Economics of Student Retention Behavior in Higher Education

Authors:

(1) **Rezwanul Parvez**
Research Economist
Colorado State University
Email: rezwanul.parvez@colostate.edu

(2) **Syed Imran Ali Meerza**
Assistant Professor
Arkansas Tech University
Email: smeerza@atu.edu

(3) **Nazea H. Khan Chowdhury**
Faculty of Social and Behavioral Science Program
Front Range Community College
Email: nazea.khanchowdhury@frontrange.edu

Selected Paper prepared for presentation for the Agricultural and Applied Economics Association (AAEA) Meeting, 2020

Abstract

The main objective of this study is to develop and empirically test a comprehensive list of factors affecting student retention behavior at an institutional level. We collect available student data from a 4-year flagship institution at Colorado to build a data mining model that can assess student retention behavior. Both econometric and machine learning methods are employed to determine the factors affecting student retention. Empirical results indicate that grade point average (GPA), institution's primary campus, first-generation students, age, and academic advisor are significant factors in explaining student retention behavior.

Keywords: Student retention, econometric method, machine learning methods

1. Introduction

The concept of retention and graduation rates, as identified by the "Higher Education Act" is well-recognized as key measures to assess institutional effectiveness. Postsecondary institutions allocate money and resources to explore the factors affecting student retention behavior and how to better understand dropout rates and student's unwillingness to return to the same institution. From an institution's perspective, the retention rate is a measurement (expressed as a percentage) at which students persist in their educational program. In other words, retention rates measure the percentage of first-time undergraduate students who return to the same institution in the following fall. The economics of education literature indicates a downward slope trend line of college retention and complete graduation across the United States since the early '90s. Existing literature indicates that the rate of retention does vary across institutions (62% to 96% at public 4-year institutions while 54% to 81% at 4-year private institutions) (NCES, 2019). The trend of retention is even more alarming in the case of 2-year degree-granting institutions, including community colleges (62% at public 2-year and 67% at private 2-year) (Hussar & Bailey, 2019; NCES, 2019). This declining trend of graduation and retention has also been broadly discussed in the literature during 1991-2012 (Allen

& Bir, 2012). These trends of persistence across institutions are not promising for the economy, raising concerns for the public interest, and reflect a significant loss of time and monetary capital among proponents of achieving a better student persistence rate at all (individual, state, and national) levels. Thus, academic communities, including higher learning institutions, continue to confront a myriad of serious challenges on how to identify metrics and implement effective measures to improve student retention (Hone & Said, 2016). Besides, there are only a handful of studies that depicted intervention programs to mitigate student dropout rates (Seidman, 2005). As mentioned in the literature, the existence of low retention rates is still on the rise across academia (Yu, DiGangi, Jannasch-Pennell, & Kaprolet, 2010; Scott & Hoover, 2014; Kimbark et al. 2017; Windham et al. 2014). As a result, a comprehensive assessment of exploring the factors affecting student retention behavior is worth pursuing.

Retention rate is considered as the topmost contributing factor of assessing an institution's effectiveness and financial stability. An institution with a better retention rate indicates quality programs, better student engagement, and enhancement of student success. From the public policymakers' perspective, the retention rate is a key measure to evaluate institutions' accountability and institutional effectiveness (Fike & Fike, 2008). Also, academic institutions have an obligation to ensure students' successful college experience by providing quality instructions and better student engagement. Surprisingly, majority institutions tend to consider retention as a subject of lesser importance (Tinto, 1999). Each academic institution has different management practices in terms of addressing student retention behavior. Literature indicates a higher portion of incoming students entering community colleges (CC) are underprepared to make a successful transition and are continued to face challenges in college transition (Lu, 1994). This is due to the fact that student enrollment is limited at 4-year institutions due to hard-core admission standards and required test scores. Most CC ended up enrolling a higher number of underprepared students due to its' open-door policy. Despite this, CC plays a vital role in its local economy as colleges provide services to its students by successfully transferring to a 4-year college or complete their degrees or certificates (Bahr, 2018). Additionally, CC provides unique support services (e.g., skill builders, developmental education, and remedial courses) to enhance student success. Despite having all these, there are still growing concerns on how to improve student retention and the rate of graduation across institutions.

Understanding the theory behind a student's decision to stay or leave an academic institution is critical. According to Tinto (1993), student development theory (student integration model) defines how a student makes progress through various stages. Students' success can be influenced by multiple factors (e.g., group study, social and academic integration, family background, abilities, and former schooling) of whether a student is going to complete their program of study. Tinto's model presents a relationship between pre-college entry factors, institutional characteristics, integration with goals and outcomes, and institutional incidents (e.g., faculty interactions, activities on campus, peer interactions) (Tinto, 1987).

