Student Retention Modelling: An Evaluation of Different Methods and their Impact on Prediction Results

Joe J.J. Lin
Purdue University, Indiana, U.S.A.
linjj@purdue.edu

P.K. Imbrie
Purdue University, Indiana, U.S.A.
pkimbrie@purdue.edu

Kenneth J. Reid
Ohio Northern University, Ohio, U.S.A.
k-reid@onu.edu

Abstract*: Entering engineering students' cognitive data from high school and their non-cognitive self-beliefs can be influential factors affecting their academic success and retention decision. Effectively modelling the relationships between these early available factors and student's future status of persistence in engineering can be particularly valuable to improve student retention in engineering. In this paper, twenty retention modelling systems were developed based on a combination of five retention models and four prominent modelling methodologies. These five retention models contain different collections of cognitive and/or non-cognitive factors, ranging from 9 to 71 input variables. The four modelling methodologies compared are: neural networks, logistic regression, discriminant analysis and structural equation modelling. Prediction performance results from these twenty modelling systems show that 1) neural network method produced the best prediction results among these four methods consistently, and 2) models combining both cognitive and non-cognitive data performed better than cognitive-only or noncognitiv-only models.*

Introduction

Every year a group of good quality graduates from high schools entered the engineering programs across this country, with remarkable academic record in terms of grade point average and standardized test scores. However, as reported in various studies, the number of students switching of engineering continues to be a major issue (Augustine, 2005; Beaufait, 1991). In a study of over 300 universities, Astin found that only 47% of first-year engineering students eventually completed their engineering degree (Astin, 1993). For the long-term competitiveness of United States, this attrition problem in engineering programs is too critical to ignore.

To effectively assist the first-year students with timely advising and intervention starting from their first semesters, an accurate predictive model of retention that use only pre-college data are highly desirable. In this work, the authors developed new predictive systems based on four different modelling methods and five different sets of pre-college factors. These systems are aimed to help discover the non-persistent students in early stage. The predictive performances from different systems were then compared to evaluate the strength and weakness of competing modelling methods and collections of predictor variables. Discoveries from this research will be valuable in helping future researchers develop more effective models of student persistence in engineering. It is our belief that, with an effective predictive system on student retention, a well designed intervention program can then be performed in time to help retaining more quality students in engineering.

Research question

How do retention models, which make use of methods such as neural networks, logistic regression, discriminant analysis or structural equation modelling, compare in their performance in predicting first-year students' retention in engineering after one year?

Theoretical framework

Imbrie et al. have proposed the Model of Students' Success (MSS) in engineering as a framework of important factors and major outcomes related to engineering students' success in academics and career (Imbrie, Lin, & Malyscheff, 2008). In this study, the main scope of our investigation is demonstrated in Figure 1, which contains a subset of the factors and outcomes from the aforementioned MSS framework. Our focus is on predicting students' retention in engineering after one year with pre-college cognitive and non-cognitive variables.

Retention modelling systems based on neural networks (NN), logistic regression (LR), discriminant analysis (DA) and structural equations modelling (SEM) methods were developed independently to capture the relationship between potential factors and the outcome of student's retention after one year. Detailed description of these factors and outcome will be provided in following methodology section.

Methodology

Data collection

Independent Variables: The students' ***non-cognitive*** measures were collected across nine scales in a self-reported online survey completed prior to the freshman year (Immekus, Maller, Imbrie, Wu, & McDermott, 2005). These scales are: Leadership, Deep vs. Surface Learning Types, Teamwork, Self-efficacy, Motivation, Meta-cognition, Expectancy-value, and Major decision.

The following eleven ***cognitive*** items were also collected: overall GPA and core GPA from high school, standardized test results, average high school grades in mathematics, science, and English classes and finally the number of semesters taking mathematics, science, and English.

