Paper 2632-2018

A Study of Modelling Approaches for Predicting Dropout in a Business College

Xuan Wang, Helmut Schneider, Louisiana State University.

ABSTRACT

Graduation rate is of interest for stakeholders of higher education including educational researchers and policymakers; and as widely acknowledged, retention rates are driving graduation rates. This study explores the use of predictive analytics in an academic institution for improving retention rates using data from the Louisiana State Business College as an example. The study applies SAS Enterprise Miner to build predictive models for identifying the students at risk of dropping out at different stages in their program. The institutional administrators can use the results from the models to identify students who need advising and remedial actions to help retain students and lead them to graduation. Preliminary findings from the study show that an ensemble model is the best for predicting student dropout. In addition, the predictive models can be further improved by collecting additional information about behavioral issues and study habits during the first year. Besides the practical implication, this study also shows the effectiveness of the analytics tools in improving graduation rates.

INTRODUCTION

The rise of personal computing in the 1980s and its acceleration through the Internet in the 1990s had a profound effect on the educational needs throughout the world. It not only created the demand for a technology skilled labor force but also led to a globalization of the economy allowing demand to be met anywhere in the world where highly skilled labor force is available and subsequently affected the choice of industrial locations. As a consequence of this development, government reporting and the funding mechanism for higher education went through a transformation from “complete input based systems to the adaption of more competitive outcome based approaches” (Alexander, 2000) and government interest in performance funding and budgeting for higher education has substantially increased in OECD nations (Brennan, 1999; Schmidtlein, 1999; Alexander, 2000). Political leaders in these countries have realized that to strengthen the competitiveness of their constituents they must increase their involvement in the development of human capital, specifically in higher education (Alexander, 2000). This economic motivation is energizing states to reassess their relationships with higher education, pressuring institutions to become more accountable, more efficient and more productive in the use of publicly generated resources (Alexander, 2000). Thus, accountability in college education has become a focal point of public debate (Elton, 1988; Keams, 1998; Dill, 1999; Alexander, 2000; Kitagawa, 2003; Huisman, 2004; Bailey, 2006). The objective of this study is to explore the use of predictive analytics in an academic institution with the aim of improving retention rates using data from a business college as an example. The remainder of the paper is organized as follows. Section 2 reviews the literature on the subject of graduation rates and dropout rates to provide the domain knowledge for our analytics case that is necessary in the application of analytics in order to avoid the misinterpretation of results. Section 3 uses predictive analytics for developing models that can be used for improving retention rates on an individual level. Section 4 closes with recommendations regarding the application of predictive models in academics to increase retention rates.

RETENTION RATES LITERATURE

Although graduation and retention rates have been the focus of researchers (Tinto, 1975; Cabrera, 1992; Braxton, 2003) for many years, student retention continues to be a difficult problem (Lau, 2003; Talbert, 2012). Some research indicates that retention and graduation rates are more accurately predicted at nationally selective schools than at less selective institutions (Schmitz, 1993). Resident status was also found to effect graduation rates and students taking a “Freshman Orientation Course” had a lower risk of dropping out (Murtaugh, 1999).

Although many campuses have focused on increasing retention and graduation rates largely because of external reasons (rankings, e.g., U.S. News & World Report), very few assessments of campus retention initiatives exist and evidence is thus scarce as to whether these initiatives are effective (Hossler, 2008). This is partly due to the slow adoption of advanced data management systems by colleges. However, in recent years, as new low cost analytics solutions have become available, there has been a growing interest in using analytics to gain better and timely insight into what drives student retention and to allow for the tracking of the effects of new initiatives (Pirani, 2005), specifically of high risk students (Talbert, 2012). The use of descriptive analytics has focused mainly on analyzing the admission process and pre-college factors and their impact on graduation rates and retention rates. However, predictive analytics can be used to identify students at risk of dropping out of college and can therefore allow for early corrective actions to be taken to increase student retention and subsequent graduation rates (Campbell, 2007). As new technologies are adopted and available data grow larger, more complex models can be used to monitor and predict student success. This research is focusing on applying the predictive models to identify the students at risk of dropping out of Business College, in order to advise the College counselors to take actions in time.