Another theory (psychological model of retention) suggests environmental variables and student's intention as drivers of retention (Bean, 1990; Hossler et al. 1990). Additionally, Astin (1985, 1991, 1993) is well mentioned in the literature for his "student involvement theory (SIT)" and "input-environment-outcome model (IEO)" theories explaining retention. These theories include macro-level data across hundreds of academic institutions. The SIT theory suggests the dependence of student retention on student demographics (e.g., age, race, gender), and institutional characteristics (e.g., location, size, level of academic and social involvement). However, the IEO theory focuses on predicting retention using relationships between outputs (e.g., degrees earned, number of graduates) and inputs (e.g., gender, age, student ability, primary areas of study). Besides this, Bean's (1980) model of student attrition highlights the fact that student satisfaction is derived from services offered at the institution (e.g., grades, values, peer support, rewards) (Ishitani & DesJardins, 2002). The necessity of including key institutional data points (e.g., registered credits, campus location, award recipients, academic performance, parental education, student demographics) along with output variables is a must to build a robust model of student retention. Further, Existing literature mentioned a vast majority of first-generation students' presence across institutions (Thayer, 2000). Literature also indicated parental education as a key driver of student retention as first-generation students are less likely to persist. Moreover, a student from low-income families is less likely to succeed as compared to students from high-income families (Fike & Fike, 2008). So, the inclusion of financial aid variables to a retention model can help better predict retention behavior.

Post-secondary institutions are still dealing with cross-disciplinary challenges to identify and implement effective measures to increase retention rate (Hone & Said, 2016; Lau, 2003). Also, the value of student success for building a strong economy is well understood by the Federal and state governments (FSG). As a result, FSG offers multiple programs to all layers of students across higher academic institutions to enhance student success. Besides, academic institutions have designed student interventions and support programs to mitigate dropout rates (Seidman, 2005). Existing literature has made attempts to identify metrics driving student retention using empirical methods (Reason, 2009; Scott & Richard, 2014; Slanger et al. 2015; Raju & Schumacker, 2014-2015; DeBerrad et al. 2006). However, existing retention models investigated retention behavior without including the importance score of predictor variables. For example, Windham et al. (2014) adopted a post-facto quasi-experimental method to assess factors (ethnicity, age, gender, socio economic status, ACT score, and skills course participation) affecting student retention at an academic institution. Also, a chi-square test of independence model and logit model was employed to measure success course effectiveness on student performance (Kimbark et al. 2017). Additionally, Barbatis (2010) adopted a constant comparative method to examine the effectiveness of student engagement on student persistence and retention at a large urban academic institution. Thus, it will be worthwhile to examine student retention behavior by including additional key indicator variables, compare results, and whether the findings are supportive across similar institutions.

This study attempts to contribute to the literature by assessing factors, including the importance score of key predictor variables that are predictive of college student retention at a 4-year institution. This inclusion of key factors and proposed machine learning techniques might bring better ways to understand the shifting trends of retention and accurately predict student retention. This methodology of selecting predictor variables to predict retention is also consistent with existing retention theories. Key findings reported here will provide institutions' personnel with actionable information and best management practices to assess retention predictors. Additionally, using a combination of data (e.g., student engagement, student performance, and student demographics) for retention prediction will help the institution adopt targeted intervention measures to establish knowledge-based and data valued decision making. Similar institutions across higher academia can share findings reported here to define factors affecting retention and allocate resources efficiently across the institution. The structure of this research paper is organized as follows. Section one presents the introduction, while Section two presents the sample data and variables. Section three describes the empirical models, and Section four discusses the empirical results. The last section concludes the paper.

2. Data and Variable

We collect available student data from a 4-year flagship institution at Colorado to build a data mining model that can assess student retention behavior. We identify and explore the key factors associated with student retention. This study relies on a cross-sectional dataset covering the academic year 2019-2020. The dataset includes both first-time full time/part-time degree-seeking students. The primary data sources that have been used to conduct this research are the national center for education statistics (NCES), student unit record data system (SURDS), and integrated postsecondary education data system (IPEDS). The dependent variable includes student retention behavior (See Table 1).