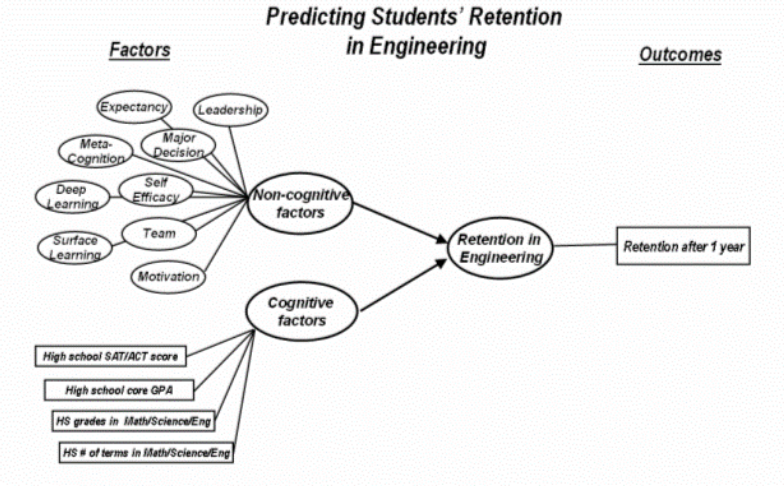

Figure 1: Predicting Student's Retention in Engineering

Dependent Variables: Students' persistence in engineering was collected at the beginning of semester following their freshman year. Students remaining in the lower-division and upper division engineering programs were considered as "retained" students. The students transferred to majors other than engineering, or leave the university completely were classified as "not-retained".

Participants: The participants in this study included 1,508 incoming first-year engineering students (289 females, 1,219 males) at a large Midwestern university during the 2004-2005 academic year. Ethnicity was as follows: 2.05% African American, 0.51% American Native, 10.18% Asian/Pacific Islander, 2.64% Hispanic, 82.43% Caucasian, 2.20% Other.

Modelling methodologies

Through literature reviews, several modelling methods were found to be applied in prior educational researches to predict students' retention. The more frequently used ones are logistic regression, discriminant analysis and structural equation modelling (SEM). These three statistics based methods, plus neural networks from machine learning community, were applied to develop retention models in this study.

Logistic regression (LR) has been broadly used in educational studies to predict student's retention or graduation status. Levin and Wyckoff (1991), House (1993), Schaeffers et al. (1997), Beserfield-Sacre et al. (1997), Zhang & RiCharde (1998) have all used logistic regression models to study student persistence in colleges. More recently, Besterfield-Sacre et al. (2002) developed a logistic regression model to predict first year engineering student's first-term probation and reported an overall classification accuracy as 68.8%. French

et al. (2005) studied the enrollment status in engineering after 6 or 8 semesters using logistic regression model and reported a 65% correct classification rate. Among these studies on students retention using LR models, only Schaeffers et al. (1997) reported a correct classification rate on retention that is higher than 70%. However, their model requires the use of college cumulative GPAs as the most important factor to predict the 3-5 year persistence, and therefore is less suitable for implementing early proactive advising for freshman students.

Discriminant analysis (DA) is another method used in modelling college student retention in prominent literatures. Pascarella and Terenzini (1983) studied students' withdrawal status at the end of freshman year using discriminant analysis, and reported correct classification rates from 77% to 81%. However their factors were collected during the student's first year and therefore less suitable for early intervention. Fuertes and Sedlacek (1994) used discriminant analysis and pre-college cognitive and non-cognitive factors to study retention for college Asian students. They reported 64% and 68% correct classification for 5th semester and 7th semester retention. Burtner (2005) studied the enrollment status after one year for engineering students and reported 85.2% correction classification. However, his data were collected in the later part of second semester (April), which also makes his approach less suitable for early intervention with freshman students.