When using analytics for improving retention rates, it is important to distinguish between pre-college factors and in-college factors that affect graduation rates throughout college. While pre-college factors are used for selective admission, in-college factors are time dependent, measuring the student’s progress towards graduation. As a rule, college students fail classes, take courses out of the recommended sequence and change their curriculum. The ultimate role of analytics is to provide administrators with tools for corrective actions that can be taken to bring the student back on track to graduation at any time during the student’s life cycle at a college.

PRE-COLLEGE FACTORS

Pre-college factors include academic factors such as High School GPA and ACT assessment scores as well as non-academic factors such as socioeconomic status, self-confidence, achievement motivation, and academic goal orientation which attempt to measure personal traits. Academic factors have been shown to be important for college success and have been widely adopted for selective admission (Schnell, 2003), (DesJardins, 2003). Although research shows that besides academic factors, the socioeconomic status, self-confidence and motivation for achievement also play a large role (Lotkowski, 2004), but these factors are difficult if not impossible to evaluate during the admission process.

IN-COLLEGE FACTORS

In-college factors are significant keys for effecting the performance after the student enroll the college. In addition, the factors could be varied the students’ performance by time. Some of the in-college factors such as student persistence and family encouragement have received much attention and have been shown to play a significant role in student retention (Hossler, 2008). Research also indicates that student coaching in college and one-on-one tutoring elevates students’ retention rates. A randomized trial with 13,555 students in eight different institutions showed that student retention rates with coaching increased by 14 percentage points after a two year coaching period (Bettinger, 2011). Thus it becomes critical for colleges to identify students at risk who need coaching or tutoring while they can still be helped before it is too late.

First-year retention rates have received special attention because of the considerable adjustment from a high school to a college environment and is therefore critical for college success (Hossler, 2008). Some research has examined specific strategies to increase first-year retention. For example, first-year seminars have been shown to increase first-year retention rates (Schnell, 2003). Therefore, this study is developing both first semester and second semester predictive models for identifying the risky students.

PREDICTIVE ANALYTICS

Predictive analytics includes techniques that are used to make individual predictions. For instance, forecasting is a predictive method. The objective of forecasting is to make accurate predictions judged by the forecasting error without testing any true relationship between factors and the target variable. Several data mining methods were used to predict whether a student may dropout from the business college. As in many business applications, surrogate measures are needed that are timely performance indicators for the future result of a target variable. In this case study, student dropout with pre-college factors and in-college factors during each of the first two semesters are predicted.

There are many predictive algorithms available, some of which are included in the standard off-the-shelf statistical software packages. We do not intend to do an exhaustive discussion of the benefit of each method here. The purpose of this article is to discuss the modeling issues and provide an assessment of which method(s) is/are the best. In particular, SAS Enterprise Miner is used for building the predictive models by each semester with pre college factors and in college factors. The predictive methods in consideration are Decision Tree, Gradient Boosting, Neural Network, Regression and Ensemble Models. In business applications, selecting an appropriate method is one of the most important steps, and SAS Enterprise Miner also offers the option for assessing the method model fit.

A total of three models are studied on a data set of 9,913 students who entered a business college between 2006 and 2015. The first model uses only pre-college factors to predict the first semester dropout rate. The pre-college factors include ACT score, overall High school GPA, Gender, Greek Status, Distance from home to school, size of high school, high school type, campus status, pelican status and race. All of the variables are categorical, and ACT score and High school GPA are defined as ordinal variables, and other variables are defined as nominal variables. This model can be used as decision support for selective admission. It also allows administrators to design remedial measures to help those students to stay in the college. The second model includes first semester grades in addition to all pre-college factors to predict students who may not return for the second semester. In addition to the pre-college factors in the first model, in the second model, first year's curriculum is added as nominal variable. the grade of first semester's MATH1021, the grade of first semester's MATH1431, the grade of first semester's ISDS1102 and the grade of first semester' ECON2000 are added as ordinal variables. The third model uses all pre-college factors plus the grades received by the end of the second semester to predict whether a student may return for the third semester. In addition to the pre-college factors in the first model, in the third model, first year's curriculum is added as nominal variable. the grade of second semester's MATH1021, the grade of second semester's MATH1431, the grade of second semester's ISDS1102, the grade of second semester's ECON2000, the grade of second semester's ACCT2001 and the grade of second semester's ISDS2000 are added as ordinal variables. And this model only uses students who remain in the college during the second semester. Before we start building the predictive models, we split the data into a 75% training set and a 25% validate set. With the steps of model comparison, we set up validation misclassification rate as the criteria for comparison in SAS Enterprise Miner. Below are the diagram flows and comparison tables for each model.

