SA23025

Predictive Analytics in Education: Considerations in Predicting versus Explaining College Student Retention

Kevin Paterson, DBA Student
University of the Incarnate Word

Adam Guerrero, Ph.D.
University of the Incarnate Word

Abstract

Data from a moderately-selective state university in the Midwest is used to cross-examine the most appropriate data analytical techniques for predicting versus explaining college student persistence decisions. The current research provides an overview of the relative benefits of models specializing in prediction versus explanation with particular emphasis on estimation methodologies, model specification by estimation technique, and model diagnostics, including classification tables and measurements of the goodness of fit. The predictive validity of a model of college student retention estimated using logistic regression is compared with that of discriminant analysis estimated using cognitive and non-cognitive predictors of retention. Key contributions to the literature include a unique analysis sample, a unique set of independent variables, and statistical estimation methodologies that build upon traditional frameworks, including machine learning techniques. The currents study ends with a discussion that will allow leaders in higher education and policymakers to make better data-informed decisions surrounding prediction and explanation so that they can proactively intervene with the most appropriate attrition-minimization policies.

Keywords: discriminant analysis, predicting student success, college retention, student persistence

INTRODUCTION

Most students who drop out of college do so during their first year of study. As of 2020, 2,278 colleges and universities, including public, nonprofit, and for-profit institutions, offered four-year degrees, thus providing prospective students with a wide range of options (NCES, 2022). Among these institutions, 718 were designated public institutions and competed for the same students as their private and for-profit counterparts. As compared to 2019, this represents a slight decrease from 2,230 and 730, respectively (NCES, 2022). The United States has also experienced a decline in the number of children under 18 years old, further intensifying the level of competition in higher education over the past decade (Ogunwole et al., 2021). The number of first-time freshmen (FTF) enrolled in 4-year public institutions reached over 1.21 million in 2019, with 74.8 percent remaining enrolled after one year (NCES, 2021). After just a year, public institutions lost just over 305 thousand FTF students, unchanged from 2018.

Due to these factors, improving retention rates of first-time freshmen (FTF) students remains an important research topic, as it provides a strategic competitive advantage to an institution of higher education. In addition, it has been shown that retention and graduation rates are ranking factors that distinguish colleges and influence where students apply for admission and how institutions make their selections (Sanoff et al., 2007). Further, higher retention and graduation rates are the outcomes of student persistence, which result in higher revenue generation for colleges (Tinto, 2006, 2017), making it possible for them to accomplish their respective missions. Given the importance of college student success to stakeholders in higher education, it is essential to understand the factors that drive admissions through to retention and graduation.

The current research compares the accuracy of predictions of persistence probabilities estimated using logistic regression to those obtained by discriminant analysis. Persistence probabilities by estimation technique are obtained for FTF students enrolled in classes at a moderately selective Midwestern public institution of higher education (MDHE, 2022). Retention model diagnostics including classification tables obtained using logistic regression are compared to those generated using a discriminant analysis based on cognitive and non-cognitive predictors. Finally, emphasis is placed on how leaders in higher education can use different data analytical techniques to approach the problem of prediction versus explanation of persistence decisions, especially considering the problem of omitted variable bias.

A REVIEW OF THE LITERATURE

College student retention is a multidisciplinary problem that has been researched from sociological, psychological, and economic perspectives for over fifty years (Tinto, 2006). For instance, Tinto's Interactionalist Theory and Bean's Industrial Model of Student Attrition are two of the most recognized sociological and psychological theories, respectively (Aljohani, 2016). A survey of the literature revealed that most studies focus on the explanation versus prediction of college student retention. Specifically, papers typically focus on the drivers of college student retention identified in the literature categorized as academic, social, psychological, economic, and student background variables (Astin, 1999; Bean, 1980, 1983; Bean & Eaton, 2000; Braxton & Hirschy, 2005; Cabrera et al., 1992, 1993; Kerkvliet & Nowell, 2005; Leppel, 2001; McCormick, 1997; Montmarquette et al., 2001; Paterson et al., (2022); Reason, 2003; Singell, 2004; Spady, 1970; Tinto, 1975, 1993).