Structural equation modelling (SEM): Aitken (1982) developed a four equation structural model of student satisfaction, performance, and reported that 19.4% of the variance in the student retention can be explained by his model. Nora et al. (1990) studied the relation between retention and pre-college factors and reported the factors in their SEM model accounted for 15.3% of the variance in retention. Cabrera et al. (1993) also use SEM to model college student retention after one year. They reported 45% of the observed variance in retention can be accounted by their model, with the most significant factors as college GPAs after first year. French et al. (2003) studied the relation between enrollment in engineering with factors including high school rank, SAT scores, university GPAs, motivation, and faculty/student integration. They found their SEM model accounted for 11% of the observed variance in enrollment in engineering.

Neural Networks (NN) is a well developed modeling approach among the various tools within the Artificial Intelligence (AI) community. During the past decades it has been widely used in technical applications involving prediction and classification, especially in areas of engineering, business and medicine (Kukar, Kononenko, Groselj, Kralj, & Fettich, 1999; Smith & Gupta, 2002; Tsoukalas & Uhrig, 1997). The neural network model is especially attractive for modeling complex systems because of its favorable properties: universal function approximation capability, accommodation of multiple non-linear variables with unknown interactions, and good generalization ability (Coit, Jackson, & Smith, 1998). More modelling details on applying NN to predict student retention in engineering can be found in Imbrie et al. (2008).

Retention Models:

Five different forms of the base retention model (models A, B, C, D and E as shown in Table 1) were used in this investigation to evaluate the influence of modelling methodology on predicted results.

Table 1: Five retention models with different input factors

	Models				
Model ID	**A**	**B**	**C (=A+B)**	**D**	**E (=B+D)**
Input factors (No. of independent variables)	Averaged scores from each of the 9 non-cognitive constructs (9)	11 cognitive items from pre-college academic records (11)	Combination of inputs from model A and B (20)	Selected 60 items from the 168 item non-cognitive survey (60)	Combination of inputs from model B and D (71)
Output result (No. of dependent variable)	Persistence status in engineering after one year (1)				

Prediction performance indices

Five performance measures are used to present the prediction performance of these retention systems:

Overall Accuracy for prediction measures the fraction of accurate predictions within the total number of all observations. Its range is 0 to 100%. The perfect score is 100%.

POD Retained: Probability of detection (POD) for retained student measures how well the model predicts over those who are actually retained. Its range is 0 to 100%, with a perfect score of 100%. POD Retained equals to 100% means 100% of the retained students were predicted correctly.

POD Not-Retained: Probability of detection for not retained student measures how well the model predicts over those who are actually not retained. Its range is 0 to 100%, with a perfect score of 100%. POD Not-Retained equals to 100% means 100% of the not retained students were identified correctly. Other studies may refer to this measure as "sensitivity" for detecting not-retained students.

Bias Retained measures the ratio of over-estimation or under-estimation on the number of predicted retained students over the number of actually retained students. Similarly, ***Bias Not-Retained*** measures the ratio of over-estimation or under-estimation on the number of predicted not-retained students over the number of actually not-retained students. An over-estimation of 25% will be expressed as Bias Not-Retained = +0.25%. A negative Bias value indicates an under-estimation. Perfect score of 0 means there is no over or under estimation.

Findings and discussion

Twenty sets of prediction results from five different models (A, B, C, D and E as described in Table 1) and four different prediction methods (NN, LR, DA and SEM) are presented in Table 2. These performance results are obtained and validated through k-fold cross-validation procedure with *k=10*.

The risk of reporting only the overall accuracy

Overall prediction accuracy is traditionally reported in literature. However, results from discriminant analysis (DA) in Table 2 show that there is a huge risk of reporting only overall prediction accuracy. If the authors choose to report only the overall accuracy values in Table 2, DA will be the best performing method with overall accuracy values around 80% across all five models. However, a carefully examination of the Bias Not-Retained values for DA results (from -96% to -99%) revealed a serious under-estimation of not-retained students. In another word, these DA models classified very few students as not-retained (i.e., at risk). This is further confirmed with the extremely low probability of detection (POD) for not-retained students in DA results. Therefore, these DA systems achieved a high prediction accuracy of 80% mostly due to the fact that there are only about 20% of not-retained students (even if they were mostly misclassified), in stead of possessing the true capability of identifying students with tendency to leave engineering. For that reason, the authors strongly believe that when reporting results from prediction models in similar studies, the probability of detection for both groups of students, as well as the bias values should be included to allow rigorous comparison. That practice will result in a more consistent way of presenting performances in predictive studies.

