The AIR Professional File

Winter 2025 Volume

Supporting quality data and decisions for higher education.

air | ASSOCIATION FOR INSTITUTIONAL RESEARCH

PREFACE

Degree completion among traditional first-time freshmen has long been a central focus in postsecondary education research. However, as higher education continues to evolve, there is increasing attention on nontraditional learners, such as those returning to school after workforce experience or starting their academic journey at community colleges. The rising demand for advanced skills in the workforce has led more professionals to pursue graduate degrees. The Winter 2025 issue of the ***AIR Professional File*** features articles exploring degree completion among transfer and graduate students.

Fikrewold Bitew and Lauren Apgar investigate the factors influencing master's degree completion within three years using machine learning (ML) algorithms and data from a large, public, Hispanic-serving university. Their study highlights the importance of both academic factors, such as cumulative GPA, and non-academic factors, including age, enrollment status, and financial aid. The study also stresses the importance of the department's composition. Departments with higher proportions of female faculty and faculty of color may foster inclusive cultures and support student integration.

From the methodological perspective, this study offers the review of different machine learning models and demonstrates the effectiveness of machine learning in predicting academic outcomes. The detailed discussion of the analytical strategy used in this study might be of special interest to IR professionals interested in replicating this study at their institutions.

Shulin Zhou, Yihui Li, Margot Neverett, Beverly King, Kyle Chapman, and Sharon McNair use survival analysis to examine factors influencing the persistence of transfer students at a four-year institution. The findings from the study suggest that higher transfer GPA, greater number of credit hours transferred, younger age, full-time enrollment, and enrollment in STEM majors lead to a greater likelihood of graduation. As demographics shift and the pool of first-time freshmen declines, ensuring transfer student success is vital for institutional sustainability and equitable educational outcomes. This research underscores the importance of proactive measures to support transfer students, ensuring they have equitable opportunities to complete their degrees and meet their educational goals. Use of survival analysis in this study might be of interest to IR professionals seeking to emphasize the timing of the event, handling censored data, or including time-varying predictors.

Together, these institutional studies contribute valuable insights into the experiences of transfer and graduate students and offer innovative methodological approaches that may appeal to IR professionals seeking to replicate these studies at their own institutions.

Iryna Muse
Editor

Editors

Iryna Muse
Editor
University of Alabama System

Inger Bergom
Assistant Editor
Harvard University

Leah Ewing Ross
Managing Editor
Association for Institutional Research

ISSN 2155-7535

Table of Contents

A Machine Learning Approach to Predicting Master's Degree Completion at the University of Texas at San Antonio

Fikrewold Bitew and Lauren Apgar

About the Authors

Fikrewold Bitew, PhD, and Lauren Apgar, PhD, have a combined 20 years of experience in the field of institutional research. They currently work in Institutional Research and Analysis at the University of Texas at San Antonio.

Acknowledgments

We would like to thank the Institutional Research and Analysis team at the University of Texas at San Antonio for their feedback and suggestions. Thank you to Khoi To for his insightful comments on an earlier draft of this manuscript.

Abstract

The pursuit of a master's degree is a significant academic endeavor, one that is influenced by a complex interplay of factors extending beyond traditional academic performance. In this study, we estimate the determinants of timely master's degree completion (i.e., within 3 years) using modern machine learning models such as random forest, decision tree, extreme gradient boosting, gradient boosting, and AdaBoost. After analyzing 15 years of master's cohort data from the University of Texas at San Antonio, a large, public, Hispanic-serving university, our findings indicated that gradient boosting with hyperparameter tuning was a reasonably superior machine learning model for predicting master's degree completion at our institution. The selected model accurately predicted more than 80% of the cases in the study and demonstrated superior predictive performance compared to the traditional logistic regression model. In support of nontraditional student retention theory, the model identified that students with higher GPAs, younger students, full-time students, and students who took out student loans were more likely to graduate within 3 years than students with lower GPAs, older students, part-time students, and students without loans, respectively. Furthermore, demographic-structural components, which are often overlooked in machine learning models, proved to be important: students in departments with a larger number of faculty and higher representation of female and non-White faculty members had a greater likelihood of completing their master's degree successfully.

Keywords: master's students, 3-year completion, machine learning, gradient boosting

The AIR Professional File, Winter 2025
Article 176

https://doi.org/10.34315/apf1762025

INTRODUCTION

The number and percentage of the U.S. adult population with graduate-level degrees have grown substantially over the past decade. Between academic years 2011–12 and 2021–22, the number of master's degrees awarded increased 16%, from 170,200 to 203,900 (National Center for Education Statistics, 2024). Despite the growth in the number of master's degrees conferred, the United States lacks nationwide data on which students begin a master's program but do not complete it. State-level data suggest that non-completion is a critical issue for students seeking master's degrees. For example, almost one-quarter (24%) of students seeking master's degrees from public institutions in Texas do not complete their degree within 5 years, according to our (the authors') analysis of data retrieved from the Texas Higher Education Coordinating Board (2022) Accountability System.

Completing a master's degree offers both advanced expertise in a field and increased earning potential. On average, graduate degree holders earn more than individuals with a 4-year degree (Pyne & Grodsky, 2020; Valletta, 2018). Students take substantial risks in pursuing a graduate-level degree, however, since graduate and professional students are more likely than undergraduate students to pay full tuition for their degrees (Woo & Shaw, 2015). Graduate and professional degree seekers have been taking out progressively larger student loans to finance their degrees over the past 20 years (Pyne & Grodsky, 2020). The increasing demand for advanced degrees, coupled with financial risks to students, underscores the importance for researchers and university administrators to understand the factors that influence master's degree completion.

Undergraduate degree completion represents a major area of focus in postsecondary education research, and universities worldwide have used educational data mining to predict students who are at risk of dropping out (Shafiq et al., 2022). Educational data mining often relies on administrative records as sources, then applies machine learning models to predict whether undergraduates will drop out of the institution (Shafiq et al., 2022). Supervised machine learning approaches in educational data mining include a wide array of predictors in their models (Shafiq et al., 2022), from social and academic integration within the institutional setting (Tinto, 1975) to student finances, family responsibilities, and outside employment (Bean & Metzner, 1985). Identifying factors that differ or are similar to those that influence undergraduate students can allow postsecondary institutions to develop targeted strategies to improve master's degree completion.

Additionally, systematic reviews of models in educational data mining show that most predictors included are at the individual-student level; only one study incorporated instructors' actions, such as posting grades (Shafiq et al., 2022). Studies lack predictors at the department level, but research in higher education on doctoral students highlights the importance of departmental environment and faculty mentoring (Council of Graduate Schools, 2013). Adding these variables to analyses can increase prediction accuracy in modeling, and so offer insights on the impact of institutional structure on completion.

