AC 2010-1711: COMPARISON OF FOUR METHODOLOGIES FOR MODELING STUDENT RETENTION IN ENGINEERING

P.K. Imbrie, Purdue University

Joe Jien-Jou Lin, Purdue University

Kenneth Reid, Ohio Northern University

Comparison of Four Methodologies for Modeling Student Retention in Engineering

Abstract

Several methodologies based on statistical methods or machine learning theories have been applied in previous studies for the modeling of student retention. However, most prior studies were based solely on a specific modeling method of authors' choice. Direct comparison of competing methods using identical collection of student retention data was rarely provided.

The purpose of this paper is to present a direct comparison of prominent methods for modeling student retention using the same data. Four modeling methodologies (neural networks, logistic regression, discriminant analysis and structural equation modeling) are included in this study. These competing methods were implemented on five retention models with various collections of cognitive and non-cognitive factors, ranging from 9 to 71 variables. The retention data in this study were collected from more than 1500 first year engineering students in a large Midwestern university. The eleven cognitive attributes include high school GPAs, standardized test scores, and the grades and number of semesters in math, science and English courses in high school. The non-cognitive variables were collected through Student Attitudinal Success Instrument (SASI), covering the following nine constructs: Leadership, Deep Learning, Surface Learning, Teamwork, Academic Self-efficacy, Motivation, Metacognition, Expectancy-value, and Major Decision.

The following findings are found during this study. First, among the five retention models, the two hybrid models with both cognitive and non-cognitive factors always perform better than models consisting of either only cognitive, or only non-cognitive factors. Second, the addition of non-cognitive items can significantly improve the prediction performance of a cognitive-only model when applied properly. Third, neural network methods perform better than the other three methodologies in performance indices, followed by logistic regression. However, logistic regression may be attractive to some researchers for its ease in implementation and lower requirements for computation power. Finally, the authors found the commonly used threshold (0.05) for including variables in stepwise selection process in logistic regression may not result in the best model for prediction performance. The authors strongly suggest that researchers explore beyond this typical threshold in order to find the best performing collection of variables.

Introduction

Exceptional high school graduates with excellent grade point averages and standardized test scores enter engineering programs across this country. However, as reported in various studies, the number of students switching out of engineering majors continues to be a major issue[1,2]. In a study of over 300 universities, Astin found that only 47% of first-year engineering students eventually completed their engineering degree[3].

To effectively assist the first-year students with timely advising and intervention beginning with their first semester, an accurate predictive model of retention using only pre-college data is highly desirable. The authors have developed new prediction systems based on four different modeling methodologies and five different sets of pre-college factors. These systems are aimed to help discover the non-persistent students in early stage. The prediction performances from different systems were then compared to evaluate the strength and weakness of competing 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 early to help retaining more quality students in engineering.

Research Question

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

Methodology

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[4]. In this paper, the main scope of our investigation focuses on a subset of the factors and outcomes from the aforementioned MSS framework.

Retention modeling systems based on neural networks (NN), logistic regression (LR), discriminant analysis (DA) and structural equations modeling (SEM) methods were developed independently to capture the relationship between these predictive factors and the outcome of student's retention after one year.

A. Data Collection

Independent Variables*:** The students' ***non-cognitive measures were collected across nine scales in a self-reported online SASI survey completed prior to the freshman year[5,6]. These scales are: Leadership, Deep vs. Surface Learning Types, Teamwork, Self-efficacy, Motivation, Metacognition, 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 the number of semesters taking mathematics, science, and English.

Dependent Variables: Students' persistence in engineering was collected at the beginning of semester following their first academic 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.

B. Methodologies for Prediction

Through literature reviews, several modeling methods were found to be employed in prior educational research to predict students' retention. The most frequently used are logistic regression, discriminant analysis and structural equation modeling (SEM). These three statistics based methods, plus neural networks from artificial intelligence and machine learning techniques, were applied to develop retention models in this study.

Logistic regression (LR) has been broadly used in educational studies to predict student retention or graduation status. Levin and Wyckoff[7], House[8], Schaeffers et al.[9], Beserfield-Sacre et al.[10], Zhang & RiCharde[11] have all used logistic regression models to study student persistence in colleges. More recently, Besterfield-Sacre et al.[12] developed a logistic regression model to predict first year engineering student first-term probation. Their results showed 86% of first term probation students were identified, with an overall classification accuracy of 68.8%. French et al.[13] 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 student retention using LR models, only Schaeffers et al.[14] reported a correct classification rate on retention that was higher than 70%. However, Schaeffers' model requires the use of college cumulative GPA 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 modeling college student retention in prominent literature. Pascarella and Terenzini[15] 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[16] 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[17] studied the enrollment status after one year for engineering students and reported 85.2% correction classification. However, Burtner's data were collected in the later part of second semester , which makes his approach less suitable for early intervention with freshman students.