As a result of these obvious flaws of discriminant analysis (DA) systems in this study, further comparison of performances from now will focus on systems based on NN, LR and SEM only.

Table 2: Prediction results from four different prediction methods for all five models

Model Input	Methods	Performance measures				
		Overall Accuracy	POD Retained	POD Not-Retained	Bias Retained*	Bias Not-Retained*
A: Non-cognitive variables only (9 scales)	Neural networks	**68.1%**	**76.5%**	**33.2%**	-7.3%	31.6%
	Logistic regression	68.0%	76.4%	33.0%	-7.4%	32.0%
	Discriminant A.	80.6%	99.9%	*0.0%*	24.0%	*-99.7%*
	SEM	67.5%	76.2%	31.6%	-7.3%	31.6%
B: Pre-college cognitive variables only (11 items)	Neural networks	**70.3%**	**78.0%**	**38.3%**	-7.3%	31.6%
	Logistic regression	69.5%	77.4%	36.5%	-7.5%	32.3%
	Discriminant A.	80.4%	99.8%	*0.0%*	23.8%	*-98.9%*
	SEM	69.9%	77.7%	37.1%	-7.3%	31.6%

C: Both cognitive and Non-cognitive (20 items)	Neural networks	**71.9%**	**79.0%**	**42.4%**	-7.3%	31.6%
	Logistic regression	70.3%	78.0%	38.1%	-7.3%	31.6%
	Discriminant A.	80.4%	99.7%	*0.3%*	23.6%	*-98.2%*
	SEM	71.3%	78.6%	40.4%	-7.3%	31.6%
D: Non-cognitive variables only (60 items)	Neural networks	**68.7%**	**76.9%**	**34.6%**	-7.3%	31.6%
	Logistic regression	67.6%	76.1%	32.3%	-7.6%	32.7%
	Discriminant A.	80.6%	99.6%	*1.4%*	23.3%	*-96.8%*
	SEM	68.5%	76.8%	34.2%	-7.4%	31.9%
E: Both cognitive and non-cognitive (71 items)	Neural networks	**71.7%**	**78.8%**	**42.0%**	-7.3%	31.6%
	Logistic regression	71.5%	78.7%	41.7%	-7.3%	31.6%
	Discriminant A.	79.9%	98.9%	*1.1%*	22.7%	*-93.9%*
	SEM	71.0%	78.4%	40.3%	-7.3%	31.6%

* To achieve a foundation for direct comparison between different modelling methodologies, the authors have purposefully maintained a similar level of Bias Not-Retained for NN, LR and SEM systems.

Comparing models with different sets of cognitive and non-cognitive independent variables

So, which collection of variables provides the better input to predict retention in engineering after one year?

Cognitive-only model performs better than non-cognitive only models: From Table 2 and Figure 2-3, we found cognitive-only model B performed better than non-cognitive-only model A and D in the three major performance indices (Overall Accuracy, POD Retained and POD Not-Retained), while maintaining same levels on the two controlled bias values. This advantage is true across all three modelling methods (NN, LR, SEM) as shown in Figure 2-3.