Predictive analytics includes techniques that are used to make individual predictions. For instance, forecasting is a predictive method. The objective of forecasting is to make accurate predictions judged by the forecasting error without testing any true relationship between factors and the target variable. Several data mining methods were used to predict whether a student may dropout from the business college. As in many business applications, surrogate measures are needed that are timely performance indicators for the future result of a target variable. In this case study, student dropout with pre-college factors and in-college factors during each of the first two semesters are predicted.

There are many predictive algorithms available, some of which are included in the standard off-the-shelf statistical software packages. We do not intend to do an exhaustive discussion of the benefit of each method here. The purpose of this article is to discuss the modeling issues and provide an assessment of which method(s) is/are the best. In particular, SAS Enterprise Miner is used for building the predictive models by each semester with pre college factors and in college factors. The predictive methods in consideration are Decision Tree, Gradient Boosting, Neural Network, Regression and Ensemble Models. In business applications, selecting an appropriate method is one of the most important steps, and SAS Enterprise Miner also offers the option for assessing the method model fit. A total of three models are studied on a data set of 9,913 students who entered a business college between 2006 and 2015. The first model uses only pre-college factors to predict the first semester dropout rate. The pre-college factors include ACT score, overall High school GPA, Gender, Greek Status, Distance from home to school, size of high school, high school type, campus status, pelican status and race. All of the variables are categorical, and ACT score and High school GPA are defined as ordinal variables, and other variables are defined as nominal variables. This

model can be used as decision support for selective admission. It also allows administrators to design remedial measures to help those students to stay in the college. The second model includes first semester grades in addition to all pre-college factors to predict students who may not return for the second semester. In addition to the pre-college factors in the first model, in the second model, first year's curriculum is added as nominal variable. the grade of first semester's MATH1021, the grade of first semester's MATH1431, the grade of first semester's ISDS1102 and the grade of first semester' ECON2000 are added as ordinal variables. The third model uses all pre-college factors plus the grades received by the end of the second semester to predict whether a student may return for the third semester. In addition to the pre-college factors in the first model, in the third model, first year's curriculum is added as nominal variable. the grade of second semester's MATH1021, the grade of second semester's MATH1431, the grade of second semester's ISDS1102, the grade of second semester's ECON2000, the grade of second semester's ACCT2001 and the grade of second semester's ISDS2000 are added as ordinal variables. And this model only uses students who remain in the college during the second semester. Before we start building the predictive models, we split the data into a 75% training set and a 25% validate set. With the steps of model comparison, we set up validation misclassification rate as the criteria for comparison in SAS Enterprise Miner. Below are the diagram flows and comparison tables for each model.

Pre-College Model	Misclassification	ROC Index
Ensemble	**21.82%**	**0.57**
Gradient Boosting	21.82%	0.532
Decision Tree	21.82%	0.5
Logistic Regression	21.86%	0.568
Neural Networks	21.98%	0.564

Table 1. Comparison of Predictive Models – Pre-college factors vs. First semester dropout