Table 1. Description of Model Variables

Variables	Descriptions	Units
retained	whether a student is retained or not in the following year (fiscal year to fiscal year)	1=yes, 0=otherwise Dummy
acad_advisor	student met an academic advisor in their first semester	Y = 1; N=0
trio_advisor	enrolled in student support services program	Y =1; N=0
primary_camp	student taking classes only at primary campus	
first_gen	(first generation student or not)	1=yes, 0=otherwise
gpa_cal	first semester GPA	Number
age_cal	age at college entrance	Years
current_age	students' current age	Years
race_white	students' race	1= white 0=Non-white
parenteduc	parental education	1=mother, 2=farther, 3=both, 4=Neither
academicachievement	whether a student is member of honors society	1=yes, 0=otherwise
total_credits	total registered credits	Number
credits_earned	total earned credit hours	Number
student_status	active or graduating	1=active, 2=graduating
primary_program_ind	student declares a primary program	Y=1; N=0
gender	students' gender	Female = 1, otherwise = 0
residency_ind	students' residency status	1=state resident, otherwise= 0
fin_aid_applicant_ind	use of financial aid	1=Yes; N=0

Explanatory or predictor variables include student demographics (e.g., age, race, first-generation, parental education, residency status, and gender), institution characteristics (e.g., campus location, academic advisor, and primary program indicator), academic achievement, student status, registered credit hours, earned credit hours, grade point average (GPA), enrolled in student support services program (trio advisor), and use of financial aid (see Table 1). We hypothesize the above-mentioned independent variables may influence student retention behavior. All model variables are explained and reported in Table 1.

3. Empirical Estimation Strategies

This study uses both econometric method (i.e., Logit model) and machine learning approach (i.e., random forest model) to predict student retention behavior. A binomial logistic (BL) model can be specified as:

$$\ln \frac{\mathrm{Pi}}{1-\mathrm{Pi}} = \sum \beta_i X_i \qquad (1)$$

Equation (1) can be restated as

$$L_i = \ln \frac{\mathrm{Pi}}{1-\mathrm{Pi}} = Z_i = \beta_1 + \beta_2 X_i + u_i \qquad (2)$$

where Pi = the probability (students chose to stay or leave) that an event occurs for an observed set of variables X_i; β_i = the coefficients to be estimated; and X_i= all explanatory variables of the model and u_i is the stochastic error term. L_i, the log of the odds ratio is not only linear in X, but also linear in the parameters. L_i is called the logit, and hence equation (2) is the logit model. The initial specifications of the logit models include considerably greater number of predictor variables. However, numerous combinations of variables are being tested with the model before settling to a final test.

This study also employs random forest (RF) model (Breiman, 2001), a machine learning method, to check the robustness of the results estimated by the earlier model. As Meerza et al. (2019) summarizes that while binomial logistics regression outperforms the RF method when the data are relatively balanced between classes (Muchlinski et al. 2015), the RF method performs well in the presence of non-linear features and complex interactions among predictor variables (Culter et al. 2007).

The random forest method splits data into training and validation subgroups. In this study, 70 percent of our observations were used to train the model, while the rest 30 percent were used for validation. After splitting data into two subsets, the random forest method performs the following steps (Zhang & Ma, 2012):

1. Draw a bootstrap sample of size N
2. Select randomly k number of predictors out of p available predictors and determine the best binary splits out of all binary splits on k predictors
3. Utilizing the split from step 2, split the node into two descendant nodes
4. Repeat steps 2 and 3 n number of times to create n number of trees
5. Make prediction at point x_i as $\hat{f}(\, x_i) = \arg max_y \sum_{j=1}^{J} I(\widehat{h_j}(x_i) = y)$ for classification

The random forest method also measures the importance of predictor variables in predicting the response variable. For each observation, this method gives two types of out-of-bag predictions: (1) obtained from real data, and (2) obtained from variable k permuted data. To determine the importance of predictor variables, the difference between error rates of predictions are computed for each observation and averaged over all observations (Zhang & Ma, 2012).

4. Results and Discussion

Table 2 shows the descriptive statistics of model variables. According to Table 2, the mean college entrance age of students is 24.02, whereas the current age is 24.43. The mean GPA of students is 2.20, with a standard deviation of 1.47. On average, a student registered 8.00 credits while attending college; however, the student's mean earned credit is 1.58. Additionally, a vast majority of students' is resident in the state of Colorado (mean value of 0.89) while attending college. In other words, only

11 percent of students are out of state students. The total number of female students is higher as compared to male at this institution and are active students. Majority students have registered for classes at primary campus and met with an academic advisor in their first year. Finally, the majority of students declare their primary major while attending the institution and are also eligible to receive financial aid.

