01 Nov 2016

Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy

Gene Lesinski

Steven Corns
Missouri University of Science and Technology, cornss@mst.edu

Cihan H. Dagli
Missouri University of Science and Technology, dagli@mst.edu

Recommended Citation

G. Lesinski et al., "Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy," *Procedia Computer Science*, vol. 95, pp. 375-382, Elsevier, Nov 2016.
The definitive version is available at https://doi.org/10.1016/j.procs.2016.09.348

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 95 (2016) 375 – 382

Procedia Computer Science

Complex Adaptive Systems, Publication 6
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2016 - Los Angeles, CA

Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy

Gene Lesinski[a,*], Steven Corns[b], Cihan Dagli[b]

[a]*United States Military Academy, West Point, NY 10996, USA*
[b]*Missouri University of Science and Technology, Rolla, MO 10996, USA*

Abstract

This paper presents a neural network approach to classify student graduation status based upon selected academic, demographic, and other indicators. A multi-layer feedforward network with backpropagation learning is used as the model framework. The model is trained, tested, and validated using 5100 student samples with data compiled from admissions records and institutional research databases. Nine input variables consist of categorical and numeric data elements including: high school rank, high school quality, standardized test scores, high school faculty assessments, extra-curricular activity score, parent's education status, and time since high school graduation. These inputs and the multi-layer neural network model are used to classify students as: graduates, late graduates, or non-graduates. Several neural network architectures are examined and compared by run time, minimum mean square error achieved (MSE), mean correct classification rate, precision, recall, and specificity. A multi-layer neural network with 50 hidden neurons, momentum value of 0.8, and learning rate of 0.1, with hyperbolic tangent hidden neuron activation functions was able to accurately predict graduation success and achieved the best performance with classification accuracy exceeding 95%, high recall, high precision, and high specificity. This prediction model may be used to inform admission decisions and identify opportunities for required remediation with the potential to improve graduation rates, increase student retention, reduce late graduation, and reduce first-term course failures.

Keywords: Neural network; backpropagation; enrollment management; student retention; classification

doi:10.1016/j.procs.2016.09.348

1. Introduction

All colleges and universities are concerned with graduation rates and retention of their students. Graduation and retention rates are typically used by organizations like Forbes and US News and World Report as proxy indicators of school quality which indirectly impact the institution's bottom line. Graduation and retention rates are particularly important at the United States Military Academy where a retention loss is ultimately a loss to Army officer end strength. Each year, more than 15,000 candidates, from all 50 states, apply for admission to West Point. Approximately 1,200 applicants are accepted each year and receive the equivalent of a four year full scholarship with a Government Accounting Office (GAO) estimated value of $327,000 [7]. Significant effort is applied to graduate a majority of students within four years to satisfy Army officer manning requirements. Recently there has been a spike in the number of first term course failures for entering freshmen at West Point. This has generated interest in reexamining the decision criteria and models that inform admissions decisions. Given the magnitude of commitment associated with admission and the emphasis on four year completion, it is important to closely examine and periodically revalidate the criteria used to make these important admission decisions. Accurately modeling graduation success can ultimately improve graduation rates, increase student retention, reduce late graduation, and reduce first-term course failures. An accurate prediction model can both inform admission decisions as well as identify students requiring remediation. In this research we utilize a multi-layer feedforward neural network with nine selected input variables to model and classify student graduation status to inform admission decisions and identify opportunities for required remediation.

1.1. Related Research

In studies of college graduation success, vast amounts of research are focused on identification of significant predictor variables/factors as well as different mathematical models utilizing these factors to predict successful completion of college. There are numerous studies in the literature regarding factors that may predict successful college graduation. These factors are generally divided into pre-admission and post-admissions considerations. Pre-admissions factors can be further categorized as academic and non-academic. Academic pre-admission factors often include, high school rank, high school grade point average, and standardized test scores.

