1970; enables us to avoid theoretical concerns associated with linear probability (OLS) and logit models for dichotomous outcomes such as inherently heteroskedastic variance. In addition they were preferred based on reasoning similar to Delen, in that decision trees portray a more transparent model structure and explicitly illustrate the logical process associated with outcomes as opposed to neural networks or ensemble models (Delen, 2010). For these reasons, we chose to move forward with decision trees as our champion model for implementation.

The predictive accuracy of our models improved with each shift in time from the additional information that we were able to add to the training data as it became available (such as financial aid, course performance, absenteeism etc.).

Variables	Model	Overall	Not Retained	Not Retained
		% Correct	% Correct	% Target Captured
Pre Enrollment	Logit	0.71	57.2954	26.3072
	Neural Network	0.71	58.9286	26.9608
	AutoNeural	0.72	59.8662	29.2484
	Decision Tree	0.70	53.4014	25.6536
	Ensemble	0.71	58.209	25.4902
5th Week	Logit	0.75	66.0274	39.3791
	Neural Network	0.74	64.9171	38.3987
	Autoneural	0.75	67.1348	39.0523
	Decision Tree	0.73	69.1244	24.5098
	Ensemble	0.75	66.185	37.4183

Full Semester	Logit	0.79	75.2427	50.6536
	Neural Network	0.79	77.095	45.098
	AutoNeural	0.79	77.0718	45.5882
	Decision Tree	0.79	75.1185	51.7974
	Ensemble	0.80	77.5561	50.817

With each split created by the decision tree, a metric for importance based on the Gini impurity index is calculated. The *best variable* or analogously the one that explains the most variation in the dependent variable is receives a value of 1 and the remaining variables are represented as a fraction relative to that variable.

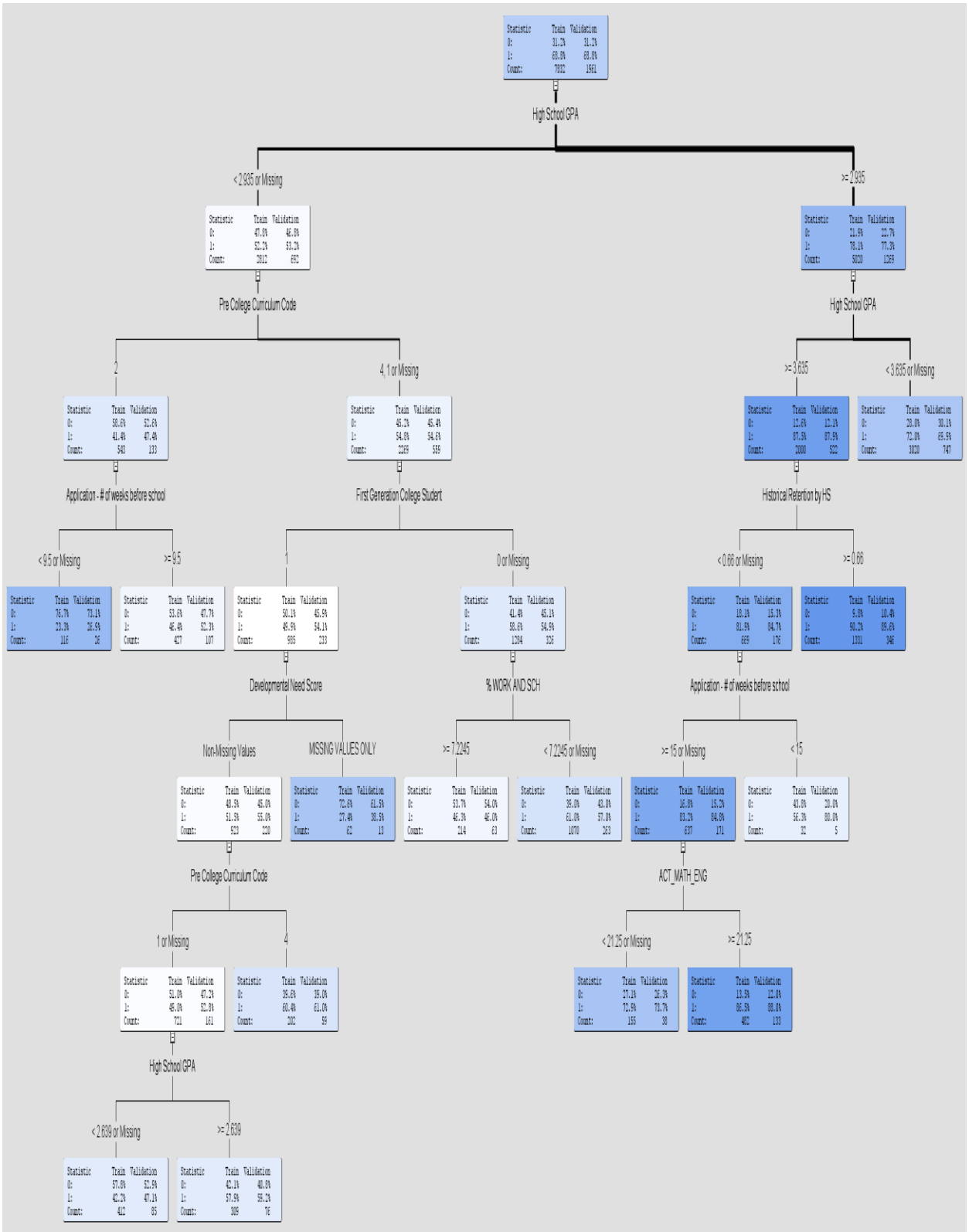
Pre-Enrollment	**Importance**
High School GPA	1
Pre College Curriculum Code	.25
Application Time	.21
First Generation	.16
% Work &. College	.15
Developmental Needs	.15
High School Retention	.14
ACT Math & English Composite	.11