However, studies that examine the predictive validity of models constructed using cognitive and non-cognitive factors incorporate machine learning techniques, including logistic regression and discriminant analysis, with an emphasis on model diagnostics (Williams et al., 2018). In addition to cognitive factors, such as math grades earned in college, non-cognitive factors including subject matter confidence have been used to predict persistence probabilities of first-year full-time engineering students using discriminant analysis (Burtner, 2005). Findings from the discriminant analysis at Prince George's Community College were 78.2% successful in predicting persistence (Hawley et al., 2005). The most likely to drop out were students who took several developmental courses or anticipated that English proficiency would be a problem during college.

METHODOLOGY

In this section, logistic regression and discriminant analysis are used to explore the predictive capabilities of models of college student persistence. Since most studies are limited in their ability to explain why students leave college because they do not include in analyses variables from all categories of determinants found in the literature, researchers should focus on prediction in addition to explanation, as discussed in Paterson et al. (2022). As shown in Table 1, in a population of 2,511 first-time, full-time freshmen enrolled during the 2017-18 and 2018-19 academic years, 2,352 will be analyzed. Approximately 78 international students were eliminated from the sample, and another 81 were removed due to missing data.

Table 1: 2017-2019 Tabulation of Student Participants

Group	Freq.	Percent
cohort_17-18	1,118	47.53%
cohort_18-19	1,234	52.47%
Total	2,352	100.00%

Logistic regression estimated using maximum likelihood estimation (MLE) is a machine learning technique that can be used to classify objects. In the context of college student persistence, objects are students classified into two categories, those who persist between their freshman and sophomore years and those who drop out. For analyses that contain a binary dependent variable, such as whether a student remains enrolled at a college or university, Tinto (1993) and Wetzel et al. (1999) recommend using logit regression analysis versus ordinary least squares. Discriminant analysis is also used to solve classification problems and understand the drivers of group membership (Finnegan, 2008; Isiaka, 2019). In contrast to logistic regression, discriminant analysis can be applied to multi-class classification problems.

RESULTS

The sample is comprised of 2,433 first-time freshmen enrolled at a four-year public university located in the Midwest. Of the 2,433 first-time, full-time freshmen students enrolled, 2,352 are included in the analysis sample. Approximately 81 students were lost due to listwise deletion of students missing data on variables included in the study, such as high school rank percentiles, ACT scores, or ACT subject scores. Variable categorization, descriptive statistics, and measurements for cognitive and non-cognitive predictors of retention are outlined in Table 2 below.

Table 2: Retention Predictors from Midwestern Institution Database Records

Variable and Category	Measurement	Mean	SD
Dependent Variable:			
drop	1 if a student dropped, 0 otherwise	0.223	0.416
High School Variables:			
HSGPA	High school grade point average	3.391	0.454
HSPercentile	High school percentile rank	65.644	21.775
HSRank	High school rank	111.851	139.641
HSSize	High school size	309.165	287.187
Standardized Test Variables:			
act	Student ACT composite scores	22.393	3.718
act2	ACT composite scores squared	515.256	173.384
ACTPercentile	ACT percentile rank	65.416	21.629
ACTENGL	ACT english score	22.273	4.838
ACTMATH	ACT math score	21.784	4.147
ACTReading	ACT reading score	23.468	4.888
ACTScience	ACT science score	22.809	3.703
College Variables:			
gpa	College grade point average	2.755	1.115
fgpa	College fall grade point average	2.983	0.849
sgpa	College spring grade point average	2.728	1.164
Additional Control Variable:			
Cohort	1 if 17-18 academic year, 0 otherwise	0.475	0.499

Of the 2,352 students observed, 525 did not persist the following year. This corresponds to a 22.3 percent drop rate, which closely aligns with the drop rate one would expect for public 4-year institutions with an admissions acceptance rate of 75 to 89.9 percent between 2017 and 2019 (NCES, 2021). Categories of variables include high school variables, standardized test variables, and college variables. The dependent variable is a dichotomous measure called “drop,” which equals one if the student does not remain at the same institution for the next year and zero otherwise.