1.2. Graduation Prediction Factors

Burton and Ramist found that the best combination of SAT scores to be the best predictor of graduation success.[4] Geisler and Santelices conclude that high school GPA was not only the best predictor of first year grades but also for degree completion [8]. Niu and Tienda argue that another measure of high school achievement, high school rank, is a better predictor of college performance than standardized test scores [12]. Black, et al. found a significant correlation between high school quality and student success at college and believe that high school achievement should be adjusted relative to high school quality [2]. Some examined non-academic indicators of college success include social economic status (SES), parental education, faculty references, and high school extra-curricular involvement. Several sources note the strong correlation between parental level of education and the propensity to attend college. Additionally, Nelson identified a significant relationship between parental education and student college success [11]. Willingham identified faculty references and high school activity involvement as two significant non-academic indicators of college success [13]. In this research we high school rank in conjunction with high school quality, SAT scores, parental education, high school faculty assessments, candidate activity scores, and time since high school as factors to predict college graduation success.

1.3. Mathematical Prediction Models

Within the literature there are also a wide variety of modeling approaches applied to prediction of college graduation. Bowen and Bok utilized a logistic regression model to predict graduation within six years using gender, ethnic group, SES, selectivity of the college, SAT scores and high school records as predictive factors [3]. Kanarek achieved successful results using discriminant function analyses to classify students into graduates and non-graduates with a combination of pre-admission and post-admission factors [9]. Yingkuachat, et al. use Bayesian belief networks to determine important college graduation success prediction variables with resulting high prediction accuracy [14]. Karimi, et al. utilize a hybrid decision tree and cluster analysis model to identify at-risk college students [10]. Barker, Trafalis, and Rhoads use neural networks and support vector machines to classifying successful student graduation rates at a 4-year institution utilizing student demographic, academic, and attitudinal information [1]. In this research we employ a multi-layer feedforward neural network with backpropagation learning to predict graduation success.

1.4. Current Methodology Employed

West Point currently uses a proprietary linear combination of five factors to quantify candidate quality and inform admissions decisions. The five factors are CEER, Faculty Assessment Score (FAS), Candidate Activity Score (CAS), Candidate Leadership Score (CLS), and Candidate Fitness Assessment (CFA). Each factor score ranges from a possible 200 minimum to 800 maximum. The CEER is a score intended to capture academic performance/potential and includes factors such as: HS rank, HS quality, SAT/ACT scores. FAS is a score assigned based upon 1 x English, 1 x Mathematics, and 1 x Science teacher assessment of academic potential. CAS is a score assigned based upon depth and breadth of extra-curricular activities. CLS is a score assigned based upon demonstrated leadership duties and activities CFA is a score based upon a standardized physical fitness test. These five factors are combined to formulate a Whole Candidate Score (WCS). A general risk level is established for each individual factor (~500) as well as the WCS (~5200). These levels were determined by a series of linear regression equations. If a candidate has a risk in a sub-factor or WCS, additional analysis is conducted by the admissions committee to make a final determination of qualification status or remediation requirements.

2. Data Requirements and Pre-Processing

Since academic failure is the primary reason for departure or extended duration stay, we primarily consider academic indicators as model inputs. Academic indicators utilized in this research include: HS rank, HS quality, SAT/ACT Math scores, SAT/ACT English scores, and Faculty Assessment scores. Additionally, we include other factors that previously presented research indicates a high correlation to graduation rates/success: HS extra-curricular activity, parent education, and time since HS. The major outputs of the model are whether a student graduates, does not graduate, or graduates late. Figure 1 below highlights the model inputs and outputs. The data descriptions and definitions are included at Appendix A.

2.1. Data Pre-Processing

The required data for this research was collected from two primary sources: West Point admissions database and the annual Cooperative Institutional Research Program (CIRP) survey. The CIRP survey data provides candidate parent’s level of education. All other data elements were retrieved from the West Point admissions database. The combined data was “cleaned” by screening for errors and missing data elements. We delete samples with missing data elements or errors. In our initial data set, the most common missing data elements were parent’s level of education. Removal of samples did not significantly decrease the overall number of data samples or skew the data set. An additional consideration was cadets that left the Academy for their Mormon mission and returned a year or more later. These data samples were also removed as they would skew the results for late graduating cadets. After cleaning the data set there were 5100 data samples from the Classes of 2012-2015 which consisted of 9 input variables and 3 outputs.

