May 2011

# Empirical Methods for Predicting Student Retention- A Summary from the Literature

Matt Bogard
*Western Kentucky University*, matt.bogard@wku.edu

## Empirical Methods for Predicting Student Retention- A Summary from the 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 Univeristy, 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, in terms of 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).

**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

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.

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

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

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

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.