Table 2. Descriptive statistics

Variables	Mean	S.D.*
gpa_cal	2.20	1.47
age_cal	24.02	8.07
current_age	24.43	8.09
race_white	0.32	0.47
total_credits	7.99	3.83
credits_earned	1.58	1.76
acad_advisor	0.75	0.43
trio_advisor	0.00	0.09
primary_camp	0.56	0.50
first_gen	0.62	0.49
parenteduc	3.41	1.03
academicachivement	0.05	0.22
student_status	1.03	0.17
gender	0.59	0.49
residency_ind	0.89	0.31
primary_program_ind	0.98	0.14
fin_aid_applicant_ind	0.89	0.31

*S.D. refers to Standard deviation

Running the binominal logistic regression model (using STATA), we obtain p-values for each explanatory variable, and we find that majority of explanatory variables are statistically significant. So, there's evidence that each of these has an independent effect on the probability of a student being retained (rather than just a difference observed due to chance). Key regression results (logistic) indicates that "academic advisor" has a positive and statistically significant impact on student retention behavior. In other words, if a student meets an advisor during their 1st year at the institution, he/she is about 25 percent more likely to come back for the following year as compared to students who didn't meet an advisor during 1st year at the institution. Moreover, students who enrolled in the student support services program were approximately 58 percent more likely to come back for the following year. Students who are taking classes in "primary campus" tend to stay longer with the institution as compared to students who are taking classes in "non-primary campuses." Intriguingly, first-generation students are 7 percent more likely to come back for the following year as compared to those who are not first-generation students. Students with academic achievements (e.g., received awards while attending the institution) are around 30 percent more likely to retain for the following academic year. Further, empirical results reveal that while student's resident tuition eligibility and academic standing (e.g., GPA) are positively correlated with student retention, some demographic characteristics, such as gender, race, and parental education, have a negative impact on student retention.

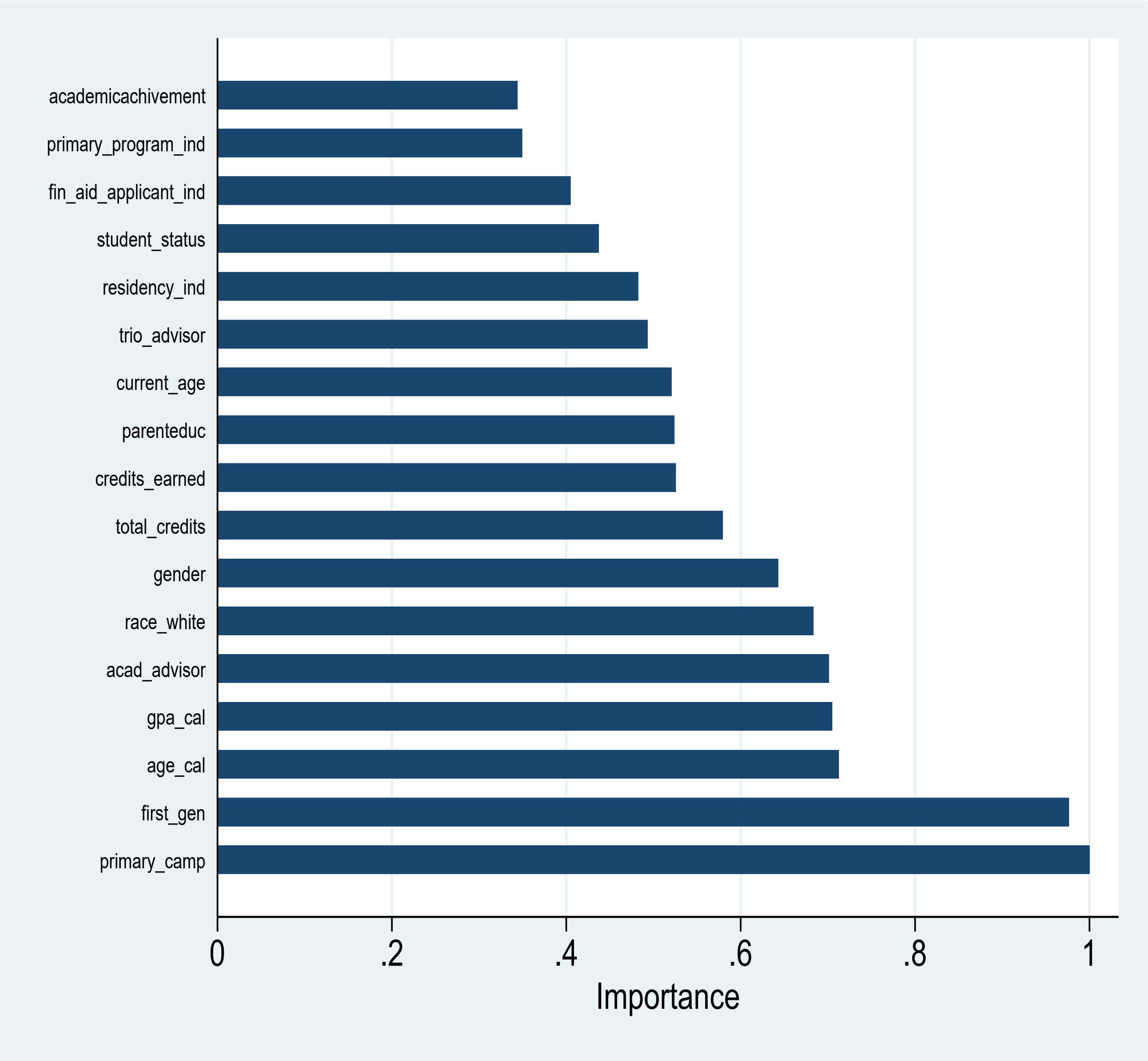
Table 3. Effects of predictor variables on student retention behavior (dependent variable = student retention)