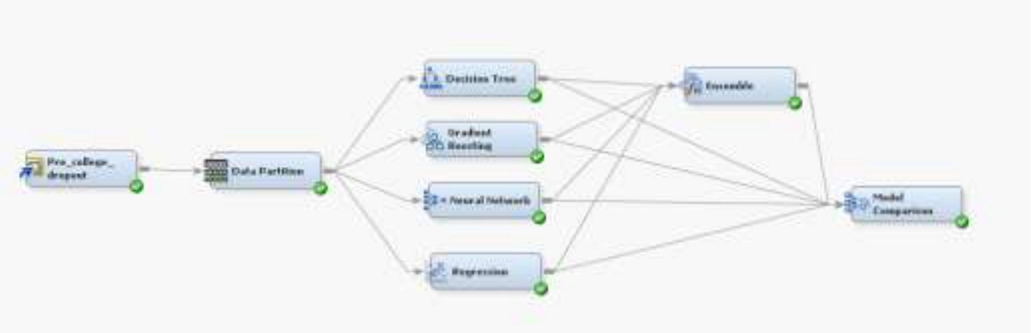

Figure 1. Diagram Flow in SAS Enterprise Miner – Pre-college factors vs. First Semester Dropout

Table 1 and Figure1 show the comparison results between all the methods for the predictive model of pre-college factors vs. first semester dropout. It clearly shows that Ensemble, Gradient Boosting and Decision Tree have the same misclassification rate. ROC index is equal to the probability that a classifier will rank a randomly chosen positive one higher than a randomly chosen negative one. And a perfect ROC index is 1. In addition, with the comparison of ROC index, Ensemble model will be the best choice because the value of ROC index is closest to 1. Meanwhile, SAS Enterprise Miner also indicates that the Ensemble model is the first choice among these models. Therefore, for the predictive model of pre-college factors vs. first semester dropout, we should choose ensemble model for predicting the future dropout after the first semester.

1st_semester_Model	Misclassification	ROC Index
Ensemble	**21.82%**	**0.716**
Gradient Boosting	21.82%	0.692
Decision Tree	21.82%	0.5
Logistic Regression	21.94%	0.711

Neural Networks	22.10%	0.709

Table 2. Comparison of Predictive Models – Pre-college & first semester factors vs. first semester dropout

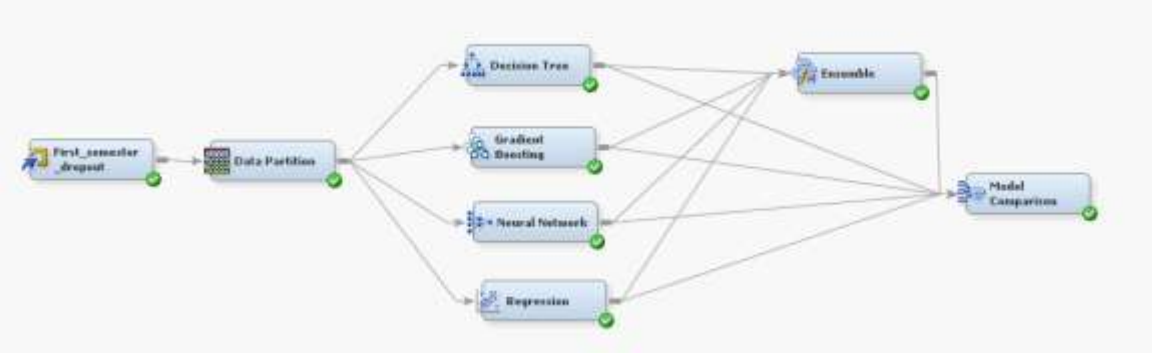

Figure 2. Diagram Flow in SAS Enterprise Miner – Pre-college & first semester factors vs. First Semester Dropout

Table 2 and Graph2 show the comparison results between all the methods for the predictive model of pre-college & first semester factors vs. first semester dropout. It shows that the Ensemble, Gradient Boosting and Decision Tree have the same misclassification rate as Table 1. In addition, with the comparison of ROC curve, the Ensemble model still is the best choice because the value of ROC index is closest to 1. Meanwhile, SAS Enterprise Miner also indicates that the Ensemble model is the first choice among these models. Therefore, for the predictive model of pre-college & first semester factors vs. first semester dropout, we should choose ensemble model for predicting the future observations. It is interesting to note that the addition of first-semester grades does not improve the misclassification rates, but increases the ROC index.