Considering that the dropout rate is 22.3 percent (i.e., not 50 percent), this represents an imbalanced dataset, which necessitates determining the optimal cutoff. The logit model has a pseudo-R-squared of 0.3325. At all possible cutoffs, a receiver operating characteristic (ROC) curve was used to evaluate the model’s ability to classify positive and negative outcomes. In this case, the area under the ROC curve was 84.09 percent, which indicates that the logit model can discriminate between the two groups. As shown in Figure 1, the model's optimal cutoff of 0.174161 was determined by plotting the sensitivity and specificity against the probability cutoff.

Figure 1: Optimal Probability Cutoff for Logit Model

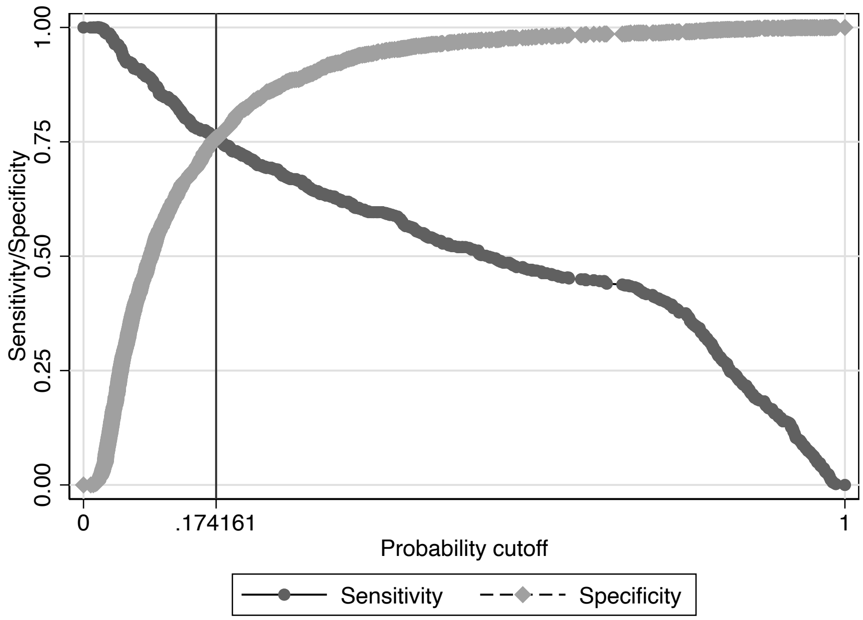


Table 3 illustrates the classification table generated by sequentially placing independent variables into the logit model. The probability threshold is assigned to a positive outcome set to the optimal cutoff, resulting in 75.89 percent of students being classified correctly. In addition, the pseudo-R-squared of 0.3325 indicates a strong fit.

Table 3: Logit Model for Drop

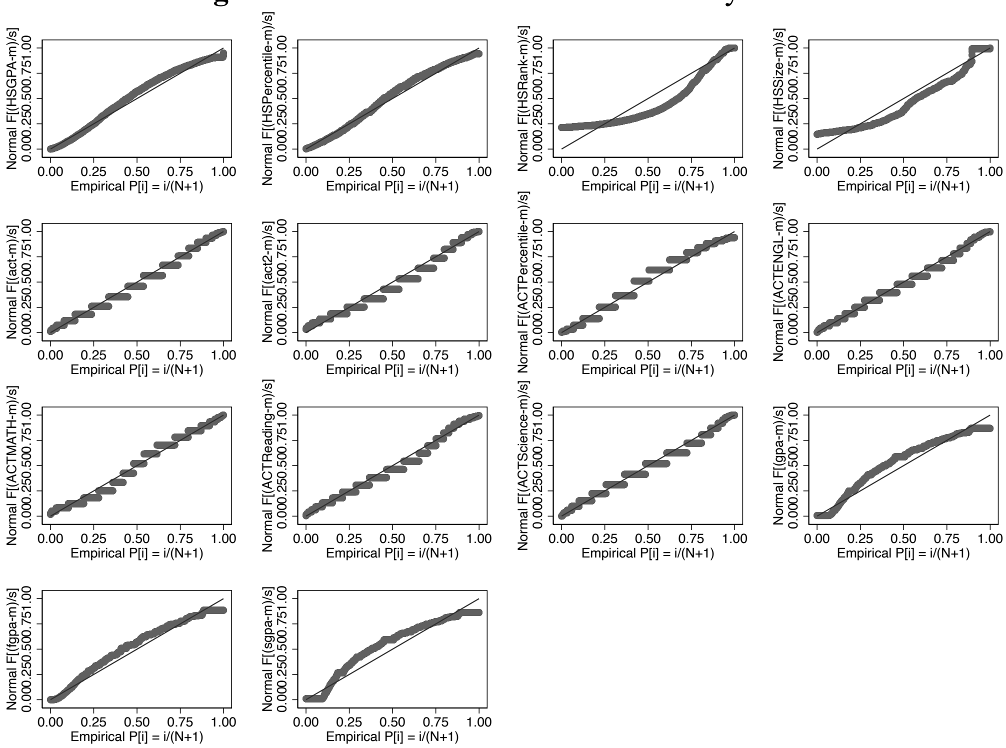
	True		
Classified	D	~D	Total
+	398	440	838
-	127	1387	1514
Total	525	1827	2352