In this study, we estimated several machine learning models to predict master's degree completion within 3 years from the University of Texas at San Antonio (UTSA), a Hispanic-serving, large, public institution in Texas. We evaluated the prediction accuracy across different models, identified students at risk of non-completion, and highlighted the 10 most important factors that predicted on-time degree completion (i.e., graduating within 3 years of starting a master's degree). In addition to variables such as GPA, age, and enrollment (i.e., whether students are part time or full time), demographic-structural components at the department level were important factors for accurate estimates of degree completion.

LITERATURE REVIEW

The theoretical framework behind undergraduate and doctoral student retention and degree completion can inform master's degree completion. Theoretical underpinnings stem from student–institution fit (Spady, 1970) and social integration (Tinto, 1975): as students develop peer and faculty relationships inside and outside the classroom, they become more attached to the university and are more likely to persist. However, the environment outside the institutional setting has more sway over nontraditional undergraduate students (Bean & Metzner, 1985) than it does over their traditional colleagues. Even if nontraditional students are socially integrated at the university, financial difficulties, outside employment, or family responsibilities conflict with their degree completion (Bean & Metzner, 1985). This pattern may also hold for master's students.

Although a thorough accounting is beyond the scope of this literature review (see other reviews, e.g., Mayhew et al. 2016), studies have used regression models to demonstrate quantitative support for these theories at the undergraduate level. In support of Bean and Metzner's (1985) theory, studies have shown that part-time enrollment correlates with work and family commitments (Nicklin et al., 2019), which can extend the timeline for degree completion or prevent it altogether. Full-time undergraduates often complete their degrees more quickly due to the continuity in learning and progression that their full-time status allows (Taniguchi & Kaufman, 2005). Undergraduates who are employed while in college are less likely to complete their degrees; among those who do complete them, however, they take longer to complete their degree than their nonworking colleagues (Ecton et al., 2023). Low- and moderate-income undergraduate students who receive need-based institutional aid are more likely to graduate within 6 years than are those who do not receive this type of aid (Price & Davis, 2006).

Not as many studies have focused on master's student retention and completion as undergraduate students; those studies that have done so largely use regression models and their findings support Bean and Metzner's (1985) theory. In Lenio's (2021) study of online master's student retention, employer financial support, student household income, student overall satisfaction with an institution's offices and support services, and student self-efficacy, as measured by a self-reported item on the importance of graduating from the institution, significantly predicted 1-year retention. Older master's students enrolled in a large, northeastern university were more likely to drop out and were less intent on persisting than were their younger colleagues (Cohen, 2012). Age may have served as a proxy for external environmental difficulties not measured by the study, including child care and/or work conflicts (Cohen, 2012).

Regression models have also been used to examine the relationship between social and academic involvement and undergraduate student retention and degree completion. Although Tinto (1975) views social integration as a psychological construct, research often measures student involvement in various activities. Using the Beginning Postsecondary Students (BPS:96/01) dataset (National Center for Education Statistics, 2003) in a multilevel event history model, Chen (2012) shows that social involvement (e.g., participation in fine arts activities, intramural sports, varsity sports, school clubs, and social activities with friends from school) and academic involvement (e.g., participation in study groups, meeting with an academic advisor, social contact with faculty, and talking with faculty about academic matters outside of class) decrease the odds of student dropout.

Despite the common use of regression analysis in the aforementioned studies, machine learning techniques have gained prominence as valuable tools for predicting and understanding factors that contribute to undergraduate student retention within U.S. institutions (Huo et al., 2023). Machine learning models incorporate a diverse range of variables that influence retention, including academic performance, financial aid, student demographic information, institutional enrollment patterns, and engagement with academic resources. Machine learning models analyze historical data to generate predictions that inform educators and administrators of the likelihood that any undergraduate student drops out. Often machine learning favors forms of modeling besides linear and logistic regression, since other types of models and computational models uncover different patterns and trends that more-accurately predict which students are likely to be retained, and so will eventually complete their degree.

While it is important to apply machine learning models to master's student degree completion because those students are an overlooked population, it is also important to incorporate a measure of social integration, which many educational data mining models lack (Shafiq et al., 2022). Mentoring and advising are not commonly collected institutional data points, as suggested by the lack of studies that include these types of indicators (Shafiq et al., 2022). However, Main (2018) has demonstrated that the structural-demographic composition of a department is related to doctoral degree completion. Drawing from Kanter's (1977) theory of proportions, Main proposed that, as faculty sex-ratios become more balanced within departments, tokenism, which evokes sex-typed or stereotypical roles, lessens. Main finds that female doctoral students are more likely to complete their degree in departments with higher proportions of female faculty. Similarly, racial/ethnic diversity among faculty members correlates with higher student graduation rates across 4-year institutions and community colleges (Stout et al., 2018). If direct measures of interactions with faculty are not available, structural-demographic department composition could serve to approximate the type of environment that would encourage student integration.

DATA AND METHODS

This study uses 15 years of master's student cohort data (entering Fall 2005 to Fall 2019) at our institution: the University of Texas at San Antonio (UTSA), a large, public Hispanic-serving institution located in the southern United States (N = 21,182). The outcome of interest is a dichotomous variable: completion of a master's degree from the institution within 3 years of entering. At our institution over this period, 59% of master's students completed their degree within 3 years (see Table 1). The data include individual-student level demographics, academic performance measures, and student financial aid information available in the university's student information system. We used the Python programming language (version 3.0) for data processing and analysis. We chose Python because it has many libraries for machine learning tasks, the coding language is relatively simple, and because it easily incorporates SQL, which our office relies on to pull student data out of our student information system. In addition, we chose it because it is a freely accessible program.

Table 1. 3-Year Completion Status of Master's Students, by Background Characteristics

	Completed Degree within 3 Years					
	Yes		No		Total	
Background Characteristics	#	%	#	%	#	%
Gender						
Female	7,123	59%	4,990	41%	12,113	100%
Male	5,326	59%	3,742	41%	9,068	100%
Unknown		0%	1	100%	1	100%
Race/Ethnicity						
American Indian or Alaska Native	20	34%	39	66%	59	100%
Asian	520	65%	286	35%	806	100%
Black or African American	729	57%	553	43%	1,282	100%
Hispanic or Latino	4,461	55%	3,634	45%	8,095	100%
International	1,852	81%	439	19%	2,291	100%
Native Hawaiian or Other Pacific Islander	23	64%	13	36%	36	100%
Two or More Races	213	59%	151	41%	364	100%
Unknown or Not Reported	462	58%	333	42%	795	100%
White	4,169	56%	3,285	44%	7,454	100%
First-Generation Status						
First Generation	5,193	56%	4,113	44%	9,306	100%
Not First Generation	6,693	60%	4,371	40%	11,064	100%
Unknown	563	69%	249	31%	812	100%
Full-time/Part-time Status						
Full-Time Status	7,453	73%	2,777	27%	10,230	100%
Part-Time Status	4,996	46%	5,956	54%	10,952	100%
Received Scholarship						
Yes	1,552	74%	537	26%	2,089	100%
No	10,897	57%	8,196	43%	19,093	100%
Received Grant						
Yes	1,774	64%	982	36%	2,756	100%
No	10,675	58%	7,751	42%	18,426	100%

