A Generalizable Neural Network for Predicting Student Retention

Cameron Ian Cooper, Ph.D. – Fort Lewis College

Session: Tools, techniques, and best practices of engineering education for the digital generation

Abstract

This research revisits a neural network-based decision support system presented at the ASEE 2009 Northeast Section Conference at the University of Bridgeport. The decision support system identified students who are “at-risk” of not retaining to their second year of collegiate study. At that time, the positive preliminary results presented were based upon a small out-of-sample dataset. This research updates the results using the full 2008 freshman cohort at Fort Lewis College (N = 800). Overall, the system correctly predicted retention for approximately 70% of the freshmen. Additionally, over three fourths of the retainers were correctly identified. This research offers a confirmed generalizable framework for the development of an early alert system that will allow institutions to identify accurately students who are “at-risk” of not retaining to their second year of study. The eleven salient predictors including ethnicity and socioeconomic status also offer insight into how an institution might improve its retention efforts. This research finally proffers possible means (e.g. STEM support classes, student clubs, peer-mentoring, etc...) to assist students identified as “at-risk” in taking steps to mitigate their risk for non-retention.

Introduction

This research follows a line of research started in 2008 [1]. The primary goal of this research is to understand the dynamics of a student body for the purposes of predictive modeling via neural networks. As evidenced, behavioral modeling of a student body in regards to success can be a difficult enterprise [2]. Student body data can become quickly obsolete and of little use for predicting student success.

The initial stage of this research attempted to create a generalizable model for predicting first-semester persistence (i.e. identify students most likely to persist and conversely those students “at-risk” of not persisting to their second semester of study). This model utilized the 2005 and 2006 freshman cohorts to train and test a neural network for production with the 2007 freshman cohort. The neural network was utilized to predict persistence for the entering 2008 freshman cohort. Unfortunately, the neural network did not result in generalizable predictions [3]. The neural network performed only slightly better than following the naïve rule (i.e. predict persistence for all students, the most prevalent category ≈80% of the students persist to their second semester).

Under the hypothesis that going back two years for training data is too far and leads to erroneous predictions via stale data, the author chose to create a neural network using only the 2007 cohort with a small out-of-sample dataset. The resultant predictive system performed well on the 11% out-of-sample data set with predicting non retention at 76.5% accuracy and retention at 76.9% for an overall accuracy of 76.7%. See the corresponding confusion matrix below:

Output / Desired	*Retain(0)*	*Retain(1)*
Retain(0)	39	12
Retain(1)	12	40
Accuracy	77%	77%

Methodology & Results

The first step in the continuation of this research was to test the neural network above constructed with only the latest data with retention results (2007 cohort) on the entire 2008 freshman cohort. The results of this test are given in the confusion matrix below:

Output / Desired	*Retain(0)*	*Retain(1)*
Retain(0)	193	119
Retain(1)	129	361
Accuracy	60%	75%

The predictive model fell short of the expected 77% predictive accuracies found for the out-of-sample dataset test. This discrepancy can likely be explained by the exclusion of the financial aids variable #10. Acc_Priv_Loans – Dollar Amount of Accepted Private Loans and #11 Off _Priv_Sch – Dollar Amount of Offered Private Scholarships listed in the table below. From the time of the out-of-sample test to the time when the retention results for the 2008 freshman cohort were available, the Financial Aid office had recoded and reworked their data entry such that the values for these variables were aggregated into other financial aid measures, thus making the values effectively impossible to retrieve without a tremendous amount of work for disaggregation.

Per the results from the entire 2008 freshman cohort test, the neural network stemming from the single, most current cohort provides a generalizable predictive system for at least the subsequent year's cohort. The system correctly identified three quarters of the retainers and 60% of the non-retainers for an overall predictive accuracy of 69%. Thus, using a neural network constructed via the cohort of the previous year, one can expect to predict correctly the retention of nearly 70% of the freshman cohort at the conclusion of the first semester of study.

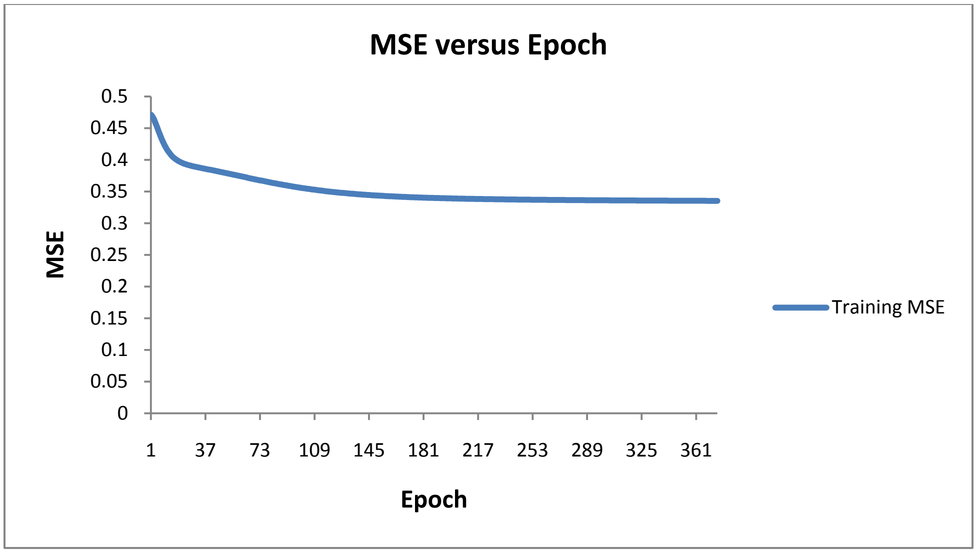
The second step in the continuation of this research was to test the repeatability of the previous steps. In other words, can a neural network from the 2008 freshman cohort be created to predict the retention for the 2009 freshman cohort? Using the same predictors/variables sans variable #10 and variable #11 found in the creation of the neural network described above, a neural network was built from solely the 2008 freshman cohort.