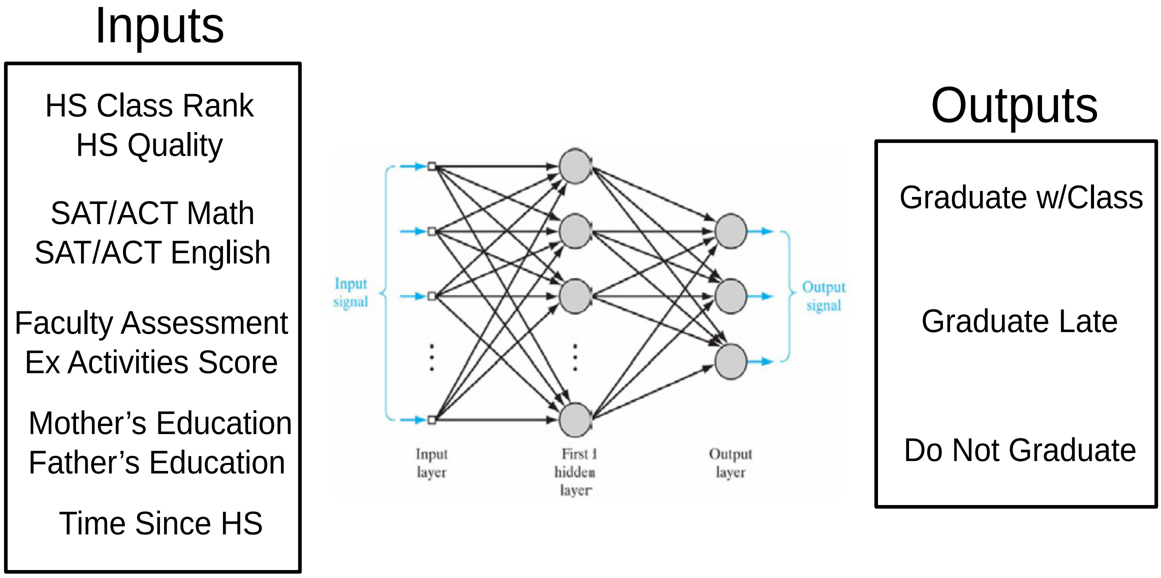


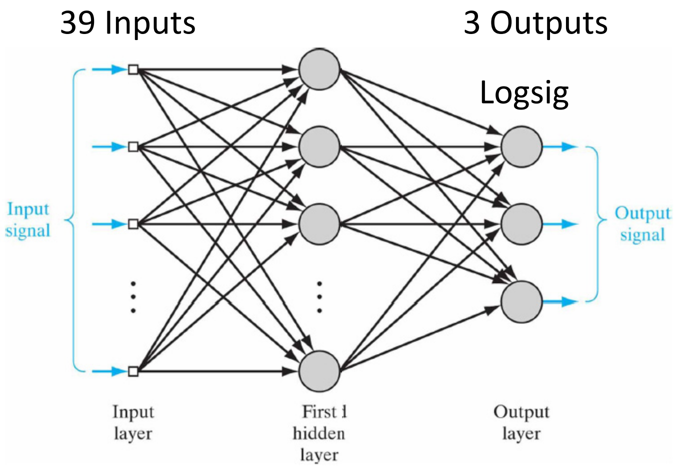
Fig. 1. Model inputs and outputs.

The major data pre-processing tasks required prior to training a neural network for the data previously described are: conversion of SAT/ACT test scores to national percentiles, conversion of categorical variables to binary values, and normalization of numerical data elements. The College Board provides a mapping of SAT/ACT scores to national percentile values [5,6]. Of the nine input variables, five are categorical (HS Rank, HS Quality, Mother's Education, Father's Education, and Years since HS). To convert the categorical variables into binary representations requires transforming a categorical variable into an equivalent number of binary variables. Binary representation of categorical variables was chosen to facilitate future reduction of model variables while minimizing the impact on model structure. The final data pre-processing step is standardizing the SAT data, Faculty Assessment scores, and Activity scores.