2nd_semester_Model	Misclassification	ROC Index
Ensemble	**16.21%**	**0.769**
Gradient Boosting	17.00%	0.761
Decision Tree	16.63%	0.668
Logistic Regression	16.00%	0.756
Neural Networks	16.54%	0.767

Table 3. Comparison of Predictive Models – Pre-college & second semester factors vs. second semester dropout

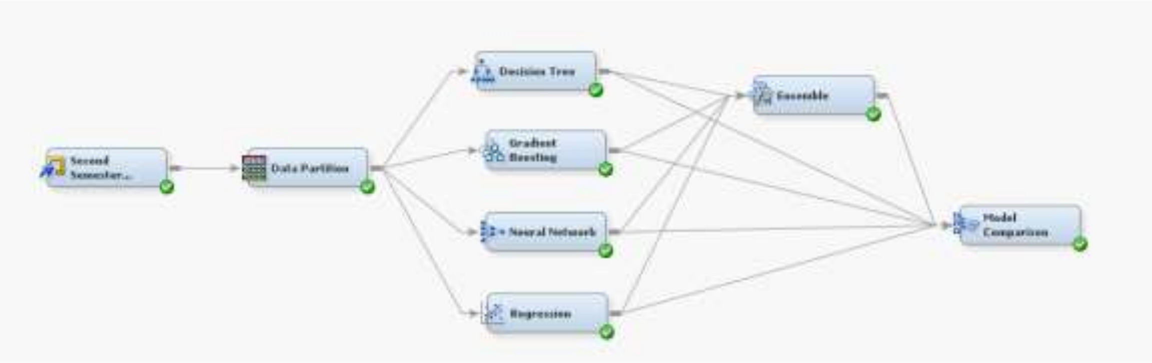

Figure 3. Diagram Flow in SAS Enterprise Miner – Pre-college & second semester factors vs. Second Semester Dropout

Table 3 and Graph 3 show the comparison results between all the methods for the predictive model of pre-college & second semester factors vs. second semester dropout. Logistic Regression holds the lowest misclassification rate. The Ensemble model holds the best ROC curve. Because of the criteria that we set in the step of model comparison, SAS Enterprise Miner indicates that the Logistic Regression model is the first choice among these models. However, the difference of misclassification rate between Logistic regression and Ensemble is only 0.2%. Therefore, for the predictive model of pre-college & second semester factors vs. second semester dropout, we could choose either ensemble or logistic regression for predicting the future observations.

			Cut-off value
Model	Dependent Variable	Metric	0.5
Pre-College*	1st Semester	Misclassification	21.82%
	Dropout	ROC Index	0.57
1st Semester*	1st Semester	Misclassification	21.82%
	Dropout	ROC Index	0.716
2nd Semester**	2nd Semester	Misclassification	16.21%
	Dropout	ROC Index	0.769

Table 4. Summary of Three Models

Table 4 shows the misclassification rate and ROC index for the three models. As the table shows, the model using only pre-college information is the poorest among all the three models. It has the highest misclassification rate and the lowest ROC index. Therefore, with the additional information about grades in the second semester we are able to build better models with lower misclassification rates.

CONCLUSION

Predictive models support the identification of first and second semester dropouts. The predictive model can be improved by collecting additional information about behavioral issues and study habits during the first year. Future research should include a survey of students to identify additional factors that increase the ROC index of the model and the accuracy of the predictions. Analytics has been proven successful in industry and it holds the promise of allowing university administrators to make more effective decisions regarding factors affecting retention rates specifically and university operations in general.

REFERENCES

ACT (2014) National Collegiate Retention and Persistence to Degree Rates, accessed, [available at http://www.act.org/research/policymakers/pdf/retain_2014.pdf].

Alexander, K. F. (2000). The Changing Face of Accountability: Monitoring and Assessing Institutional Performance in Higher Education. *The Journal of Higher Education*, 71(4), 411-431.

Astin, A. W. (1997). How "Good" is Your Institution's Retention Rate? *Research in Higher Education*, 38(6), 647-658.