Structural equation modeling (SEM): Aitken[18] 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.[19] 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.[20] also used 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 GPA after first year. French et al.[21] studied the relation between enrollment in engineering with factors including high school rank, SAT scores, university GPA, 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 machine learning 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[22,23]. 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[24]. More modeling details on applying NN to predict student retention in engineering can be found in Imbrie et al.[4].

C. Retention Models

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

Table 1. Retention models A through E with different input factors

	Models				
Model ID	**A**	**B**	**C**	**D**	**E**
Input factors (No. of independent variables)	Non-cognitive constructs (9)	Non-cognitive survey items (60)	Pre-college Cognitive factors (11)	Both cognitive constructs and non-cognitive factors (20)	Both cognitive survey items and non-cognitive factors (71)
Description of factors	Average scores of each of the 9 non-cognitive constructs from the 168- item SASI survey	Selected 60 items from the 168 item SASI survey	11 pre-college cognitive items as described in data collection	Combination of model A and C; 20 input variables in total	Combination of model B and C; 71 input variables in total
Output result (No. of dependent variable)	Persistence status in engineering after one year (1)				

D. Prediction Performance Indexes

Five performance related indexes are used to present the prediction performance of these retention prediction systems. The detailed mathematical formulas for each can be found in Imbrie et al[4]. Among these indexes, the first three are used to express the prediction performance with different focus on the groups (whole population, retained students, and not retained students). The remaining two mainly measure the bias levels. These biases are meant to be controlled to provide a foundation for fair comparison between different model and methods.

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. An over-estimation of 25% will be expressed as Bias Retained = +0.25%. A negative Bias value indicates under-estimation. Perfect score is 0, which means there is no over or under estimation.

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 under-estimation. Perfect score is 0, which means there is no over or under estimation.

Results and Discussion

A. The Risk of Reporting Only "Overall Accuracy"

Overall prediction accuracy (or classification accuracy, correct classification rate) is traditionally reported in literature, sometimes alone and sometimes with other indexes. However, results from discriminant analysis (DA) shown in Table 2 illustrate a very important, but sometimes overlooked risk for only reporting overall prediction accuracy results. The warning message here: overall prediction accuracy value alone can be very misleading.

Why may the prediction system with the highest overall accuracy not be the most desirable one?

If only the overall accuracy values in Table 2 were reported, the method 3a with discriminant analysis (3a-DA in later discussion) will be the best performing method with overall accuracy 80.4%. Accuracy of 80% on student retention can be considered respectable when compared with previously published work. However, when we examine the probability of detection for not-retained students (POD Not-Retained) and bias values, they reveal another side of the story. In this example, the Bias NotRetained value for 3a-DA results (-98.2%) indicates a serious under-estimation of not-retained students. In other words, this 3a-DA model predicted very few students (about 2% of the actual number) as not-retained. Therefore, there is an extremely low probability of detection (POD) for not-retained students (0.3%) in this 3a-DA model. Since the early identification of at risk students is important for our purpose, 3a-DA is actually the least performing model in our opinion even it seems to provide a misleadingly high overall prediction accuracy. In conclusion, this 3a-DA system achieved a high prediction accuracy of 80% mostly due to the fact that there are only about 20% of not-retained students (even though almost all of them were misclassified with 3a-DA), instead of possessing the effective power to distinguish students with different persistence tendencies.

Table 2. The risk of reporting only overall accuracy

Model D: Both cognitive and Non-cognitive factors (20 items)	**Performance measures**				
Prediction Method	**Overall Accuracy**	**POD Retained**	**POD Not-Retained**	**Bias Retained[1]**	**Bias Not-Retained[2]**
1. Neural networks modeling	71.9%	79.0%	42.4%	-7.3%	31.6%
2a. Logistic regression (forward stepwise selection for variables)	70.3%	78.0%	38.1%	-7.3%	31.6%
2b. Logistic regression (keep all factors in model)	71.7%	78.8%	41.5%	-7.3%	31.6%
3a. Discriminant Analysis (forward stepwise, group membership results generated directly by SPSS 17.0)	***80.4%***	***99.7%***	***0.3%***	***23.6%***	***-98.2%***
3b. Discriminant Analysis (forward stepwise, grouping threshold selected by user to control bias)	70.1%	77.8%	37.5%	-7.3%	31.6%
3c. Discriminant Analysis (keep all factors, grouping threshold selected by user to control bias)	71.4%	78.6%	40.8%	-7.3%	31.6%
4. Structural equation modeling	71.3%	78.6%	40.4%	-7.3%	31.6%