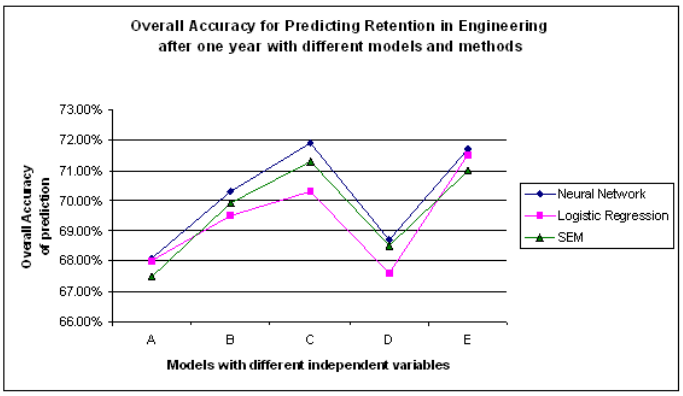

Figure 2: Overall Accuracy for predicting retention in engineering after one year with different models and methods

Combination models performs better than both cognitive-only and non-cognitive-only models: Similarly, we found model C, with combined inputs from model A and B, predicted better than both A and B across all three major indices, by all three methodologies (NN, LR, SEM). Also, combination model E performed better than model B and D individually. This finding supported the advantage of combining both cognitive and non-cognitive factors in one retention model over using non-cognitive only, or cognitive only factors. This is consistent with results reported in previous studies by **Besterfield-Sacre et al. (1997)**.

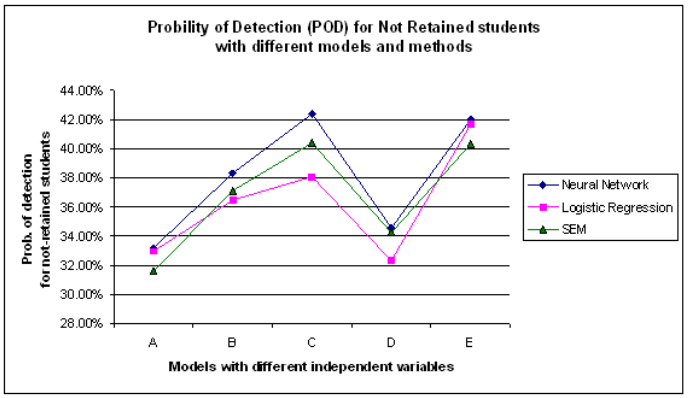

Figure 3: Probability of detection for not retained students after one year with different models and methods

Comparing modelling methods to predict student retention

Performance results through k-fold cross-validation showed systems developed with NN consistently outperform LR and SEM models in all five input models (A, B, C, D and E) in all three major performance indices. This is consistent with the findings from Dreiseitl and Ohno-Machado (2002). When comparing methods between LR and SEM, however, the results are mixed. Models with SEM performed better than LR in three models (B, C and D), but

worse in model A and E. The remaining method DA, with its inflexible **dichotomous outputs and linear nature of discriminant functions,** did not demonstrate the ability to detect at-risk students. Therefore we suggest future researchers use caution and avoid applying DA in student retention models similar to the nature of this study.

Recommendations

Modelling/predicting matriculation of beginning engineering students has an obvious benefit of identifying at-risk students early-on who could benefit from tailored intervention programs to improve retention. Model results can also be used to provide faculty and advisors with informed course selection advice to first-year engineering students. Future work will attempt to improve POD Not-Retained students.

Reference

Aitken, N. D. (1982). College Student Performance, Satisfaction and Retention: Specification and Estimation of a Structural Model. *Journal of Higher Education*, *v53*(n1), p32-50.

Astin, A. W. (1993). Engineering Outcomes. *ASEE Prism*, 27-30.

Augustine, N. (2005). *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future.* Washington, D.C.: National Academies Committee on Prospering in the Global Economy of the 21st Century.

Beaufait, F. W. (1991). *Engineering education needs surgery*, West Lafayette, IN, USA.

Besterfield-Sacre, M., Atman, C. J., & Shuman, L. J. (1997). Characteristics of freshman engineering students: Models for determining student attrition in engineering. *Journal of Engineering Education, 86*(2), 139-149.

Besterfield-Sacre, M., Shuman, L., Wolfe, H., Scalise, A., Larpkiattaworn, S., Muogboh, O. S., et al. (2002). *Modeling for Educational Enhancement and Assessment*. Paper presented at the Annual Conference of American Society for Engineering Education.