Table 1. 3-Year Completion Status of Master's Students, by Background Characteristics (continued)

	Completed Degree within 3 Years					
	Yes		No		Total	
Background Characteristics	#	%	#	%	#	%
Took a Loan						
Yes	5,842	59%	4,051	41%	9,893	100%
No	6,607	59%	4,682	41%	11,289	100%
Research/Teaching Assistantships						
No	11,776	58%	8,494	42%	20,270	100%
Yes	673	74%	239	26%	912	100%
College						
Business	3,295	72%	1,281	28%	4,576	100%
Education and Human Development	3,831	57%	2,867	43%	6,698	100%
Engineering and Integrated Design	1,606	66%	812	34%	2,418	100%
Health, Community, and Policy	1,786	50%	1,794	50%	3,580	100%
Liberal and Fine Arts	794	46%	948	54%	1,742	100%
No College		0%	1	100%	1	100%
Sciences	1,137	52%	1,030	48%	2,167	100%
GMAT (Average)	550		546		549	
GRE (Average)	299		299		299	
GPA (Average)	3.7		3.4		3.6	
Age (Average)	28		31		29	
White, non-Hispanic Faculty (Average)		56%		59%		57%
Female Faculty (Average)		42%		47%		44%
Total	12,449	59%	8,733	41%	21,182	100%

Source: 15 years of entering master's cohort data from UTSA.

Variables that assess the nontraditional student model and highlight the financial environment that the student faces include dichotomous indicators of whether or not a student received a grant, scholarship, or loan during their first year in the master's program. We include an indicator for on-campus employment as Research/Teaching Assistantships. If a student ever worked as research or teaching assistant while enrolled at UTSA, then we considered them to be employed. Additionally, we include enrollment status: students enrolled in at least 9 credit hours during their first term were full time, and students enrolled in 8 or fewer credit hours were part time.

Student academic performance is measured through the last available cumulative GPA on record for the student. GRE and GMAT scores are added as continuous variables and categorical variables categorized into quintile groupings. Students who did not take a test were grouped into an additional "no test" category. GRE and GMAT scores are optional for admission into many master's programs at UTSA. Categorical statistics are not shown in Table 1, but are available upon request. An advantage of machine learning methods is that these models will accept both continuous and categorical measures in the same dataset. All dichotomous and categorical variables were encoded using either one-hot encoding or label encoding.

Variables that assess the structural-demographic composition model are department size and the demographic composition of faculty in each department for master's students entering cohort year. We include a variable measuring the percent of female faculty in a department and another variable measuring the percent of White, non-Hispanic faculty in a department. Indicators for broad fields (engineering, sciences, business, social science, education, and liberal and fine arts) are also included in the model.

The individual-student level demographic variables include a dichotomous indicator for female gender and race/ethnicity measured through the Integrated Postsecondary Education Data System (IPEDS) categories. IPEDS first identifies students who are not citizens or legal permanent residents of the United States as International. For the remaining students, Hispanic/Latino is prioritized, followed by racial identification as American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, or White. Students sometimes identify as Two or More Races. Students who do not identify their race or ethnicity are classified as Unknown or Not Reported. First-Generation Status refers to students whose parents (or parent) have not obtained a bachelor's degree. A continuous variable for age is also included.

Authors debate the use of demographic variables in predictive machine learning models. Some promote the use of demographic variables as a means to validate the fairness of model, instead of using them as predictors (Baker et al., 2023). Other authors promote the use of demographic variables as predictors in models because it results in better prediction; the inclusion of structural racism or sexism results in different outcomes for students that are not captured by other predictor variables (Wolff et al., 2013). Excluding demographic variables may obstruct opportunities to recognize racist practices. Not all models can measure every system and policy an institution has in place, and researchers' interpretations of model results with group disparities in degree completion should emphasize unmeasured structural factors. Similarly, measures of department demographic composition would point to leaders and administrators examining the types of mentoring opportunities and faculty–student interactions that occur within a department.

Analytic Strategy

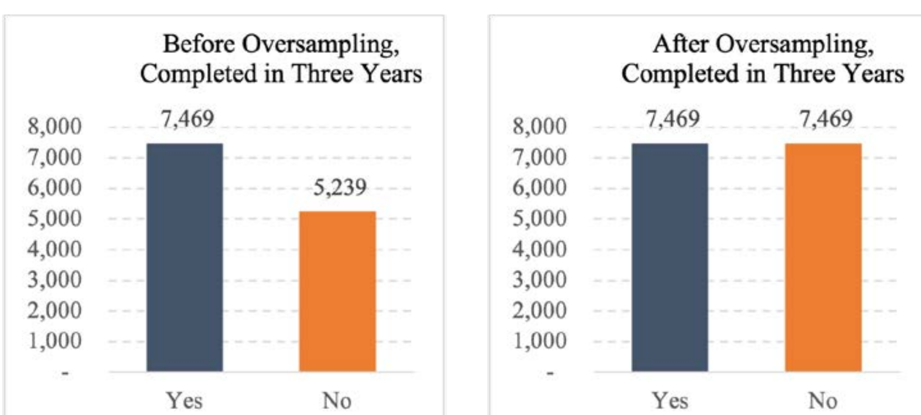
We applied five machine learning models (random forest, decision tree, extreme gradient boosting [XGBoost], gradient boosting, AdaBoost) and one traditional model (logistic regression) to identify the most appropriate model with the highest predictive power of a master's student degree completion. Decision tree models take tables as input, where tables can be numeric or categorical attributes (Safavian & Landgrebe, 1991). The attributes split the study sample, and splitting is repeated in a top-down manner to attain pure nodes, or the most homogeneous subset of data, based on a purity score. Random forest is a supervised ensemble learning method that acts based on decision trees (Ho, 1995). The random forest model repeatedly samples the variables in the training dataset and forms trees. After many of these trees are formed, the predictive performance of each variable is measured, and the best set of variables is obtained.

In contrast, boosting models (e.g., XGBoost, gradient boosting, and AdaBoost), build models sequentially in an adaptive manner, then combine them with a deterministic strategy. AdaBoost creates a strong classifier by combining weak classifiers, which are predictors that perform poorly but are better than random guessing (William, 2021). In the gradient boosting model, subsequent models attempt to reduce the errors of the previous model. For a dichotomous outcome, the gradient boosting classifier is used to minimize the loss function (Saini, 2021). Finally, XGBoost is a scalable implementation of the gradient boosting framework (Chen & He, 2018); compared to prior models, it offers better controls against overfitting by using more-regularized algorithm formalization.

Following standard methods for machine learning techniques, data were split into two sets: training and testing. Models were calculated from the training data, then applied to the test dataset, and model accuracy was assessed. Seventy percent of data were used for training, while the remaining 30% of data were held out as a test or validation set (N = 4,236); this 70–30 split is recommended for training and validation since it enables enough data points to be used for training to ensure a sensitive and complex model (Gholamy et al., 2018).