1. In this study, the bias levels were controlled by selecting proper grouping threshold values so that only 25% of students are considered as at risk by each method. This yields a relatively consistent bias level across all methods (except 3a) for a fair comparison in this study. Method 3a uses direct grouping/prediction output from SPSS 17.0 without controlling of biases. It is included for illustration purpose.

2. The attrition rate for this freshmen cohort (2004, *N*=1508) is 19%. When systems are controlled to identify 25% of population as at risk, the over-estimation of attrition is 6% of students (i.e., 25%-19%). This translates into *Bias Not-Retained* of 31.6% (i.e., 25%/19% -1).

Which type of output from prediction systems is preferred? Should we process the probabilities of group membership, or just use the predicted group membership?

Continuing the previous discussion, if the prediction model was tested using a different population where only 60% of students were retained, the overall prediction accuracy of 3a-DA will likely drop significantly. This is because 3a-DA lacks the ability to detect not-retained students, as *POD Not-Retained* and *Bias Not-Retained* values in Table 2 suggest. To correct this problem in 3a-DA, instead of using the direct grouping status (0 or 1) from the output of discriminant analysis by SPSS, the authors take over control on selecting the threshold for the grouping probabilities (real numbers between 0 and 1, also generated by DA/SPSS) to classify students into two groups. This practice enables us to obtain a better control on how many students we will considered as at risk, and therefore the control of classification biases. The results are also shown in Table 2 as method 3b, with bias values much more in line with other methods for effective comparison. The experience above lead us to the following suggestion: even though the statistical software packages (such as SPSS) do provide outputs in both final group membership (0 or 1) and the probability of group membership, use extra caution when you adopt the final group membership prediction directly. In the example of 3a-DA, 80% of overall prediction accuracy does not mean much value to us if the system only detects 0.3% of not-retained students. On the other hand, the probability of group membership values for each student will offer more detailed information that works much better for our needs. As a result all final prediction performance discussed in this study are based on the probability of group membership, rather than the dichotomous (0 or 1) status prediction directly from software packages.

Table 2 and the above discussion illustrate the risk of reporting only the overall prediction accuracy in similar prediction models. A retention prediction model can have a high overall accuracy number but possess extremely low power to identify students of our interest if these students are in smaller proportion of the whole population.

Recommendation on reporting performance results from prediction models/systems

The authors therefore strongly recommend that when reporting results from prediction models in similar researches, the probability of detection (or *sensitivity*) for both groups of students, and the bias values (over/under estimation ratio) should be included as required information in addition to the overall prediction accuracy. That practice will provide a more consistent way of presenting prediction/classification results and facilitate rigorous comparisons across different studies.

B. Comparing Models with Different Collections of Variables

Retention prediction systems based on four prominent prediction methodologies and five different models are developed in this study. For logistic regression and discriminant

analysis methods, both the "enter all variables" and "stepwise selection" approaches for determining variables in model were implemented for the purpose of comparison. Therefore, 6 variations of methods for each of these 5 input models are studied, as shown in Table 3. This results in a matrix of a total of 30 competing prediction systems. For every combination of methods and models, the authors purposefully controlled the classification threshold to allow only 25% of students be predicted as at risk. This helps us maintain a consistent bias level across all prediction systems to facilitate a fair comparison throughout this study.

Table 3. Variations of prediction methodologies

Method ID	Prediction Methodologies
NN	Neural networks modeling
LR_All	Logistic regression; enter all variables in the model
LR_Step	Logistic regression; forward stepwise selection for selecting variables in model
DA_All	Discriminant analysis; enter all variables in the model
DA_Step	Discriminant analysis; forward stepwise selection for selecting variables in model
SEM	Structural equation modeling

Figure 1 to Figure 3 present the prediction results across all 30 prediction systems. Every system is a combination of one prediction methodology and one model type. All 30 prediction systems go through same K-fold cross-validation process (with K=10) using the identical set of student data (N=1508). It is worth mentioning that the K-fold cross-validation process always keeps the testing data and training data separated so that the data used for testing are never seen or used in training by the model they are tested upon. This process ensures that these testing results from K-fold cross-validation are as close to testing brand new data as possible. The average performance results of 10 cross-validation runs are then plotted in Figure 1-3 for final comparison.