Classified + if predicted Pr(D) >= .174161
True D defined as drop != 0

Sensitivity	Pr(+ D)	75.81%
Specificity	Pr(-~D)	75.92%
Positive predictive value	Pr(D +)	47.49%
Negative predictive value	Pr(~D -)	91.61%
False + rate for true ~D	Pr(+~D)	24.08%
False - rate for true D	Pr(- D)	24.19%
False + rate for classified +	Pr(~D +)	52.51%
False - rate for classified -	Pr(D -)	8.39%
Correctly classified		75.89%

Emphasis is placed on the diagnostic ability of predictive models constructed based on cognitive and non-cognitive factors to obtain predictions. The remaining interpretation will use the same unbalanced dataset, observations, and variables used in the logit model for discriminant analysis using prior proportional probabilities. The prerequisite assumptions for discriminant analysis are verified using histograms, normality probability plots, and multivariate tests for equal variances and covariances.

Figure 2: Standardized Normal Probability Plots



The canonical discriminant analysis examines the two-group discriminant model outlined in Table 4 below. As you can see, the discriminant function is statistically significant below the five percent level of statistical significance. A canonical correlation coefficient of 0.6063 indicates that the discriminant model accounts for approximately 36.76 percent of the total variance. Canonical group means indicate that students who drop out have a mean of -1.42, and those that do not drop out have a mean of 0.41. The results also indicate that sgpa, act2, act, gpa, and ACTPercentile significantly impact discriminant scores. College spring grade point average (spga) has the greatest impact on whether a student will persist, followed by ACT composite scores squared (act2), ACT composite scores (act), college grade point average (gpa), and ACT percentile rank (ACTPercentile).

Table 4: Canonical Linear Discriminant Analysis

Like-

Fcn	Canon. Corr.	Eigen-value	Variance Prop.	Variance Cumul.	lihood Ratio	F	df1	df2	Prob>F	
1	0.6063	0.581386	1.0000	1.0000	0.6324	97	14	2337	0.0000	e

H0: This and smaller canon. corr. are zero; e = exact F

Standardized canonical discriminant function coefficients

	function1
HSGPA	.0004564
HSPercentile	-.127374
HSRank	-.0180327
HSSize	.0341593
act	.7065039
act2	-.9275643
ACTPercentile	-.4726439
ACTENGL	.0065441
ACTMATH	.2518204
ACTReading	.2046847
ACTScience	.2343275
gpa	-.5685998
fgpa	.149135
sgpa	1.526032

Group means on canonical structure

drop	function1
0	.4085619
1	-1.421795

Classification accuracy increases to 85.76 percent using the discriminate model, with 52.19 percent of students who drop out classified correctly (see Table 5) despite an imbalanced dataset comprised of 22.32 percent of students who drop out. Sequentially adding variables to the discriminant model increased the positive predictive value, which was maximized when all variables were included. As a result of incorporating all variables in the discriminant model, the classification summary accuracy of 52.19 percent of students who drop out is maximized.