Bailey, T. C., Juan Carlos; Jenkins, Davis; Leinbach, Tiimothy; Kienzl, Gregoy (2006). Is student right-to-know all you should know? An analysis of Community College graduation Rates. *Research in Higher Education*, 47(5), 491-519.

Bettinger, E., Rachel Baker (2011). The effects of students coaching in college: An evaluation of a randomized experiment in student mentoring. *NBER working paper No. 16881*.

Braxton, J. M., Amy S. Hirschy, Shederick A. McClendon (2003). Understanding and Reducing College Student Departure: ASHE-ERIC Higher Education Report.

Brennan, J. (1999). Evaluation of higher education in Europe. In M. L. Henkel, B. (Ed.), *Changing relationships between higher education and the state*. London: Athenaeum Press.

Cabrera, A. F., Maria B. Castaneda, Amaury Nora and Dennis Hengstler (1992). The convergence between two theories of college persistence. *The Journal of Higher Education*, 63(2), 143-164.

Campbell, J. P., Diana G. Oblinger (2007a). Academic Analytics. EDUCAUSE.

Campbell, J. P., Peter B. DeBiois, Diana Oblinger (2007b) Academic Analytics: A New Tool for a New Era, accessed, [available at http://www.educause.edu/ero/article/academic-analytics-new-tool-new-era].

DesJardins, S. L., Dong-Ok Kim, Chester S. Rzonca (2003). A nested analysis of factors affecting Bachelor's Degree completion. *Journal of college student retention*, 4(4).

Dill, D. D. (1999). Academic Accountability and university adaptation: The architecture of an academic learning organization. *Higher Education*, 38(2), 127-154.

Elton, L. (1988). Accountability in Higher Education: The danger of unintended consequences. *HIgher Education*, 17(4), 377-390.

Hossler, D., May Ziskin, John V. Moore III, Phoebe K. Wakhungu (2008). The role of institutional practices in college student persisitence. *2008 Annual Forum of the Association for Institutional Research*. Seattle, WA.

Huisman, J. C., Jan (2004). Accountability in higher education: Bridge over troubled water. *Higher Education*, 48(4), 529-551.

Keams, K. P. (1998). Institutional Accountability in Higher Education: A Strategic Approach. *Public Productivity & Manageent Review*, 22(2), 140-156.

Kitagawa, F. (2003). New Mechanisms of Incentives and Accountability for Higher Education Institutions Linking the Regional, National and Global Dimensions. *Higher Education Management and Policy*, 15.

Lau, L. K. (2003). Instituional Factors Affecting Student Retention. *Education*, 124(1), 126-136.

Murtaugh, P. A. B., Leslie D.; Schuster, Jill (1999). Predicting the Retention of University Students. *Research in Higher Education*, 40(3), 355-371.

Schmidtlein, F. (1999). Assumptions underlying performance-based budgeting. *Tertiary Education and Management*, 5, 159-174.

Schmitz, C. C. (1993). Assessing the Validity of Higher Education Indicators. *The Journal of Higher Education*, 64(5), 503-521.

Schnell, C. A., Karen Seashore Louis, Curt Doetkott (2003). The first-year seminar as a means of improving college graduation rates. *Journal of the First-Year Experience and Students in Transition*, 15(1), 53-75.

Shreve, J., Helmut Schneider and Omar Soysal (2011). A Methodology for Comparing Classification Methods through the Assessment of Model Stability and Validity in Variable Selection. *Decision Support Systems.*

Talbert, P. Y. (2012). Strategies to Increase Enrollment, Retention, and Graduation Rates. *Journal of Developmental Education*, 36(1), 22-24, 26-29, 31, 33, 36.

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, 45, 89-125.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Xuan Wang
PhD Candidate
The Stephenson Department of Entrepreneurship & Information Systems
E. J. Ourso College of Business
Louisiana State University
Email: xwang35@lsu.edu

Helmut Schneider, PhD
Ourso Family Distinguished Professor of Information Systems
The Stephenson Department of Entrepreneurship & Information Systems
E. J. Ourso College of Business
Louisiana State University
Email: hschnei@lsu.edu