What does the probability of detection (POD) mean?
Why do we need to control the bias level when comparing prediction performances?

The performance indexes shown respectively in each figure are: overall prediction accuracy, probability of detection (POD) for retained students, and POD for not-retained students. We are especially interested in the last index, POD for not-retained students (POD_NR in later discussion). This POD_NR index expresses the system's ability to detect students at risk of attrition (being not-retained). For example, one system in our study predicted 25% of students as at risk and achieved a 46.6% POD Not-Retained. As this performance is achieved through a student population with 19% actual attrition, the Bias_NR (over-estimation for not-retained students) is then 31.6% (i.e., 25%/19% - 100%). A POD_NR of 46.6% means it successfully identified 46.6% of all not-retained students through the 25% of student population it predicted as at-risk. In comparison, a pure "random guess" predictor will be expected to hit 25% of the not-retained students with the allowed 25% population. Using "random guess" predictor as baseline for comparison, this system has identified 86.4% (46.6%/25%- 100%) more at-risk students than expected result from random guess. If this exact same prediction system was allowed to predict 50% of students as at risk, the POD_NR will significantly rise to

73.8%. However the cost will be a much higher Bias_NR which may not be desirable for practical reasons.

This explains why these performance index values are highly dependent on the bias levels the researchers control. It is not meaningful to compare systems' overall accuracy or probability of detection values unless they have the similar bias level. Therefore the performance values presented in Figure 1-3 are meant to be compared among systems with same controlled biases. When compared with prediction results with different bias values, these performance numbers will need to be adjusted.

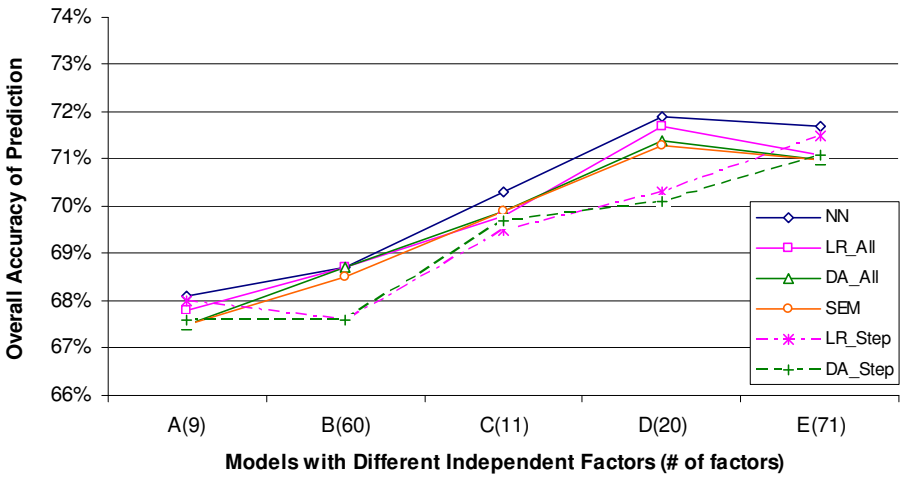

Figure 1. Overall accuracy for predicting retention in engineering after one year, with different models and methods

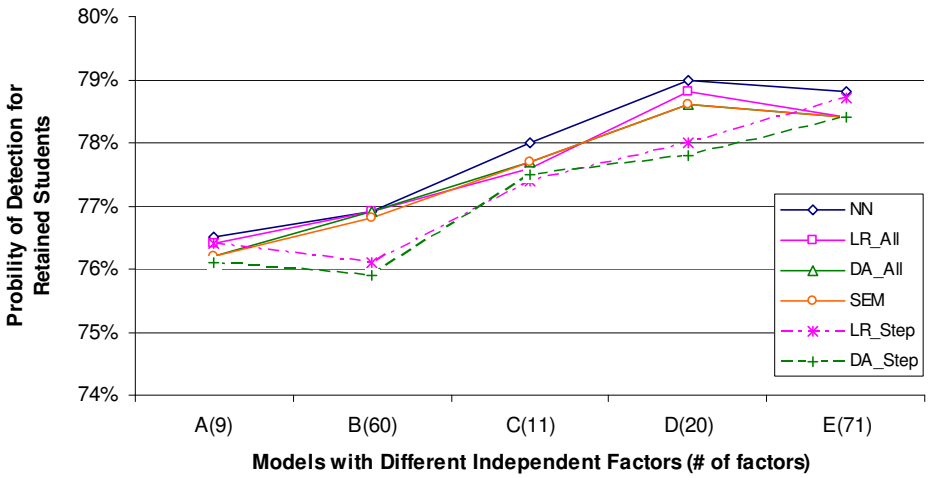

Figure 2. Probability of detection (POD) for retained students, with different models and methods