3. Model Architecture

In this research we utilize a multi-layer neural network with one hidden layer of neurons. After pre-processing, there are 39 model inputs and 3 model outputs. The number of hidden neurons are varied from 10-70 while examining the impact on model performance. The hyperbolic tangent and logistics activation functions are employed for the hidden layer while the logistic activation function is used for the outputs. Momentum values and learning rates are varied, examining the impact on model performance. Figure 2 below highlights the general neural network architecture and the experimental values. Several training algorithms are explored including: Backpropagation with momentum, Levenberg-Marquardt backpropagation, Variable learning rate backpropagation, and Resilient backpropagation. All models are developed utilizing MATLAB© Neural Network Toolbox.

The neural network utilizing backpropagation with momentum learning is initially used to identify the combination of hidden layer activation function, momentum value, and learning rate that achieves the best overall performance. The results for this initial architecture experimentation indicate that the hyperbolic tangent activation function, momentum of 0.8, and learning rate of 0.1 achieve the best performance results and are used as the values for subsequent architecture experimentation. Next, the number of hidden layer neurons are varied from 10-70 while adjusting the backpropagation learning technique to examine their impact on performance and determine the optimal architecture. The performance metrics used to evaluate the performance of the competing neural network architectures include: MSE achieved, computation time, number of epochs, classification accuracy, recall, precision, and specificity.



Momentum = 0.5, 0.7, 0.8, 0.9
Learning rate =0.001, 0.01, 0.1
Hidden Neurons: 10, 20,30 40,50,60,70
Hidden Activation Fx: Logsig, Tansig

Fig. 2. Neural network architecture and experimental values.

4. Results

The training data set was divided into 70% training, 15% testing, and 15% validation to facilitate model development, experimentation, and performance assessment. A multi-layer a neural network with 50 hidden layer neurons, variable learning rate backpropagation, momentum value of 0.8, initial learning rate of 0.1, and hyperbolic tangent hidden activation functions achieved the best results as represented by the performance metrics discussed above. Figure 3 below highlights the validation confusion matrix for the best model results.

		Predicted			
		Grad	No Grad	Late Grad	
Actual	Grad	524	30	6	93.6%
	No Grad	0	256	0	100%
	Late Grad	0	0	34	100%
		100%	89.5%	85.0%	95.8%

Fig. 3. Validation confusion matrix.

The overall classification accuracy for the best model is 95.8%. The neural network achieves a MSE of 9.9 E-05 in less than 1 second in under 200 epochs. The recall, precision, and specificity performance metrics for the best model with respect to the validation training set are highlighted in Table 1 below.

Table 1. Performance metrics for best neural network architecture.

Performance Metric	Grad	No Grad	Late Grad
Recall	93.6%	100%	100%
Precision	100%	89.5%	85%
Specificity	100%	94.9%	99.3%

5. Conclusion

In this research we employ a multi-layer feedforward neural network with backpropagation learning to predict graduation success. The model is trained, tested, and validated using 5100 student samples with data compiled from admissions records and institutional research databases. Nine input variables consist of categorical and numeric data elements including: high school rank, high school quality, standardized test scores, high school faculty assessments, extra-curricular activity score, parent's education status, and time since high school graduation. These inputs and the multi-layer neural network model are used to classify students as: graduates, late graduates, or non-graduates. The multi-layer neural network with 50 hidden neurons, momentum value of 0.8, and learning rate of 0.1, with hyperbolic tangent hidden neuron activation functions was able to accurately predict graduation success and achieved the best performance. The Levenberg-Marquardt provided quickest solutions in the fewest epochs with varied MSE and accuracy results. Back Propagation with Momentum was the slowest and consistently achieved worst MSE and accuracy results. Variable Learning Back Propagation with Momentum was consistently fast, achieved the best MSE results, and classification accuracy. There was no conclusive results on significance of different number of neurons in hidden layer although 40-60 hidden neurons appears consistently accurate. Classification accuracy in most model architectures exceeded 95%. Most misclassifications were related to identification of late graduates, suggesting that additional filtering/sub-classification may be necessary. Utilizing 2012-2015 West Point admissions data, a multi-layer neural network with 50 hidden neurons and variable backpropagation training, consistently and accurately predicts graduation success with classification accuracy exceeding 95%, high recall, high precision, and high specificity. This prediction model may be used to inform admission decisions and identify opportunities for required remediation with the potential to improve graduation rates, increase student retention, reduce late graduation, and reduce first-term course failures.