Explanatory variables	Marginal Effects
acad_advisor	0.247*** (0.012)
trio_advisor	0.584*** (0.074)
primary_camp	0.40*** (0.009)
first_gen	0.066*** (0.013)
gpa_cal	0.113*** (0.003)
age_cal	-0.018** (0.008)
current_age	0.017** (0.008)
race_white	-0.071*** (0.009)
parenteduc	-0.014** (0.006)
academicachievement	0.298*** (0.024)
total_credits	-0.002* (0.001)
credits_earned	0.002 (0.002)
student_status	0.041 (0.027)
primary_program_ind	-0.141*** (0.031)
gender	-0.026*** (0.008)
residency_ind	0.059*** (0.015)
fin_aid_applicant_ind	0.019 (0.015)
Log-likelihood	-6285.90
No. of observations	11,456

Note: Reported values are the estimated marginal effects and, in parentheses, standard errors. *** significant at 1%, ** significant at 5%, * significant at 10%.

As mentioned earlier, when non-linear characteristics and complex interactions among predictor variables exist, the random forest method performs better than the logit model (Culter et al. 2007). The overall accuracy of predictor variables in predicting student retention behavior is 78 percent. The top five important factors in predicting student retention behavior are the institution's primary campus, first-generation students, age, GPA, and academic advisor. Interestingly, comparing with the rest of the predictor variables, unlike the logit model, student support services program, academic achievement, and residency tuition eligibility are less important variables in predicting student retention behavior.

Figure 1. Importance score of predictor variables



5. Conclusion

This study employed the RF method to predict student retention behavior. Our model can successfully predict 78 percent of student retention behavior. This study also focused on identifying the factors affecting student retention behavior by utilizing both the binomial logit and random forest models. Logit regression results reveal that academic advisor, student support service program, institution's primary campus, and students' academic achievements play a significant role in retaining students. Intriguingly, while the RF method also finds academic advisor and institution's primary campus are important predictors, it shows, unlike logit regression results, that both student support service programs and students' academic achievements are insignificant in predicting student retention behavior.

A critical caveat of this study is that the findings reported throughout the study are specific to the institution. However, key findings reported here can be applicable to a greater audience of higher education researchers, stakeholders, academic institutions, and policymakers. This article provides an initial attempt to introduce a machine learning approach to predict student retention behavior. Further research on student retention behavior using different datasets and machine learning approaches will help develop a better predicting model and more complete understanding of factors affecting student retention behavior.

References

Allen, D. F., & Bir, B. (2011-2012). Academic confidence and summer bridge learning communities: Path analytic linkages to student persistence. Journal of College Student Retention: Research, Theory & Practice, 13(4), 519-548.

Astin, A. W. (1991). Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education. New York: Macmillan.

Astin, A. W. (1993). What matters in college? San Francisco: Jossey-Bass.

Astin, A. W. (1985). Achieving educational excellence. San Francisco, CA: Jossey-Bass Inc.

Bahr, P. R., & Chen, Y. (2018). Skills builders in Colorado community colleges. Ann Arbor, Michigan: Center for the Study of Higher and Postsecondary Education, University of Michigan.