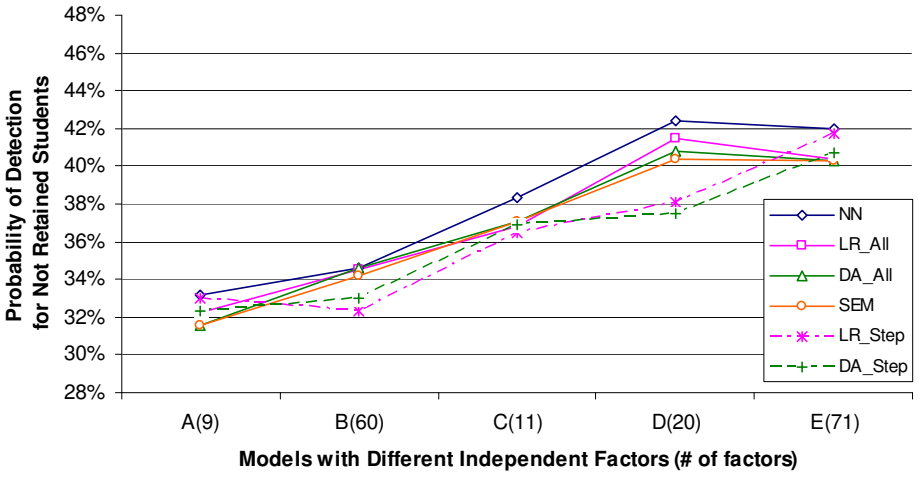

Figure 3. Probability of detection (POD) for not retained students, with different models and methods

Similar patterns across three prediction performance indexes

Although these three performance indexes reside in different numeric ranges, they noticeably present very similar patterns across figures. This observation is noteworthy to us, because it allows us to focus our discussion on one of the three indexes first, and the other two indexes will generally support rather than dispute our comparison results from the first index. Since we are especially interested in the last index, POD for not-retained students (POD_NR), most of our discussions below will be centered on POD_NR (as in Figure 3); with the other two indexes as supportive reference if necessary. This POD_NR index in essence expresses the system's ability to detect students at risk of attrition (being not-retained).

The first question of our interest is: which collection of independent variables (as realized in models A through E) offered better set of input for predicting retention in engineering after one year?

> *Finding: Hybrid models with both cognitive and non-cognitive variables predict better than cognitive-only models or non-cognitive-only models using the same methodology.*
>
> *Finding: Cognitive factor models perform better than non-cognitive factor models, regardless of which method is applied.*
>
> *Finding: Although not the best performing models when standing alone, the non-cognitive factors do improve the ability to detect at-risk students (POD_NR)*

significantly when added into cognitive only model, as the hybrid models demonstrate.

As shown in Figure 3, we found hybrid models D and E perform significantly better than cognitive model C and non-cognitive models B and A, when the same prediction method is used. This implies that, although hybrid models take more effort to collect than cognitive only or non-cognitive only models, the extra efforts does improve the model's prediction performance. Also, cognitive model C performed better than non-cognitive model A and B consistently. This is true for all six variations of methods. This suggests if we want to rely on only one category of factors, the cognitive factors we collected are more powerful predictors than non-cognitive survey factors. However, we need to also point out: since the cognitive factors are from student's high school academic history and standardized test scores, they are generally more accurate in nature than the non-cognitive results which were collected through self-reported on-line surveys. It is very possible that when the non-cognitive survey are improved through survey revision, including new constructs, or better administration strategy, those non-cognitive factors may become much better predictors in future.

Nonetheless, although not the best performing models when standing alone, the non-cognitive factors do improve the ability to detect at-risk students (POD_NR) when added to cognitive-only models as the improvement from model C to D. Even more significant improvement has been found in a newer model E' when adding proper non-cognitive items into models with cognitive items (as described later in part C). Therefore the non-cognitive variables clearly have additional influence on models' prediction of student retention beyond what the cognitive only factors can provide.