Burtner, J. (2005). The Use of Discriminant Analysis to Investigate the Influence of Non-Cognitive Factors on Engineering School Persistence. *Journal of Engineering Education, July 2005*.

Cabrera, A., Nora, A., & Castaneda, M. (1993). College Persistence: Structural Equation Modeling Test of an Integrated Model of Student Retention. *Journal of Higher Education, vol. 64*, pp. 123-129.

Coit, D. W., Jackson, B. T., & Smith, A. E. (1998). Static neural network process models: considerations and case studies. International Journal of Production Research, 36(11), 2953-2967.

Dreiseitla, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics, 35*, pp.352-359.

French, B. F., Immekus, J. C., & Oakes, W. (2003). *A structural model of engineering students success and persistence*. Paper presented at the Frontiers in Education, 2003.

French, B. F., Immekus, J. C., & Oakes, W. C. (2005). An Examination of Indicators of Engineering Students' Success and Persistence. *Journal of Engineering Education*, p.419-425.

Fuertes, J., & Sedlacek, W. (1994). Using the SAT and Noncognitive Variables to Predict the Grades and Retention of Asian American University Students. *Measurement and Evaluation in Counseling & Development, V.27*, p.74-84.

House, J. (1993). The Relationship Between Academic Self-Concept and School Withdrawal. *Journal of Social Psychology, vol. 133*, pp. 125-127.

Imbrie, P. K., Lin, J. J., & Malyscheff, A. (2008). *Artificial Intelligence Methods to Forecast Engineering Students' Retention based on Cognitive and Non-cognitive Factors*. Paper presented at the Annual Conference of American Society for Engineering Education, 2008.

Immekus, J. C., Maller, S. J., Imbrie, P. K., Wu, N., & McDermott, P. A. (2005). *Work in progress - an analysis of students' academic success and persistence using pre-college factors*, Indianopolis, IN, USA.

Kukar, M., Kononenko, I., Groselj, C., Kralj, K., & Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artif Intell Med, 16*(1), 25-50.

Levin, J., & Wycokoff, J. (1991). Predicting persistence and success in baccalaurate engineering. *Education, 111*(4), 461-468.

Nora, A., Attinasi, L. C., & Matonak, A. (1990). Testing Qualitative Indicators of Precollege Factors in Tinto's Attrition Model: A Community College Student Population. *Review of Higher Education, V. 13*(3), P.337.

Pascarella, E. T., & Terenzini, P. T. (1983). Predicting Voluntary Freshman Year Persistence/Withdrawal Behavior in a Residential University: A Path Analytic Validation of Tinto's Model. *Journal of Educational Psychology, V.75*(2), p.215-226.

Schaeffers, K. G., Epperson, D. L., & Nauta, M. M. (1997). Women's Career Development: Can Theoretically Derived Variables Predict Persistence in Engineering Majors? *Journal of Counseling Psychology, V. 44*, pp. 173-183.

Smith, K. A., & Gupta, J. N. D. (2002). *Neural networks in business : techniques and applications*. Hershey, PA: Idea Group Pub.

Tinto, V. (1975). Dropout from higher education: A theoretical Synthesis of Recent Research. *Review of Educational Research, v. 45*, pp. 89-125.

Tsoukalas, L. H., & Uhrig, R. E. (1997). *Fuzzy and neural approaches in engineering*. New York: Wiley.

Zhang, G., Anderson, T. J., Ohland, M. W., & Thorndyke, B. R. (2004). Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. *Journal of Engineering Education, 93*(4), 313-320.

Zhang, Z., & RiCharde, R. S. (1998). *Prediction and Analysis of Freshman Retention*. Paper presented at the Annual Forum of the Association for Institutional Research (AIR).

Acknowledgements

The researchers wish to acknowledge the support provided by a grant from the National Science Foundation, Division of Engineering Education and Centers (Award No. 0416113).

Copyright statement