Additionally, because 41% of our master's students failed to earn their degree within 3 years in both the overall and test samples, we faced issues of imbalanced classification. Ideally, there would be a 50–50 split of successful and unsuccessful students in the data so that models learn effectively. To address the imbalance, we replicated students who had not completed their degree and added them to our training dataset. We then synthesized these additional cases using the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). SMOTE helps to increase the size of the minority class (i.e., students who did not complete their master's degree within 3 years) while maintaining the original distribution of the majority class (students who completed their degree within 3 years). SMOTE addresses the imbalance problem and allows machine learning models to make better predictions by reducing the bias toward the majority class. As shown in Figure 1, after applying SMOTE 50% of the test dataset did not complete their degree within 3 years. After addressing imbalanced data using SMOTE, the machine learning models were trained based on 10-fold cross validation on the training set and the performance was estimated on the testing set.

Figure 1. Before and After Oversampling by Master's Degree Completion Status



Source: 15 years of entering master's cohort data from UTSA.

Hyperparameter Tuning

A hyperparameter is a type of parameter, external to the model, that is set before the learning process begins. It is tunable and can directly affect how well a model performs. In this analysis, we used the random search hyperparameter tuning method instead of the grid search method. A random search uses a large (possibly infinite) range of hyperparameter values, and randomly iterates a specified number of times over combinations of those values. The number of iterations is specified by the researcher.

In this analysis, we ran all the models first with the default parameters, then compared the models with default parameters with models we ran after choosing the best parameters using hyperparameter tuning (see Appendix A). Except for logistic regression, all models with hyperparameter tuning were found to show higher predictive ability than models with default parameters. Thus, in the Results section of this article, only models with hyperparameter tuning are presented (except for logistic regression).

Model Evaluation

To verify each model's performance in terms of classifications and to help identify the best model, a confusion matrix (also known as an error matrix) was used (see Figure 2). A confusion matrix for bivariate outcomes is a two-by-two table showing values of true negative (tn), false negative (fn), true positive (tp), and false positive (fp) resulting from the test data. With these data classified, we next calculate precision (i.e., prediction accuracy), sensitivity (i.e., recall), specificity, and F1 score rates.

Figure 2. Confusion Matrix

		Real	
		Positive	**Negative**
Predicted	**Positive**	**True Positive (tp)**	**False Positive (fp)**
	Negative	**False Negative (fn)**	**True Negative (tn)**

Source: Kulkarni et al., 2020.

Precision: What percentage of students, as predicted by the model to complete their master's degree within 3 years, truly completed their degree within that time?

$$\textbf{\textit{precision}} = \frac{tp}{tp + fp}$$

Sensitivity (i.e., recall): What percentage of students who truly completed their degree within 3 years does the model predict as completers?

$$\textbf{\textit{sensitivity}} = \frac{tp}{tp + fn}$$

Specificity: What percentage of students who truly failed to complete their degrees within 3 years does the model predict as non-completers?

$$\textbf{\textit{specificity}} = \frac{tn}{tn + fp}$$

We also estimate an F1 score that combines precision and recall into a single metric. The F1 score has been designed to work well on imbalanced data.

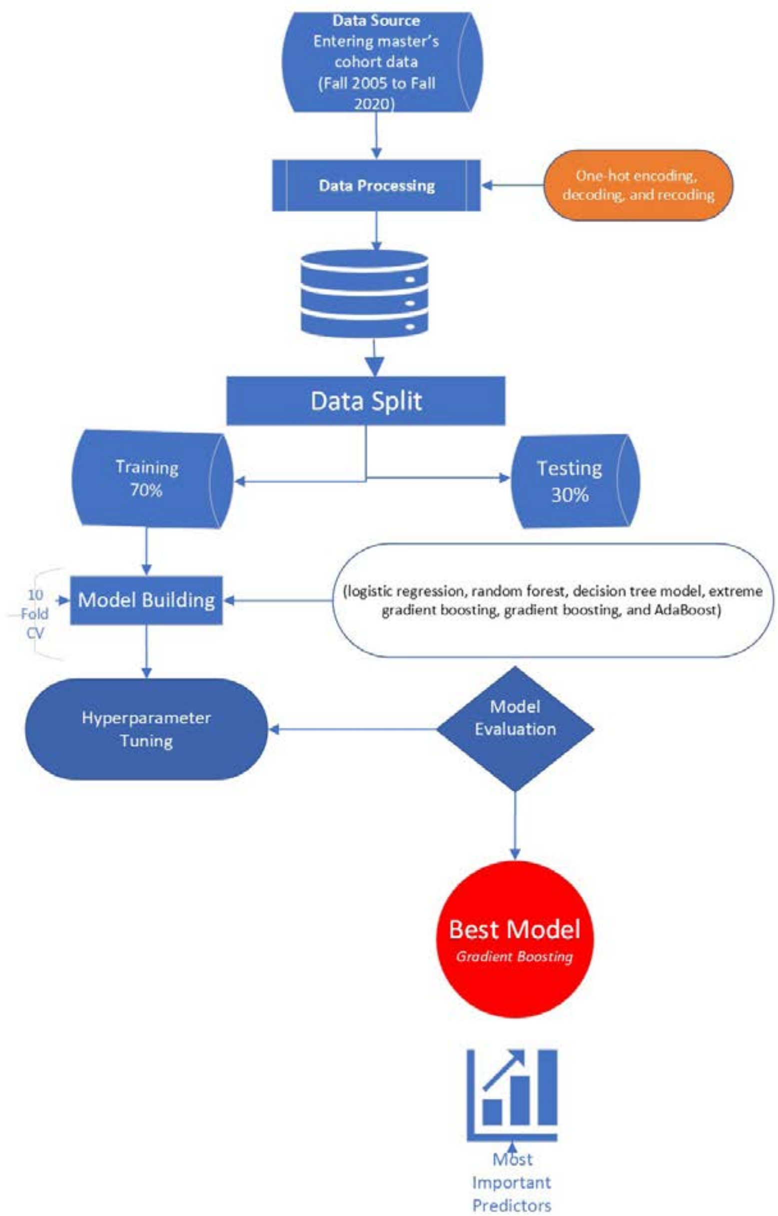
$$\textbf{\textit{F1 score}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Model Interpretation/Explanation Using SHapley Additive exPlanations

After identifying the best-fitting model using metrics from the confusion matrix, we use SHapley Additive exPlanations (SHAP) to interpret the predictions of the machine learning model (Lundberg & Lee, 2017). In machine learning research, it is rare to see explanation and interpretation of models, due to their black-box nature. The fundamental concept behind the SHAP analysis is to compute the marginal contribution of each predictor toward the outcome variable prediction result. We plot the aggregate SHAP value of the predictor for every sample to show whether that predictor increases or decreases a student's likelihood of master's completion by their 3rd year. SHAP also allows us to identify which predictors are important in predicting degree completion within 3 years by quantifying each variable's contribution to the prediction and aggregating it across the samples.