In the above case, High School GPA characterizes the first split in the tree, and Pre College Curriculum Code accounts for about 25% of the variation in retention accounted for by High School GPA.

5th Week	**Importance**
High School GPA	1
Fifth Week Assessment, Absenteeism	.69
KEES	.51
Fifth Week Assessment, D/Fs	.38
Developmental Needs	.27
Attempted Hours	.23
INST_NONNEED (non need based institutional award)	.20
ATP	.19
Application Time	.13

Full Semester	**Importance**
Adjusted Term GPA	1
KEES	.27
Greek	.18
Distance	.11
# Withdraws	.09
Attempted Hours	.09

(Pre-Enrollment) Decision Tree Example



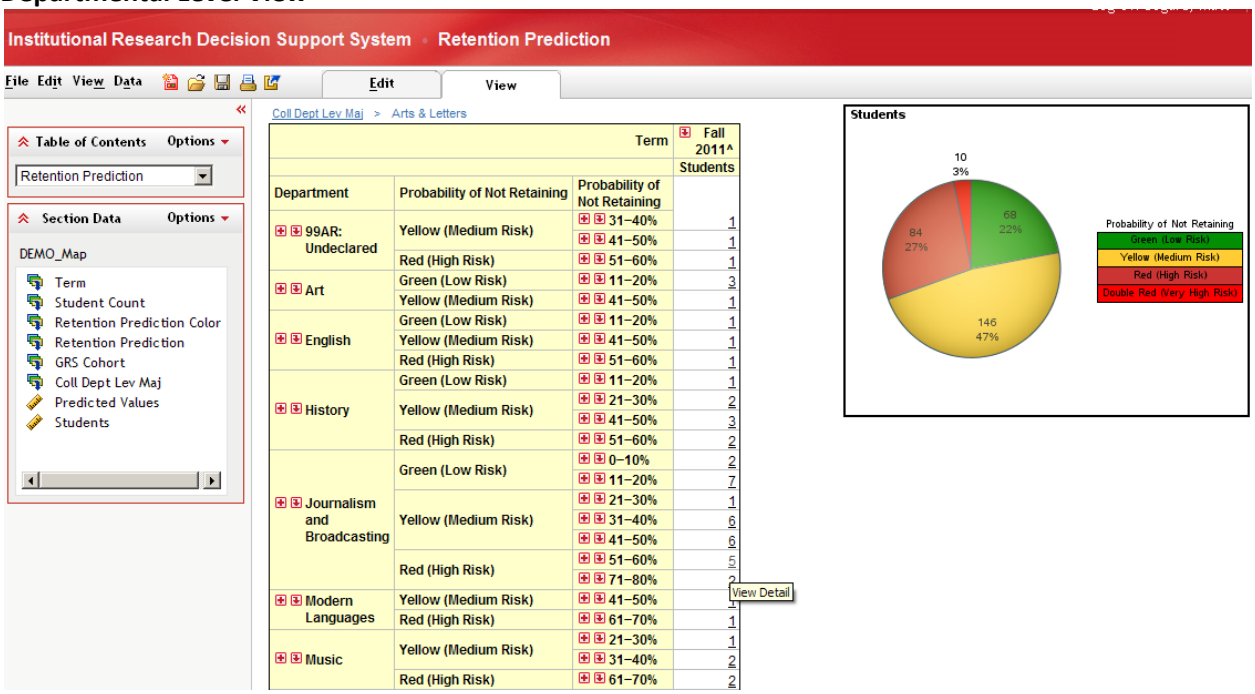
Model Implementation