C. Improving Models using a Hierarchical Logistic Regression (HLR) Approach

In Figure 3, one intriguing observation emerged. Why do the performance of the top four methods on model D drop when applied to model E, but the two lower performing methods (LR_Step and DA_Step) improve significantly when moving from model D to E? After examining several possible explanations and numerous literature searches, the authors suspect that the most likely key to that unexpected observation is the number of variables in model. As we see from Figure 3, those two prediction methods moving up significantly from model D to E are both stepwise selection variation of logistic regression and discriminant analysis. These stepwise selection models, with their nature of selecting only more important factors into the model, tend to have a much smaller number of variables in their final model formulation. On the other hand, those four methods that perform well in model D but drop lower in model E use all 71 variables (11 cognitive, 60 non-cognitive) provided in model E. To further explore this phenomenon, the authors developed a series of prediction systems based on hierarchical logistic regression (HLR). In this HLR approach, we used the 11 cognitive factors from model C as the starting point for a new model, and added subsets of the 60 cognitive items (from model B) incrementally until the model reached all 71 variables and became model E. The first block of 11 cognitive factors were entered together in the new HLR models, while the second block of non-cognitive items entered the new models in batches based on the "probability for stepwise entry" values in the forward stepwise selection process.

Generally a variable with smaller "probability for stepwise entry" indicates it has higher importance in the model. Therefore it should enter the model earlier than those variables with larger such values.

Figure 4 shows the POD_NR results for new models when using increasing "probability for stepwise entry". The solid line is plotted with average POD_NR values following the primary Y-axis on left, while the dotted line shows the average number of variables included in new models following the secondary Y-axis on right side. Again all the results in Figure 4 are obtained through K-fold cross validation with K=10. Since every point plotted is the average from 10 cross validation runs, the number of variables in model may not be an integer. Also, to display enough relevant information without overcrowding the limited space in figure, the size of increments on X-axis is not constant.

The number of variables in prediction model is not "the more the merrier".

Figure 4 shows a very interesting POD_NR curve (solid line) which starts from model C and eventually reaches model E. The POD_NR performance first improves from model C with small "probability for stepwise entry" value of 0.05, then gradually peaks at "probability for stepwise entry" of 0.175, and then drops back down until it reaches model E. Basically model E uses all 71 available variables and is equivalent to applying a "probability for stepwise entry" of 1.0 on the chart.

The dotted curve, presenting number of independent variables in model, shows a consistent trend of increasing when the "probability for stepwise entry" value increases. This is expected as a larger "probability for stepwise entry" means a lower entry threshold for items to enter the model.

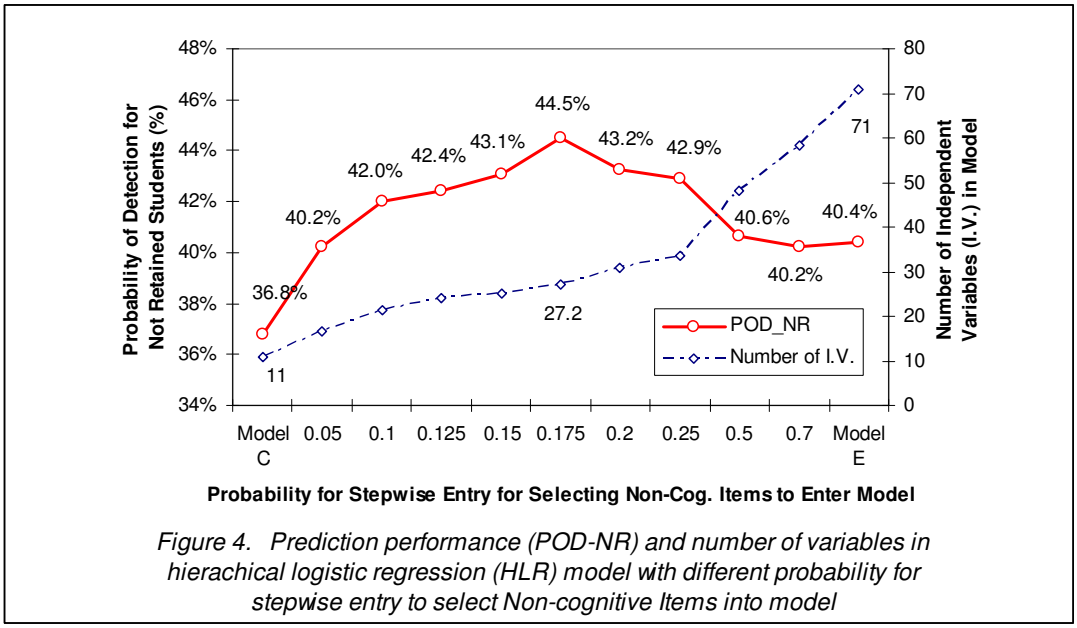

Figure 4. Prediction performance (POD-NR) and number of variables in hierachical logistic regression (HLR) model with different probability for stepwise entry to select Non-cognitive Items into model