The overall data preparation and analysis process is presented in Figure 3.

Figure 3. Flow Chart of Data Preparation and Analysis Plan



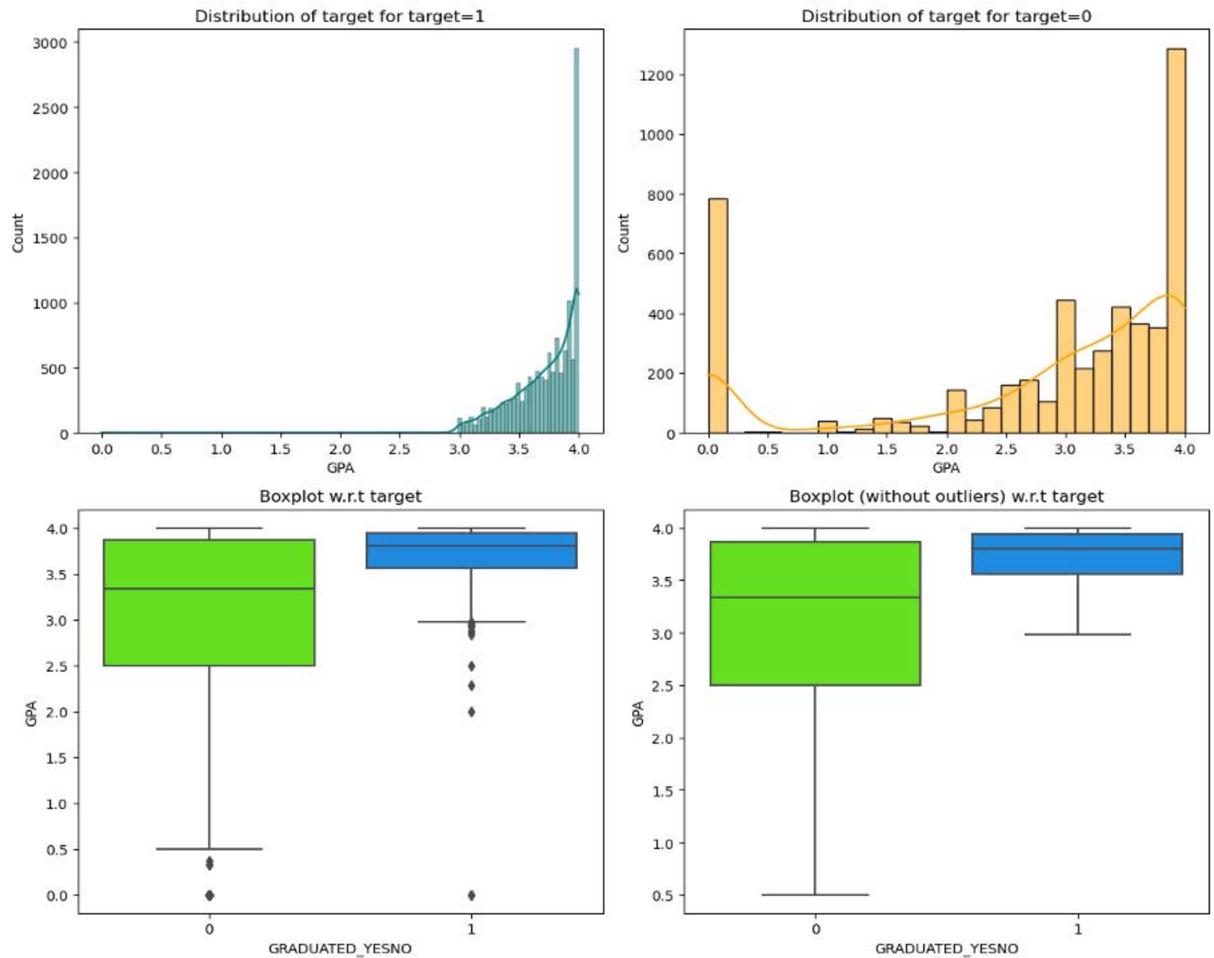
RESULTS

Table 1 presents selected descriptive statistics of our master's cohorts by their 3-year degree completion status. At our institution, more than half of the master's students were female, although there are no observable completion differences by gender. A sizeable number of students identify as Hispanic or Latino (38%), followed by White (35%); Hispanic or Latino and White students have similar master's degree completion rates at 55% and 56%, respectively. International students make up 11% of all master's students; international students have the highest master's degree completion rates at 81%. First-generation students (44% of master's students) have a lower (56%) master's degree completion rate compared to students with at least one parent who had obtained a bachelor's degree or higher (59%).

Indicators of student financial environment show that not many students were scholarship recipients; instead, almost half of all master's students took out a student loan. However, students who took out a loan completed their degree at similar rates as students who did not take out loans. Among full-time master's students, 73% completed their degree within 3 years, whereas only 46% of part-time students completed their degree within 3 years. The average age of students who completed their degree within 3 years was 28, as compared to an average age of 31 for non-completers. This difference in age suggests that students with fewer outside responsibilities are more likely to complete their degree.

Most of our master's students are either in the college of education (32%) or the college of business (22%). Students in business complete their degrees at the highest rate (72%), followed by students in engineering (65%) and education (57%); only 46% of liberal and fine arts students graduate within 3 years. Structural demographic composition of departments suggests that there is a relationship between department racial/ethnic diversity and degree completion. Among students who earn their master's degree, the departments where students pursue their degree average 56% White, non-Hispanic faculty compared to 59% White, non-Hispanic faculty among non-completers. Departments average 42% female faculty among completers compared to 47% among non-completers. Finally, higher cumulative GPA is highly correlated with higher levels of master's degree completion (Figure 4).

Figure 4. Distribution of Students' Cumulative GPA by Master's Completion Status (Yes (1) / No (0))



Source: 15 years of entering master's cohort data from UTSA.
Note: w.r.t = with respect to.

Models for Predicting Master's Completion

Tables 2 and 3 show the training and validation performance results for predicting master's degree completion for the six models estimated in this study. The five modern machine learning models (random forest, decision tree, XGBoost, gradient boosting, AdaBoost) showed a better predictive ability than the traditional model (logistic regression). We then check for overfitting to ensure that the models provide accurate predictions—not just for the training dataset, but also for testing data. When data scientists use machine learning models to estimate predictions, they often rely on 70% of their data to train their model. They then use their model fitted on their training dataset to predict outcomes for the remaining 30% of their data, or the testing dataset. When overfitting occurs, the model will show a high accuracy score on training data but a low accuracy score on test data. An overfit model can give inaccurate predictions and will not perform well for new data in the future.

Table 2. Training Performance Indicators of Five Machine Learning Models and a Traditional Model (Logistic Regression)

Measure	Logistic Regression	Random Forest	Decision Tree	Extreme Gradient Boosting	Gradient Boosting	AdaBoost
Accuracy	0.685	0.996	0.996	0.636	0.731	0.632
Recall	0.715	0.996	0.994	1.000	0.882	0.991
Precision	0.674	0.996	0.999	0.578	0.877	0.577
F1	0.894	0.996	0.996	0.733	0.768	0.730

Source: 15 years of entering master's cohort data from UTSA

.