Using each model's score code generated by SAS Enterprise Miner, the predicted probabilities associated with each student's risk of not retaining can be classified into a 'risk category' (i.e. 'double red' =very high risk, 'red' = high risk, 'yellow' = moderate risk, and 'green' = low risk.) These results can then be incorporated into the OLAP cubes and delivered through the SAS BI Server allowing custom reporting for at risk students. This allows faculty and professional staff at our institution to easily incorporate advanced analytics into their retention strategy on an ongoing basis, without having to rely on manually generated lists of at risk students, or less precise ad hoc reports generated solely on the basis of intuition.

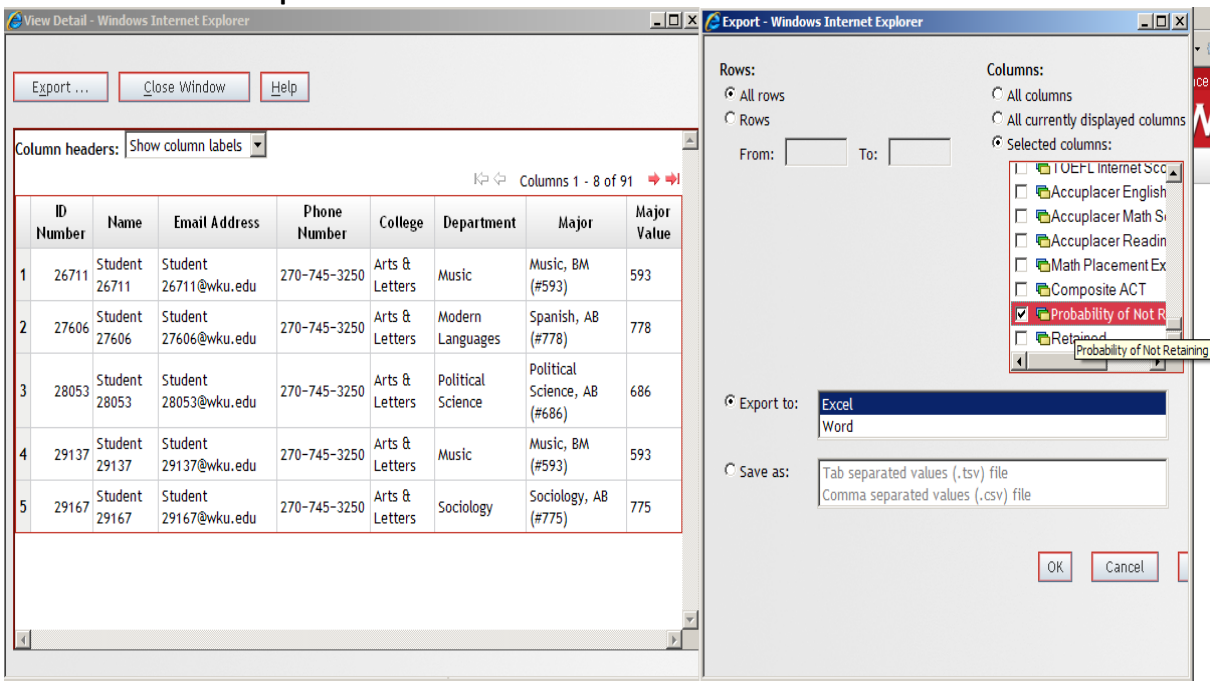
Screen Shots from Prototype Model Implementation vis SAS BI Sever Using Simulated Data