Appendix A. Data Description and Definition

Number	Variable	Description	Data Type	Location	Range	Code
1	High School Rank	# of #	Categorical	Input	0-6	0=Top 5% 1=Top 10% 2=Top 20% 3=Top 30% 4=Top 40% 5=Top 50% 6=Bottom 50%
2	High School Quality	% that go to 4 Year	Categorical	Input	0-7	0=90-100% 1=80-89% 2=70-79% 3=60-69% 4=50-59% 5=40-49% 6=30-39% 7=Less than 30%
3	SAT/ACT Math	Percentile Nationally	Numerical	Input	0-100%	N/A
4	SAT/ACT English	Percentile Nationally	Numerical	Input	0-100%	N/A
5	Faculty Assessment	Faculty Assessment	Numerical	Input	400-800	N/A
6	Extra Activities Score	Extra-curricular Activities	Numerical	Input	400-800	N/A
7	Mother's Education	Mother's education	Categorical	Input	1-8	1=Grammar school or less 2=Some HS 3= HS Grad 4= Post Secondary OTC 5= Some College 6= College Degree 7= Some Graduate School 8= Graduate Degree
8	Father's Education	Father's education	Categorical	Input	1-8	1=Grammar school or less 2=Some HS 3= HS Grad 4= Post Secondary OTC 5= Some College 6= College Degree 7= Some Graduate School 8= Graduate Degree
9	Time Since HS	Yrs since HS grad	Categorical	Input	0-3	0= None 1 = 1 year 2=2 years 3 =3 or more years
10	Performance Outcome		Categorical	Output	1-3	1= grad 2=No Grad 3=Late Grad

References

1. Barker, K., Trafalis, T., & Rhoads, T., Learning from student data. In Systems and Information Engineering Design Symposium, Proceedings of the 2004 IEEE, pp. 79-86, 2004.
2. Black, S. E., Lincove, J., Cullinane, J., & Veron, R. Can you leave high school behind?. Economics of Education Review, 46, 52-63, 2015.
3. Bowen, W. and Bok, D., The shape of the river. Princeton, NJ: Princeton University Press, 1998.
4. Burton, N. W., & Ramist, L., Predicting success in college: SAT studies of classes graduating since 1980, 2001.
5. College Board. http://www.collegeboard.com/prod_downloads/highered/ra/act/ACTPercentileRanks.pdf
6. College Board. http://www.collegeboard.com/prod_downloads/highered/ra/sat/SATPercentileRanks.pdf

7. GAO Report 03-1000, September 2003.
8. Geiser, S., & Santelices, M. V. Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. Center for studies in higher education, 2007.
9. Kanarek, E., Exploring the murky world of admissions predictions. Paper presented at the Annual Forum of the Association for Institutional Research, Baltimore, MD., 1989.
10. Karimi, A., Sullivan, E., Hershey, J., Moon, S., A Two-Step Data Mining Approach for Graduation Outcomes, 2013 CAIR Conference, 2013.
11. Nelson, J., Impact of Parent Education on Student Success. Online Submission, 2009.
12. Niu, S. X., & Tienda, M. Testing, ranking and college performance: does high school matter. Princeton University,2009.
13. Willingham, W. Success in college: The role of personal qualities and academic ability. New York: College Board, 1985.
14. Yingkuachat, J., Praneetpolgrang, P., & Kijsirikul, B. An application of probabilistic model to the prediction of student graduation using bayesian belief networks. Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology Association of Thailand (ECTI Thailand), 63-71, 2007.