A Deep Neural Network-Based Predictive Model of Undergraduate Student Retention

Taylor Dawson
Department of Electrical Engineering and Computer Science
George Fox University

Spring 2017

1 Problem Statement

In the last four decades since Tinto's seminal work on student departure (1975, 1993), research on college student retention has become one of the most prolific topics in higher education and has been a major concern for educators and administrators alike [1]. According to Dursun Delen, from the Oklahoma State University, "Student retention is an essential part of many enrollment management systems. It affects university rankings, school reputation and financial wellbeing" [2].

The question of why students leave college before completing a degree is of interest not only to institutions and scholars, but students, employers, parents of students, and spouses. It is crucial to understand a student's premature exodus from college, for those who lack a college degree will most likely have diminished lifetime earnings. At the institutional level the task of identifying the early symptoms of student failure and dropout and designing targeted strategies to support student retention and degree completion is an ongoing concern for all stakeholders [1].

We seek to create a more accurate model that can: 1) better predict the probability of student retention after the first year of enrollment, and 2) discover superior features that can better inform the predictive model of student retention.

2 Previous Work

In higher education extensive research has been conducted related to the topic of student retention, however, few focused on the employment of data mining methods for predicting student retention. Bogard points out that data mining or algorithmic approaches to prediction can provide superior results vis-à-vis traditional statistical modeling approaches [3, 4, 5, 6, 7]. There is a large amount of data that Universities posses that if analyzed properly can result in valuable information and knowledge that can aid in making critical institutional decisions. The term for this is Educational Data Mining (EDM) [8] and it has become a more widely used approach within this area of research, however, they are most often not utilizing the potential of deep neural networks.

It has been shown that in the most complex analysis, where time to graduation is estimated for new and transfer students simultaneously, the pruned neural network with three hidden layers out performed the more traditional method, logical regression. It is also known that given a large set of exploratory predictors used to estimate the degree completion time, data-mining techniques, especially deep neural networks, perform significantly better [9].

One important notion to consider in this specific field of research is generalization. Some of the studies that have been conducted are unable to be generalized to other institutions. One obvious reason for this, is that they analyze data from a single institution. In addition, there is a trend that retention rates of a private institution are significantly greater than those of their public counter parts [10].

A data mining project's success is heavily reliant on the quality and quantity of the data representing the phenomenon under consideration. For instance, the data in Delen's study covered 5 years of freshman

student records. He concludes that given a rather rich set of features, more data and more variables can potentially help to improve the data mining results [2]. That being said, our study would seek to improve upon Delen's by increasing the amount of historical data that is used in the training of the model.

Improvements definitely need to be made in terms of model accuracy at the pre-matriculation stage. For example, the model in Bogard's study was only 71% accurate in predicting, using pre-enrollement data, whether a student would be retained or not. When including data of the full semester it's results were 79% accurate [7]. A similar study by Ji-Wu Jia used Support Vector Machines (SVM) to look at data from Historically Black Colleges and Universities over the course of six years. They were able to improve their SVM retention model to 94.29% [11]. These studies were conducted using longitudinal student data. We seek to attain the same level of accuracy, 90% or above, as the aforementioned studies while only utilizing the data that is available at the start of a students freshman year.

Anther distinction of our study will be the use of deep neural networks, as opposed to the SVM. The reason being that SVMs, essential, reduce down to a single hyperplane through all of the dimensions of data resulting in a binary decision e.g. yes or no.

3 Research Methods

In this section, we outline our method for constructing and evaluating a deep neural network-based predictive model of undergraduate student retention. We describe our corpus of data, features extracted from the corpus, our model itself, and our methods for evaluating the accuracy of our model. The high-level goals of our work are to reliably predict undergraduate student retention—*what is the probability of a given student returning after her first year at the university?*—and to identify the attributes or features of a student that contribute the most to such a prediction.

3.1 Corpus

To train an accurate model of student retention, we require several years' worth of historical student data. We will work with the university's data analytics team to obtain an extract of at least five years' worth of historical student data from the university's transactional and data warehouse systems. This work will focus on traditional undergraduate student retention; therefore, degree completion and graduate students will be excluded from our corpus. We will anonymize personally-identifiable information in accordance with all applicable FERPA regulations. All data will be stored on university-owned systems and accessible only to the project's research staff.

The corpus will contain many attributes for each student, including those reported to the university as part of the admissions process (e. g., high-school grade point average, SAT scores, academic class rank, intended college major) and first year at the university (e. g., college grade point average, enrolled courses, etc), along with traditional demographic attributes such as age, gender, and hometown. We will ensure that the corpus is a fair, representative sample of the student population by stratifying across race, ethnicity, and gender distributions according to the data reported in the university's fact book for the corresponding date range.

3.2 Deep Learning Model

We will use a deep neural network (DNN) as the foundation of our predictive model. DNNs are comprised of a network of layers, each consisting of dozens to thousands of nodes. These layers encode and summarize the inputs of the previous layer, sending a transformed variation of that input to the following layer as its output. We will use the high-level Keras neural network library running on top of TensorFlow, an open-source library for numerical computation using data flow graphs, to implement the DNN [12, 13]. Our work will explore the optimal model topology, which captures mathematical operations as nodes and high-dimensionality arrays ("tensors") as edges in the graph. The flow of data along the graph transforms the input features derived from the raw data into a single output; in our case, the probability of a student returning after the first year of classes.

Once trained, our model could be used to predict the retention probability of any new, unseen student. As the training of the model is the computationally-intensive portion of the work, any trained model could be used on typical, consumer-grade computers if desired.