Table 3. Validation Performance Indicators of Five Machine Learning Models and a Traditional Model (Logistic Regression)

Measure	Logistic Regression	Random Forest	Decision Tree	Extreme Gradient Boosting	Gradient Boosting	AdaBoost
Accuracy	0.700	0.740	0.870	0.702	0.773	0.704
Recall	0.731	0.798	0.892	0.997	0.876	0.994
Precision	0.752	0.789	0.732	0.884	0.769	0.888
F1	0.741	0.783	0.712	0.797	0.819	0.798

Source: 15 years of entering master's cohort data from UTSA.

A significant degree of overfitting was detected for the random forest and decision tree models. While these models demonstrated high accuracy, recall, precision, and F1 scores on the training datasets, their scores on the testing datasets were lower than on training datasets. As a result of overfitting, these models are unable to provide precise predictions. Thus, we compared the remaining three models (XGBoost, gradient boosting, and AdaBoost) to ascertain the optimal model.

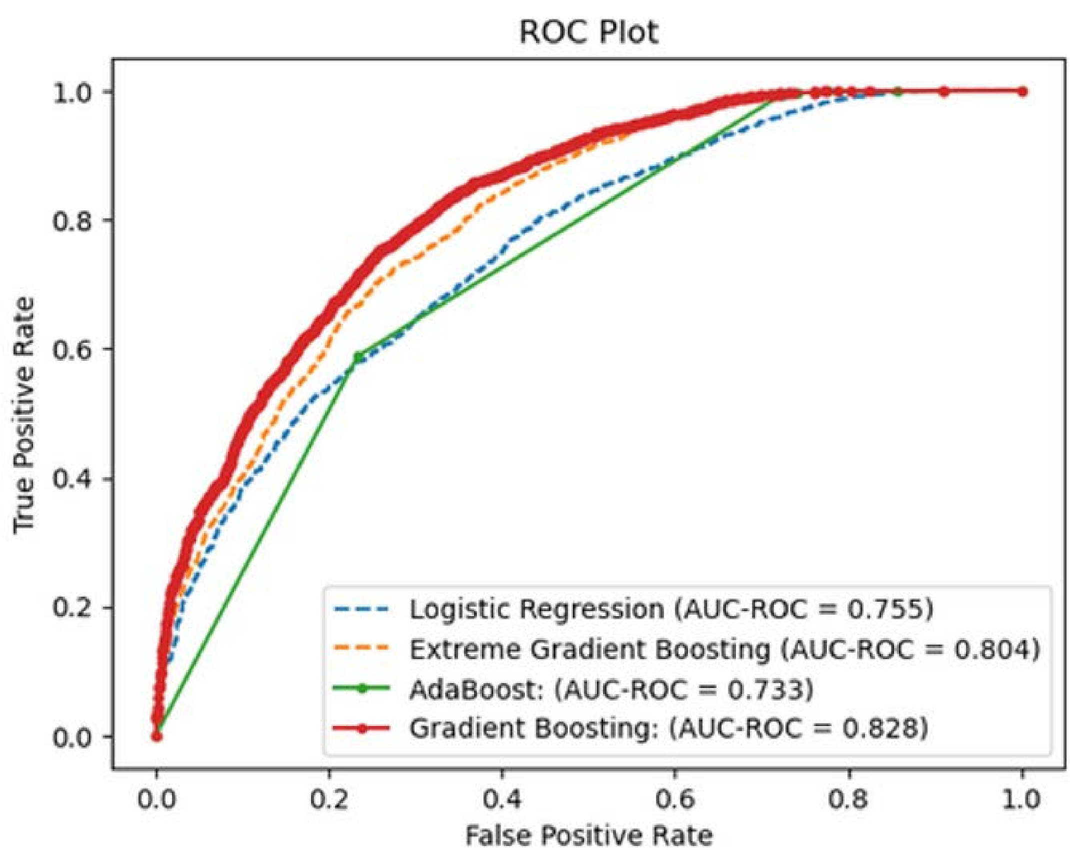
There are several evaluation metrics we can use to adjudicate between the remaining models. The accuracy metric is best used when we are interested in correctly predicting both completions and non-completions. For example, the gradient boosting model correctly predicted student degree completion outcomes 77% of the time in the testing data, compared to 70% for XGBoost and AdaBoost models. Recall is commonly used when correctly classifying an event that has already occurred, such as fraud detection, and when we are focused on identifying the true positives as often as possible. For this analysis, however, the F1 score integrates both the recall and the precision measures. Since it is a more comprehensive measure, we use the F1 score to evaluate between the three boosting models. XGBoost and AdaBoost models have relatively similar performance, with a slightly better performance observed for the gradient boosting model.

Area Under the Curve–Receiver Operating Characteristic

The Area Under the Curve–Receiver Operating Characteristic (AUC–ROC) curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve plotting the true positivity rate (sensitivity) against the false positivity rate (1 - specificity). The AUC represents the degree or measure of separability, summarizing how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting non-completers as non-completers, and completers as completers. In other words, the AUC denotes the percentage of the total cases that were predicted correctly by a model. Generally, an AUC between 0.7 and 0.8 is fair, between 0.8 and 0.9 is good, and 0.9 or above is excellent (Nahm, 2022).

The AUC–ROC curve (Figure 5) prefers the tuned gradient boosting model. The ROC curve for this model (bold red line in Figure 5) is the highest of all models, so does a better job of classifying the completers as completers. The AUC score of 0.828 is the farthest from 0.5, indicating the model is not classifying correctly, and the closest to 1, indicating the model perfectly distinguishes between completers and non-completers. The AUC score of 0.828 can be interpreted as meaning that the model correctly predicted 82.8% of total cases.