Dashboard View

Departmental Level View



Student Level –Data Export View



Further Research

While these results are encouraging, further research involves a second look at model specification including settings related to the split search algorithm utilized by SAS Enterprise Miner's decision tree node in addition to the exploration of alternative network architectures for neural networks. Improvements definitely need to be made in terms of model accuracy at the pre-enrollment stage. Other research in the Office of Institutional Research has indicated that certain course types and combinations may be related to retention. Incorporating a course index or re-weighted GPA based on this research may improve our results for models implemented once grade information becomes available. We would also like to extend the scope of our predictive modeling effort to include admissions and recruitment targeting as well as success beyond the first year. As similarly stated in Miller and Herreid, there is no reason to assume causality with regard to many of the variables in our models, and administrators may not address them (Miller and Herreid, 2008). However, through the use of survey data (perhaps similar to Dey and Astin, 1993; Stratton et al,2008; Miller and Herreid, 2008) perhaps actionable variables can be included. Despite the fact that some research has indicated that predictive power may not be improved with the addition of survey data vis-à-vis data from institutional data bases (Caison,2007), we hope that the addition of pre-enrollment survey data could improve our first model, given so little pre-enrollment data is available in our data base. Our results currently rival the predictive accuracy of those in the literature that employed logistic regression utilizing survey data. Other areas of exploration include incorporation of metrics from social network analysis and text mining as well as exploitation of geographical information to include spatial econometric techniques (see also Anselin, 1999)

References:

Adelman, C. (1999). Answers in the Tool Box, US Department of Education, Washington, DC.

Angrist, Joshua D. & Jörn-Steffen Pischke. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press. NJ. 2008.

Anselin , Luc. Spatial Econometrics. (CSISS) Center for Spatially Integrated Social Science Presentation: Bruton Center School of Social Sciences University of Texas at Dallas Richardson, TX 75083-0688 http://www.csiss.org/learning_resources/content/papers/baltchap.pdf

Arifovica, Jasmina and Ramazan Gencay. Using genetic algorithms to select architecture of a feed forward artificial neural network. Physica A 289 (2001).

Bean, J. P. (1980). Dropouts and turnover. The synthesis and test of a causal model of student attrition. Research in Higher Education, 12(2), 155–187.

Bean, J. P., & Eaton, S. B. (2000). A psychological model of student retention. In J. M. Braxton (Ed.), Reworking the student departure puzzle (pp.48-61). Nashville, TN: Vanderbilt University Press.

Braxton, J. M., Sullivan, A. S., & Johnson, R. M. (1997). Appraising Tinto's theory of college student departure. In J. C. Smart (Ed.), Higher education: A handbook of theory and research, Vol. 12 (pp. 107–164). New York City: Agathon Press.

Braxton, J. M. (2000). Reworking the student departure puzzle. Nashville, TN: Vanderbilt University Press.
Braxton, J. M., Hirschy, A.S, & McClendon, S. A. (2004). Understanding and reducing college student departure. San Francisco: Jossey-Bass. (ASHE-ERIC Higher Education Report No. 30.3).

Brewe, Eric, Kramer, Laird, and George O'Brien. Investigating Student Communities with Network Analysis of. Interactions in a Physics Learning Center. Physics Education Research Conference 2009. Part of the PER Conference series Ann Arbor, Michigan: July 29-30, 2009. Volume 1179, Pages 105-108.

Caison , Amy L. Analysis of Institutionally Specific Retention Research: A Comparison Between Survey and Institutional Database Methods. Research in Higher Education, Vol. 48, No. 4 (June 2007), pp. 435-451.