Predictor
1. HS GPA
2. Declared White
3. Academic Standing
4. ACT English
5. CCHE Index
6. Residency
7. Withdrawn
8. Campus Housing
9. Disciplinary Incidents
10. Acc_Priv_Loans
11. Off_ Priv _Sch

The author tested the best neural network on a 15% (N = 120) out-of-sample dataset. The results of this test are given in the confusion matrix below:

Output / Desired	*Retain(0)*	*Retain(1)*
Retain(0)	36	8
Retain(1)	24	52
Accuracy	60%	87%

As with the other neural networks found within this line of research, the 2008 freshman cohort network's learning curve (i.e. Mean Squared Error (MSE) versus training Epoch) leveled off supporting the belief that an adequately trained network without overtraining was found:



The overall accuracy of the neural network was 73.3% with the out-of-sample dataset. This is similar to the finding of a 76.7% accuracy found with an 11% out-of-sample dataset described earlier using the 2007 cohort neural network. Per these results, it appears this methodology of finding a neural network from previous year's cohort to predict retention of the next is both generalizable and repeatable.

Predictive Model Validation

In addition to the neural network model for identifying "at-risk" students at the school, the Student Outreach Services Coordinator also maintains a system for identifying "at-risk" students. At the beginning of each semester, the Services Coordinator sends a list of students identified as being at risk. Faculty members are sent a list of the "at-risk" students appearing on his or her roster. Faculty members are then requested to provide feedback on the progress of each student. Via the direction Student Outreach Services Coordinator, students showing some "type of concern" are then contacted by an appropriate student success service.

The Student Outreach Services Coordinator's Early Alert system identifies "at-risk" students by the following procedure from the following student populations (C.Frankhauser, personal communication, January 28, 2010):

> *Students are chosen from those in Program for Academic Advancement (PAA –[a federally-funded TRIO program]), the Freshman Advising Pilot (FAP – another program for high risk students) and Athletics. All of the staff in these three programs qualitatively picks out the 'riskiest' of their advisees/athletes and sends [their list to the Services Coordinator]. Students who are reinstated from suspension and dismissal are on this list as well.*

As a mathematics professor, the author was sent a list of ten students by the Services Coordinator. Of the ten students, the neural network predictive system identified seven of the ten students as being "at-risk." With a 70% overlap in an effectively random sample between the two methodologies, it appears both early alert systems are measuring the same concept (i.e. academic risk). This validates the use of both models. In regards to the three students where the models disagreed, the neural network outputs for these three records/students were just above the "at-risk" threshold of 0.50 within the transfer function. As evidence of further validation for both models, the course average for the ten students at midterm was 69% or a D+ letter grade. The predictions from both systems are indicative of academic performance, at least for the students in the author's course.

Conclusion and Discussion

In its current incarnation, the neural network-based predictive system for retention can only be used in a reactive means since it relies heavily on predictors determined within the first semester of study (e.g. Variables #2, #7, #8, and #9). Variable #2 – Academic Standing was by far the most influential predictor. Per the Vice President of the Division of Enrollment Management at Fort Lewis College, the most effective student success measures must be taken proactively (i.e. very early in the students' academic careers – within the first three of weeks of study). Waiting until the academic performance for the first semester has been recorded is too late to be effective. Student success resources are extremely limited and must be targeted effectively and efficiently. The Student Outreach Services Coordinator is attempting to reach student reactively in the method described earlier. If her reactive outreach efforts result in positive retention dividends, then the predictive model presented by this paper could be used in conjunction with the Services Coordinator model.

Using the neural network-based predictive system on the 2009 freshman cohort, the system identified 294 students as being "at-risk" of not retaining to their second year of study. The college currently does not have the resources allocated to address this number of students individually. If the Student Services Coordinator's retention efforts are positive, however, then an argument might be able to be made to funnel more resources into reactive retention measures. Otherwise, the next step in this line of research will be to modify the neural network to provide proactive predictions that can then harness the full body of student success research.

References

[1] Cooper C. I. (December 2008) "Predicting Persistence of College Freshmen Using Neural Networks." C. Cooper. ***International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering***.

[2] Cooper, C. I. (April 2009) "Predicting First-year Retention of College Freshmen Using Neural Networks." C. Cooper. *Proceedings* ***of the 2009 American Society for Engineering Education Northeast Conference***.

[3] Chew, B. (February 2009). Analysis of F08 FTF Persistence into W09. Unpublished internal document, Fort Lewis College.