Figure 5. Area Under the Curve–Receiver Operating Characteristic Plot

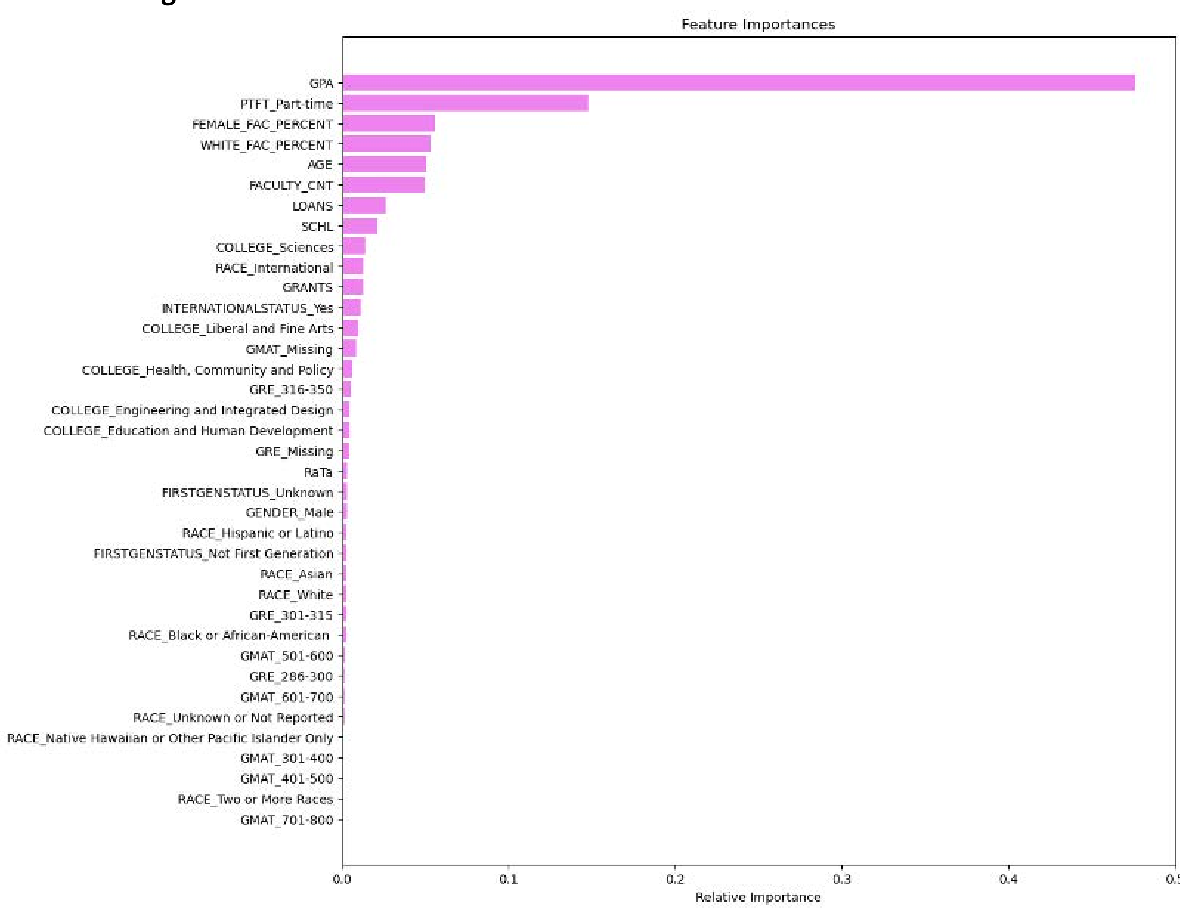


Source: 15 years of entering master's cohort data from UTSA.

Model Predictors

As described above, the accuracy results indicated that the tuned gradient boosting model was the best in predicting master's degree completion based on its F1 score, recall, and AUC–ROC. We then identify the top-10 predictor variables from this model based on the mean decrease in the Gini coefficient for master's degree completion (see Figure 6). These predictors are (1) last-earned cumulative GPA, (2) enrollment status as a part-time student, (3) the percentage of female faculty in the student's department, (4) the percentage of White, non-Hispanic faculty in the student's department, (5) student age, (6) the number of faculty in the student's department, (7) loans, (8) scholarships, (9) whether the student is studying in the college of sciences, and (10) whether the student is an international student per IPEDS race/ethnicity classification. While a strength of the tuned gradient boosting model is its ability to incorporate many predictors and to combine them to create a more accurate prediction, in order to focus on what theoretical frameworks receive the most support we present the top 10 predictors in our discussion. Additionally, a focus on the top 10 predictors allows our institution to design interventions or policy changes around the factors that are expected to have the largest impact on master's degree completion.

Figure 6. Top 20 Most Important Predictors of Master's Degree Completion from the Tuned Gradient Boosting Model

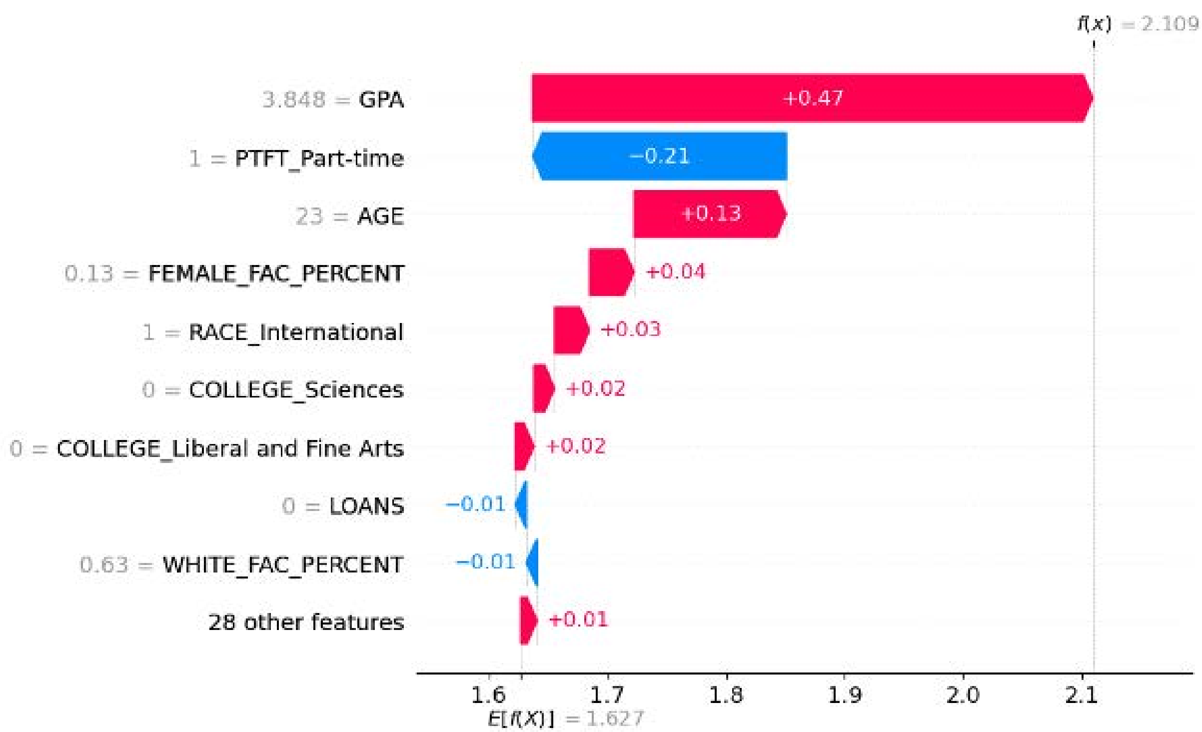


Source: 15 years of entering master's cohort data from UTSA.

We also used the model agonistic SHAP global feature importance for identifying the top predictors of master's degree completion. This technique examines the mean absolute SHAP value for each predictor across all the data, allowing us to identify the direction of the relationship between the predictor and master's degree completion.