Table 5: Discriminant Model Classification Summary

True drop	Classified 0	1	Total
0	1,766	61	1,827

	96.66	3.34	100.00
1	251	274	525
	47.81	52.19	100.00
Total	2,017	335	2,352
	85.76	14.24	100.00
Priors	0.7768	0.2232	

The logit model has a positive predictive value of 47.49 percent and overall classification accuracy of 75.89 percent. On the other hand, the overall classification accuracy of the Discriminant Model is 85.79 percent, and the positive predictive value is 52.19 percent, which is higher than Logit Model, as indicated in Table 6.

Table 6: Model Classification Accuracy

	Logit Model	Discriminant Model
Correctly classified	75.89%	85.76%
Positive predictive value	47.49%	52.19%

CONCLUSSION AND DISCUSSION

The current research was able to show that predictive models of college student retention estimated using discriminate analysis can outperform the predictive capabilities of logit regression models. Not only does the discriminant function result in stronger overall classification accuracy, but it also provides a larger proportion of positive predicted values. The positive predictive value of the logit model was approximately 47.49%, whereas the discriminant model was able to surpass the 50% threshold with a value equal to 52.19%. Hence, it is important for higher education data analysts to build upon explanatory analyses of college student retention through the creation of models that specialize in prediction. Explanatory models of college student retention help inform attrition-minimization policies and retention maximization policies centered on variables that can be influenced by university leadership.

Predictive models of college student retention can be used to flag at-risk students before the point of departure, at which point additional data can be examined with an eye on student success. Predictive models can also be used to help inform data-driven decision-making focused on cost-benefit analyses based on the most accurate predicted probabilities. The current research can be extended if universities are able to compile datasets with additional cognitive and non-cognitive variables to build upon the discriminant model presented in the current study. It is also recommended that decision-makers in higher education explore the predictive capabilities of additional machine learning techniques including regularized logistic regression. This paper concludes that leaders should continue to conduct retention analyses using traditional models for explanation but should also explore other quantitative methods, including machine learning techniques such as discriminant analysis, to maximize the effectiveness of prediction.

REFERENCES

Aljohani, O. (2016). A Comprehensive Review of the Major Studies and Theoretical Models of Student Retention in Higher Education. *Higher Education Studies*, *6*(2), 1–18. https://doi.org/10.5539/hes.v6n2p1

Astin, A. W. (1999). Student involvement: A developmental theory for higher education. *Journal of College Student Developerment*, *40*(5), 518–529. https://primo-tc-na01.hosted.exlibrisgroup.com/permalink/f/5007c2/TN_cdi_proquest_journals_195180247

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, *12*, 155–187. https://link.springer.com/article/10.1007/BF00976194

Bean, J. P. (1983). The application of a model of turnover in work organizations to the student attrition process. *Review of Higher Education*, *6*, 127–148.

Bean, J. P., & Eaton, S. B. (2000). A Psychological Model of College Student Retention. In J. M. Braxton (Ed.), *Reworking the Student Departure Puzzle* (pp. 48–61). Vanderbilt University Press.

Braxton, J., & Hirschy, A. S. (2005). Theoretical Developments in the Study of College Student Departure. In A. Seidman (Ed.), *College Student Retention* (pp. 61–87). American Council on Education and Praiger Publishers.

Burtner, J. (2005). The Use of Discriminant Analysis to Investigate the Influence of Non-Cognitive Factors on Engineering School Persistence. *Journal of Engineering Education*. *94*(3): 335-338.

Cabrera, A. F., Castaneda, M. B., Nora, A., & Hengstler, D. (1992). The Convergence Between Two Theories of College Persistence. *The Journal of Higher Education*, *63*(2), 143–164. https://primo-tc-na01.hosted.exlibrisgroup.com/permalink/f/5007c2/TN_cdi_proquest_journals_205332444

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College Persistence: Structural Equations Modeling Test of an Integrated Model of Student Retention. *Source: The Journal of Higher Education*, *64*(2), 123–139. https://www-jstor-org.uiwtx.idm.oclc.org/stable/pdf/2960026.pdf?refreqid=excelsior%3Ab64485d8800ef38f8967fab08da28a83&ab_segments=&origin=

Finnigan, C., Morris, Libby., & Lee, K. (2008). Differences by Course Discipline on Student Behavior, Persistence, and Achievement in Online Courses of Undergraduate General Education. *Journal of College Student Retention*, *10*(1), 39-54.