Barbatis, P. (2010). Underprepared, Ethnically Diverse Community College Students: Factors Contributing to Persistence. Journal of Developmental Education, 33(3), 16-26.

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.

Bean, J. P. (1990). Why students leave: Insights from research. In D. Hossler, J. P. Bean, & Associates (Eds.), The strategic management of college enrollments San Francisco: Jossey-Bass, 170-185.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J. (2007). Random forests for classification in Ecology. Ecology 88(11), 2783-2792.

DeBerrad, M. S., Spielmans, G. I., & Julka, D. C. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. College Student Journal, 38(1), 66-80.

Fike, D., & Fike, R. (2008). Predictors of first-year student retention in the community college. Community College Review, 36(2), 68-88.

Hone, Kate and Ghada, Said. (2016). Exploring the factors affecting MOOC retention: A survey study. Journal of Computers & Education, 98, 157-168.

Hossler, D., Bean, J. P., and Associates (1990). The Strategic Management of College Enrollments. San Francisco: Jossey-Bass.

Hussar, W.J. & Bailey, T.M. (2016). Projections of education statistics to 2024. U.S. Government Printing Office, Washington, DC.

Ishitani, T. T., & DesJardins, S. L. (2002). A longitudinal investigation of dropouts from college in the United States. Journal of College Student Retention, 4(2), 173-202.

Kimbark, Kris., Peters, Michelle L., Richardson, Tim. (2017). Effectiveness of the Student Success Course on Persistence, Retention, Academic Achievement, and Student Engagement. Community College Journal of Research and Practice, 41(2), 124-138.

Lau, L. K. (2003). Institutional factors affecting student retention. Education, 124(1), 126-136.

Lu, L. (1994). University transition: Major and minor stressors, personality characteristics and mental health. Psychological Medicine, 24(1), 81-87.

Meerza, S.I.A.; Yiannaka, A.; Brooks, K.R.; Gustafson, C.R. Information avoidance behavior: Does ignorance keep us uninformed? In Proceedings of the Annual Meeting—Agricultural and Applied Economics Association, Atlanta, Georgia, 21–23 July 2019. Available at https://ageconsearch.umn.edu/record/290757

Muchilnski, D., D. Siroky., J. He., and M. Kocher. (2015). “Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data,” Political Analysis, 24 (1), 87-103.

National Center for Education Statistics. (2019). Undergraduate retention and graduation rates. Available at https://nces.ed.gov/programs/coe/indicator_ctr.asp

Raju, D. & Schumacker, R. (2014-2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. Journal of College Student Retention, 16(4), 563-591.

Reason, R. D. (2009). An examination of persistence research through the lens of a comprehensive conceptual framework. Journal of College Student Development, 50(6), 659-682.

Scott, J. Mertes. & Richard, E. Hoover. (2014). Predictors of First-Year Retention in a Community College. Community College Journal of Research and Practice, 38(7), 651-660.

Seidman, A. (2005). College student retention: Formula for student success. Westport, CT: PraegerSlanger,

W. D., Berg, E. A., Fisk, P. S., & Hanson, M. G. (2015). A longitudinal cohort study of student motivational factors related to academic success and retention using the college student inventory. Journal of College Student Retention: Research, Theory and Practice, 17(3), 278-302.

Thayer, P. (2000). Retention of students from first generation and low income backgrounds. Opportunity Outlook Journal, 3(1), 2-8.

Tinto, V. (1982). Limits of theory and practice in student attrition. Journal of Higher Education, 53, 687-700.

Tinto, V. (1987). Leaving college: Rethinking the causes and the cures of student attrition. Chicago, IL: The University of Chicago Press.

Tinto, V. (1993). Leaving college: Rethinking the causes and cures of student attrition (2nd ed.). Chicago, IL: University of Chicago Press.

Tinto, V. (1999). Taking retention seriously: Rethinking the first year of college. NACADA Journal, 19(2), 5-9.

Windham, M. H., Rehfuss, M. C., Williams, C. R., Pugh, J. V., & Tincher-Ladner, L. (2014). Retention of first-year community college students. Community College Journal of Research and Practice, 38(5), 466-477.

Yu, C., S. DiGangi, A. Jannasch-Pennell, C. Kaprolet. (2010). A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. Journal of Data Science, Vol. 8, 307-325.

Zhang, C. & Ma, Y. (2012). Ensemble machine learning: Methods and Applications. Springer, Springer New York Dordrecht Heidelberg London.