Clearly, the models with larger "probability for stepwise entry" generally are the results of adding additional items to models with smaller "probability for stepwise entry" as it just lowered the entry criterion. One would think by including more factors into the larger model, it should at least perform as well as, if not better than, the smaller model with only a subset of its variables. However, this curve in Figure 4 clearly dispelled that myth. Model E obviously contains all the variables that any other model on the figure has and more, but its performance is lagging behind many models which uses only a subset of its variables. More discussion on model selection and subset size issue can be found in statistical literatures such as the work of Hastie et al.[25].

Explore beyond the ordinary path. The commonly used statistical threshold value 0.05 may not be the best choice for "probability for stepwise entry" in stepwise selection process for prediction systems based on logistic regression and discriminant analysis.

Statistical methods such as logistic regression and discriminant analysis usually include the stepwise selection module as its tool for selecting variables into the model. When developing prediction systems using these methods with stepwise selection, the user will be asked to enter the "probability for stepwise entry" as threshold for including variables. The default value in commercial packages such as SPSS is often 0.05 and that is also the value that we observed to be used most often in the literature. The performance results for the two stepwise selection variations of logistic regression and discriminant analysis (in Figure 1-3) are also based on this typical value of 0.05. However, the results in Figure 4 suggest the commonly used 0.05 for "probability for stepwise entry" does not produce the best performing prediction model for us. A better performing model actually was obtained when a much larger value 0.175 was used in the stepwise selection process in our study. Similar results were also found in developing discriminant analysis models. The authors certainly do not attempt to claim that 0.175 is the new magic number when running stepwise selection. We believe this value is relatively dependent on the model type and variables used in each study. Therefore we suggest if the goal is to improve the prediction performance, the researchers may want to explore beyond the 0.05 threshold typically used in stepwise selection process.

A further improved E' model with both cognitive and non-cognitive factors:

Based on the finding in Figure 4, a new E' model with 27 variables is developed. It contains 16 non-cognitive survey items selected through the HLR stepwise selection process and the 11 factors from cognitive model C. After implementing this new E' model with all 6 prediction methodologies, the results are shown in Figure 5.

Prediction performance of cognitive-only model is greatly improved with addition of non-cognitive items.

In contrast with Figure 3, the new E' model clearly outperform all other models in this study when implemented with same methodology. It is noteworthy that with the addition of 16 non-cognitive items into model C, the POD of not-retained students increases significantly from 38.3% in model C to 46.6% in model E', based on neural network

results on top of the figure. Similar amount of improvement is also obtained in other methodologies. This again suggests that the effort of collecting non-cognitive survey data does come with its rewards. These non-cognitive items help bringing the prediction performance of cognitive model C to a much higher level as demonstrated by model E'.

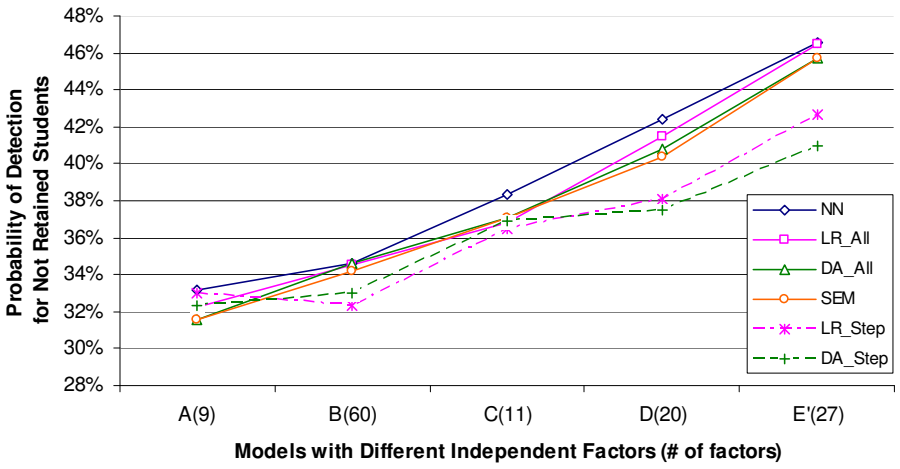

Figure 5. Probability of detection (POD) for not retained students with different models and methods, with model E' replacing E

D. Comparison between Prediction Methodologies

Which prediction methodologies in this study is the preferred choice for similar applications on modeling student retention?