3.3 Model Validation

Inherent in our method is a stratified k-fold cross-validation experimental design. We will train our model using a stratified partitioning of our corpus, with a portion of the data held out for testing and the remainder used for training for a given fold. We will use `scikit-learn`, the premiere machine learning library for Python, as it provides several standard preprocessing functions, validation models, and accuracy measures [14]. Our k-fold cross-validation technique will provide an accuracy measure given that we have the ground truth for each test: the actual retention of each student in our corpus is known. We will compare the DNN prediction for each student with the known truth and report on our level of agreement.

4 Project Benefits and Outcomes

This project will provide the university with a realistic assessment of the predictive power of a retention model built using state-of-the-art deep learning techniques. The benefit to the university is an increased understanding of factors to that contribute to student retention, both positive and negative, to aid in data-driven decision making. Moreover, we believe the project addresses an issue that is at the forefront of the university's administration, as it is one that directly impacts the individual student as well as the financial health of the entire university.

We envision several potential applications for the resulting model. The model could be used as part of the determination for merit-based financial aid for incoming (and continuing) students. Additionally, the model may eventually be utilized to monitor current students for increased risk of dropping out or otherwise leaving the university before completion. Such a use case could be supported by automatically triggering notifications in a student monitoring and reporting system—realized at our university as the "Fox360" system.

This project also has wide-reaching implications for higher education as a whole, if we succeed in creating an accurate model that generalizes well to other institutions. Such a finding could bring national recognition to George Fox University via publications and presentations in computer science and higher education administration venues. Regardless of degree of success, we plan to present the results of this project at the Consortium for Computer Science in Colleges conference in October 2017. Additionally, we plan to submit our findings to higher education retention-specific venues, such as the National Symposium on Student Success and Retention (hosted by the Consortium for Student Retention Data Exchange founded in 1994 at the University of Oklahoma), to contribute to the larger higher education community.

This project has significant benefits for me, the researcher, both academically and vocational. The experience of the process and completion of a formal research project, will aid me in my future graduate studies. In addition, I will obtain new set of skills, data analytics and utilization of deep neural networks, that will prepare me for my future career. Finally with possession of a publication in my name as an undergraduate, will assist me in achieving my professional goals.

5 Budget

Table 1 lists a summary of our proposed budget expenses. Please see our full budget proposal on the enclosed budget form. We will utilize existing computational resources within our department for initial data exploration and feature analysis. The computer science laboratory gives us access to approximately 40 computers, each containing 2–4 processor cores. While the computational power provided is sufficient, the machine learning frameworks we will use in this work are optimized to run on graphics processing unit (GPU)-accelerated systems.

To this end, we request funds to purchase a single NVIDIA TITAN X GPU—currently the fastest deep learning accelerator for the desktop computer market—enabling us to distribute our machine learning algorithm across over approximately 3600 GPU processing cores, significantly accelerating our work and allowing us to train our model on a much larger corpus of data. Please note that this is an optional, but

Item	Description	Amount ($)
1	Undergraduate research assistant stipend	3500
2	NVIDIA TITAN X 12GB GPU	1200
3	Conference registration and travel expenses	300
Total		5000

Table 1: Summary of proposed budget expenses.

highly recommended, one-time hardware expense. If authorized to purchase, the GPU will be housed within the department and made available to future teaching and research efforts at the university, including use during the department's CSIS 434 Parallel & Distributed Computing course, which is currently limited to CPU parallelization despite the industry's focus on GPU-accelerated parallelization in a wide range of applications.

References

[1] J. P. Bean and B. S. Metzner, "A conceptual model of nontraditional undergraduate student attrition," *Review of Educational Research*, vol. 55, no. 4, pp. 485–540, 1985. [Online]. Available: http://dx.doi.org/10.3102/00346543055004485

[2] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decis. Support Syst.*, vol. 49, no. 4, pp. 498–506, Nov. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2010.06.003

[3] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, Jun. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.artmed.2004.07.002

[4] D. Delen, R. Sharda, and P. Kumar, "Movie forecast guru: A web-based dss for hollywood managers," *Decis. Support Syst.*, vol. 43, no. 4, pp. 1151–1170, Aug. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2005.07.005

[5] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, vol. 30, no. 2, pp. 243 – 254, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417405001399

[6] X. Li, G. C. Nsofor, and L. Song, "A comparative analysis of predictive data mining techniques," *International Journal of Rapid Manufacturing*, vol. 1, no. 2, pp. 150–172, 2009. [Online]. Available: http://www.inderscienceonline.com/doi/abs/10.1504/IJRapidM.2009.02938

[7] M. Bogard, T. Helbig, G. Huff, and C. James, "A comparison of empirical models for predicting student retention." [Online]. Available: https://www.wku.edu/instres/documents/comparison_of_empirical_models.pdf

[8] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, Jul. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2006.04.005

[9] S. Herzog, "Estimating student retention and degree-completion time: Decision trees and neural networks vis--vis regression," *New Directions for Institutional Research*, vol. 2006, no. 131, pp. 17–33, 2006. [Online]. Available: http://dx.doi.org/10.1002/ir.185

[10] C. H. Yu, S. DiGangi, A. Jannasch-Pennell, and C. Kaprolet, "A data mining approach for identifying predictors of student retention from sophomore to junior year," *Journal of Data Science*, vol. 8, no. 2, pp. 307–325, 2010.

[11] J.-W. Jia and M. Mareboyana, *Predictive Models for Undergraduate Student Retention Using Machine Learning Algorithms.* Dordrecht: Springer Netherlands, 2014, pp. 315–329. [Online]. Available: http://dx.doi.org/10.1007/978-94-017-9115-1_24

[12] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

[13] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Vigas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," http://download.tensorflow.org/paper/whitepaper2015.pdf, 2015.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.