Chapman, D. and Pascarella, E. Predictors of academic and social integration of college students. Research in Higher Education, 1983, (19), pp. 295-322.

Correa,Alehandro, Gonzalez, Andres, and Ladino, Camilo. Genetic Algorithm Optimization for Selecting the Best Architecture of a Multi-Layer Perceptron Neural Network: A Credit Scoring Case. Paper 149-2011 SAS Global Forum 2011

Dey ,Eric L. and Alexander W. Astin. Statistical Alternatives For Studying College Student Retention: A Comparative Analysis of Logit, Probit, and Linear Regression. Research in Higher Education, Vol. 34, No. 5. 1993.
DeWitz , S. , Lynn ,Joseph M. Bruce, Woolsey W. Walsh College Student Retention: An Exploration of the Relationship Between Self-Efficacy Beliefs and Purpose in Life Among College Students. Journal of College Student Development. January/February 2009 vol 50 no 1

D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine 34 (2) (2004) 113–127.

D. Delen, R. Sharda, P. Kumar, Movie forecast guru: a web-based DSS for Hollywood managers, Decision Support Systems 43 (4) (2007) 1151–1170.

Delen, Dursun. Decision A comparative analysis of machine learning techniques for student retention management. Support Systems 49 (2010) 498–506

Gevrey, M., Dimopoulos, I., Lek, S. 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol. Model., 160: 249-264.

Hasti, Tibshirani and Friedman. Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer-Verlag. 2009.

Herzog , Serge. Measuring Determinants of Student Return vs. Dropout/Stopout vs. transfer: A First to Second Year Analysis of New Freshmen. Research in Higher Education. Vol 46 No 8 Dec 2005.

Herzog , Serge. Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis-à-Vis Regression. NEW DIRECTIONS FOR INSTITUTIONAL RESEARCH, no. 131, Fall 2006

Hosmer, David W. and Stanely Lemeshaw. Applied Logistic Regression. Second Edition. Wiley. New York. 2000.

Krogh, Anders and Peter Sollich. Statistical Mechanics of Ensemble Learning Physical Review E (Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics), Volume 55, Issue 1, January 1997, pp.811-825

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

X. Li, G.C. Nsofor, L. Song. A comparative analysis of predictive data mining techniques, International Journal of Rapid Manufacturing 1 (2) (2009) 150–172.

Miller, Thomas E. Will They Stay or Will They Go?: Predicting the Risk of Attrition at a Large Public University. College & University, v83 n2 p2-4, 6-7 2007.

Miller, T.E. and C.H. Herreid. Analysis of Variables to Predict First year Persistence Using Logistic Regression Analysis at the University of South Florida. College & University. Vol 83 No 3 2008.

Pascarella, E.T., and P.T. Terenzini (1978). The relation of students' precollege characteristics and freshman year experience to voluntary attrition. Research in Higher Education, 9, 347-366.

Ronco, Sharron and John Cahill. Does it Matter Who's in the Classroom? Effect of Instructor Type on Student Retention, Achievement and Satisfaction. AIR PRofessional File. Number 100, Summer, 2006

Stratton , Leslie S. O'Toole, Dennis M. and James N. Wetzel. Economics of Education Review 27 (2008) 319–331. A multinomial logit model of college stopout and dropout behavior.

Skahill, M.P. (2002). The role of social support network in college persistence among freshmen students. *Journal of College Student Retention, 4*(1), 39-52.

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

St. John, E. P., Cabrera, A. F., Nora, A., and E. H. Asker. (2000). Economic influences on persistence reconsidered. In J. M. Braxton (Ed.), Reworking the student departure puzzle (pp. 29–47). Nashville: Vanderbilt University Press.

Thomas, Scott L. Ties that Bind: A Social Network Approach to Understanding Student Integration and Persistence. The Journal of Higher . Education. Vol. 71. No 5. 2000.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45(1), 89–125.

Yu, Chong Ho. DiGangi, Samuel. Jannasch-Pennell, Angel and Charles Kaprolet A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. Journal of Data Science 8(2010), 307-325.