For the 6 variations of methodologies investigated in this study, we found neural networks generally come out on top with each of the three prediction performance indexes, with logistic regression (using all available variables) closely follows as shown in Figure 5. Discriminant analysis and structural equation modeling, both of which are linear models by nature, perform fairly closely to each other and below the previous two methods when given the same variables. The two stepwise variations of logistic regression and discriminant analysis (in dotted lines, with "probability for stepwise entry" as typical 0.05) fall significantly behind the other methods especially in higher performing models D and E. Since previous discussion on Figure 4 have shown us the proper use of stepwise selection with creative threshold value can help improving the model's performance, we do not want to blame this lower performance on stepwise method. Instead, we will again encourage users to question the use of 0.05 as default "probability for stepwise entry" in stepwise selection and explore different values beyond 0.05.

In conclusion, the authors consider neural networks and logistic regression as the preferred methods of choice among those studied here. With a given set of variables, neural networks have the potential to model most complex relations between variables with its non-linear modeling nature, very flexible network structures and plentiful optimization algorithms available to choose from. These advantages are realized in the better prediction performance with most models in this study. Logistic regression, on the other hand, is easier to implement with commercial packages and require less computation time for analysis. It is also easier to perform model selection on logistic regression to improve the existing models with different selection of variables. The authors will suggest keeping both available in the "modeling tool box" and choose according to the different requirements of future modeling tasks.

E. Conclusion and Recommendation

The authors wish to share the following findings:

First, among the five retention models, the two hybrid models with both cognitive and non-cognitive factors always perform better than models which consist of only cognitive or of only non-cognitive factors.

Second, the addition of non-cognitive items can significantly improve the prediction performance of a cognitive-only model when properly applied. This is demonstrated in the process of developing new model E' by adding non-cognitive survey items into cognitive model C.

In the same process we also found that models with more input variables do not always perform better than smaller models with a subset of their variables. This can be also demonstrated by the superior prediction performance of model E' over original model E.

Comparing prediction methodologies, neural network method performs better than the other three methodologies in all three performance indexes, with logistic regression as the second best performing modeling technique. However, logistic regression can be attractive to some researchers with its ease of implementation and lower requirement of computation power and time. Discriminant analysis and structural equation modeling did not perform as well as the first two methods in terms of prediction performance.

Finally, the authors found the commonly used threshold (0.05) for including variables in stepwise selection process in logistic regression may not result in the best model for prediction performance. The authors strongly suggest researchers explore beyond this typical threshold in order to find the best performing collection of variables when employing stepwise selection approach in similar statistical models.

The authors also believe other topics discussed in this paper, such as the warning about reporting only the overall prediction accuracy (classification accuracy), will be beneficial to future researchers interested in developing prediction/classification systems. It is our sincere wish to contribute these findings to the ASEE community through this paper.

Bibliography

1. 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.
2. Beaufait, F. W. (1991). *Engineering education needs surgery*, West Lafayette, IN, USA.
3. Astin, A. W. (1993). Engineering Outcomes. *ASEE Prism*, 27-30.
4. 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.
5. 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*, Indianapolis, IN, USA.
6. Reid, K. J. (2009). *Development of the Student Attitudinal Success Instrument: Assessment of first-year engineering students including differences by gender.* Doctoral Dissertation, Purdue University.
7. Levin, J., & Wycokoff, J. (1991). Predicting persistence and success in baccalaurate engineering. *Education, 111*(4), 461-468
8. House, J. (1993). The Relationship Between Academic Self-Concept and School Withdrawal. *Journal of Social Psychology, vol. 133*, pp. 125-127.
9. 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.
10. 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.
11. 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).
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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*.
18. 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
19. 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.
20. 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.

21. 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.

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

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

24. 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.

25. Hastie, T., Tibshirani, R. & Friedman J. (2009). The Elements of Statistical Learning- Data Mining, Inference, and Prediction, Second Edition, Springer Science Business Media, LLC.