Hawley, T. H., & Harris, T. A. (2005). Student Characteristics Related to Persistence for First-Year Community College. *Journal of College Student Retention.* *7*(1–2), 117–142.

Isiaka, R., Babatunde, R., Ajao, F., & Abdulsalam, S. (2019). A Machine Learning Approach to Dropout Early Warning System Modeling. *International Journal of Advanced Studies in Computers, Science and Engineering*. *8*(2), 1-12.

Kerkvliet, J., & Nowell, C. (2005). Does One Size Fit All? University Differences in the Influence of Wage, Financial Aid, and Integration on Student Retention. *Economics of Education Review*, *24*(1), 89–95.

Leppel, K. (2001). The Impact of Major on College Persistence among Freshmen. *Higher Education*, *41*(3), 327–342.

McCormick, K. (1997). An essay on the origin of the rational utility maximization hypothesis and a suggested modification. *Eastern Economic Journal*, *23*(1), 17–30.

MDHE. (2022). *Admissions Selectivity Categories*. DHEWD Policies & Guidelines. https://dhewd.mo.gov/policies/admissions-selectivity.php

Montmarquette, C., Mahseredjian, S., & Houle, R. (2001). The Determinants of University Dropouts: A Bivariate Probability Model with Sample Selection. *Economics of Education Review*, *20*(5), 475–484.

NCES. (2022). *COE - Characteristics of Degree-Granting Postsecondary Institutions*. Annual Reports and Information Staff (Annual Reports). https://nces.ed.gov/programs/coe/pdf/2022/csa_508.pdf

NCES. (2021). *Retention of first-time degree-seeking undergraduates at degree-granting postsecondary institutions, by attendance status, level and control of institution, and percentage of applications accepted: Selected years, 2006 through 2020*. Digest of Education Statistics. https://nces.ed.gov/programs/digest/d21/tables/dt21_326.30.asp

Ogunwole, S. U., Rabe, M. A., Roberts, A. W., & Caplan, Z. (2021). *U.S. Adult Population Grew Faster Than Nation's Total Population From 2010 to 2020*. United States Census Bereau. https://www.census.gov/library/stories/2021/08/united-states-adult-population-grew-faster-than-nations-total-population-from-2010-to-2020.html

Paterson, K., & Guerrero A. (2022), Factors Related to College Success: The Case of a State University in the Midwest. *Proceedings of the Academic and Business Research Institute, USA.* http://www.aabri.com/Virtual21F/F22VC002.html

Reason, R. D. (2003). Student Variables that Predict Retention: Recent Research and New Developments. *NASPA Journal*, *40*(4).

Sanoff, A. P., Usher, A., Savino, M., & Clarke, M. (2007). College and University Ranking Systems. *Institute for Higher Education Policy*, 1–53. https://files.eric.ed.gov/fulltext/ED497028.pdf

Singell, L. D. (2004). Come and stay a while: does financial aid effect retention conditioned on enrollment at a large public university? *Economics of Education Review*, *23*(5), 459–471.

Spady, W. G. (1970). Dropouts from Higher Education: An Interdisciplinary Review and Synthesis. *Interchange*, 64–85.

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition*. The University of Chicago Press.

Tinto, V. (2006). Research and Practice of Student Retention: What Next? *Journal of College Student Retention : Research, Theory & Practice*, *8*(1), 1–19. https://doi.org/10.2190/4YNU-4TMB-22DJ-AN4W

Tinto, V. (2017). Reflections on Student Persistence. *Student Success*, *8*(2), 1–8. https://doi.org/10.5204/ssj.v8i2.376

Wetzel, J. N., O'toole, D., & Peterson, S. (1999). Factors Affecting Student Retention Probabilities: A Case Study. *Journal of Economics and Finance*, *23*(1), 45–55.

Williams, R., Smiley, E., Davis, Ronnie., & Lamb, T. (2018). The Predictability of Cognitive and Non-cognitive Factors on the Retention Rate among Freshmen College Students. *The Journal of Negro Education*, *87*(3), 326-338.