Figure 7 displays the SHAP global importance scores for the top 10 factors, visualized using a Beeswarm plot and generated with the optimized XGBoost model. Higher cumulative GPAs have a significant positive influence on master's degree completion, whereas part-time study has a negative impact. Age, a higher percentage of female faculty in the student's department, being an international student, and enrolling in the college of sciences or the college of liberal and fine arts also have a positive effect on master's degree completion. In addition, not taking out loans and having a higher percentage of White, non-Hispanic faculty have a small negative impact.

Figure 7. Beeswarm Plot, Ranked by Mean Absolute SHAP Value Generated by Optimized Extreme-Gradient Boosting Model



Source: 15 years of entering master's cohort data from UTSA.

DISCUSSION AND CONCLUSION

With an increasing number of students pursuing master's degrees, it is essential to evaluate the master's student experience and identify the factors contributing to their timely degree completion. While the master's 3-year completion rate at UTSA is higher than the undergraduate 6-year completion rate, non-completion of a master's within the expected 3 years is still prevalent. Accuracy in prediction becomes even more important when completion is higher, since it is more difficult to identify potential non-completers. Our study offers further evidence that machine learning models predict degree completion more accurately than a traditional logistic regression model. With a gradient-boosting model in place, our institution can more precisely identify students who are likely to drop out or lag in their degree completion and target their services toward these students. Not only could UTSA save money by knowing which students to target with services, but it also potentially increases its alumni giving when more students graduate with a master's degree.

We identified the variables that saw the greatest gains in the gradient-boosting model's performance, combining some classic theoretical models along with an organizational demography approach. The top variables in our model included cumulative GPA, enrollment status, the demographic composition of the student's department (e.g., percent female faculty and percent White, non-Hispanic faculty), student age, and student financial aid (e.g., whether a student took out loans and/or received scholarships). These and other variables in our model predicting master's degree completion support much of what has been found in the literature, showing that theories developed for nontraditional and doctoral students also apply well to master's students.

Academic performance is key, since students with higher cumulative GPAs are more likely to complete their degree within 3 years. While cumulative GPA is an important predictor, non-academic factors and outside environment also play a crucial role in master's degree completion, as suggested by the nontraditional student model of retention. Enrollment status is the second-most impactful predictor of master's completion and is indicative of the influence of the outside environment, such as employment and/or family conflicts (Nicklin et al., 2019). Similarly, younger students often transition to their graduate studies directly from their undergraduate experience at a time when they have fewer outside conflicts, whereas older students might be balancing school, work, and family obligations. The nontraditional student model of retention also highlights the importance of student finances. Students enter the master's program with different levels of family and employer financial support, and financial aid can mitigate financial barriers. Grants and scholarships alleviate financial pressure, and students who received this type of aid were more likely to complete their master's degree within 3 years. While the accumulation of debt can increase financial stress and negatively impact a student's ability to persist (Baker et al., 2017), our study suggests that the master's students who took out loans were more likely to complete their degree, possibly signaling student commitment to their degree and its potential returns. The importance of student finances and financial aid on master's completion highlights how imperative it is for student financial needs to be met if they are to finish their degree within 3 years.

While factors in the nontraditional model (enrollment status, age), as well as GPA, have the strongest associations with degree completion, this study also highlights the importance of organizational demography. Based on Kanter's (1977) theory of proportions, higher proportions of female faculty and faculty of color might be associated with a departmental culture that facilitates the degree attainment of students of all genders and racial/ethnic backgrounds. Kanter theorized that larger proportions of previously minoritized groups would reduce tokenism and reliance on stereotypes. Other research suggests that female faculty members serve as mentors for female students, fostering a sense of student belonging and inclusion (Johnson, 2014); a similar dynamic could be in play for students of color. Department size also plays a role in degree completion, since additional faculty can lead to increased attention from and availability to students (Rujimora et al., 2023). A limitation of structural-demographic measures is that these measures only hint at the environment of the department or existing programs that could result in student integration. The relationship between faculty demographics and master's degree completion can be influenced by faculty–student interaction, mentoring relationships, and institutional support systems. Nevertheless, the demographic composition of the department can influence relationship building, and can be used to approximate student integration when more-direct measures are not available.

One limitation of this study is its reliance on institutional data instead of survey data. As a result, we do not have indicators of student belonging and social integration into the university, or good measures of faculty–student interactions. More research is needed to assess whether the impact of organizational demography on master's completion is mediated through a sense of student belonging. Furthermore, this study is a case study on one large, public 4-year institution. While the methodology may be generalized to other universities, the results and key predictors are specific to our institution. Additional research is needed to determine whether these variables also influence master's completion within 3 years at other institutions, or if different theoretical models hold sway elsewhere. Still, the use of machine learning techniques for predicting master's degree completion represents a significant step forward in educational research, along with the incorporation of structural-demographic factors. These data-driven insights hold immense potential for advancing student success and timely master's degree completion in our institution and offer an exemplar that can be replicated across other institutions in the United States.

APPENDIX A: HYPERPARAMETER DEFAULTS AND TUNING

For tree base learners, the most common parameters are

- Max depth: The maximum depth per tree. A deeper tree might increase the performance, but it also increases the complexity and chances to overfit.
  Max depth = None is used. Default is 6.
- Learning rate: The learning rate determines the step size at each iteration while the model optimizes toward its objective. A low learning rate makes computation slower, and requires more rounds to achieve the same reduction in residual error as a model with a high learning

rate, but also optimizes the chances to reach the best optimum.
The value we used here is 0.05. Default is 0.3.

- N estimators: The number of trees in our ensemble. Equivalent to the number of boosting rounds.
The value must be an integer greater than 0. Default is 100.
- Column sample by tree: Represents the fraction of columns to be randomly sampled for each tree. It might improve overfitting.
The value must be between 0 and 1. Default is 1.
- Subsample: Represents the fraction of observations to be sampled for each tree. A lower value prevents overfitting but might lead to underfitting.
The value must be between 0 and 1. Default is 1.

REFERENCES

Baker, A. R., Andrews, B. D., & Mcdaniel, A. (2017). The impact of student loans on college access, completion, and returns. *Sociology Compass, 11,* E12480.

Baker, R. S., Esbenshade, L., Vitale, J., & Karumbaiah, S. (2023). Using demographic data as predictor variables: A questionable choice. *Journal of Educational Data Mining, 15,* 22–52.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55,* 485–540.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16,* 321–357.

Chen, R. (2012). Institutional characteristics and college student dropout risks: A multilevel event history analysis. *Research in Higher Education, 53,* 487–505.

Chen, T., & He, T. (2018). XGBoost: Extreme gradient boosting. *R.* 0.71.1 Ed.

Cohen, K. E. (2012). *Persistence of master's students in the United States: Development and testing of a conceptual model* (Order No. 3543989). [Doctoral dissertation, New York University]. ProQuest Dissertations Publishing.

Council of Graduate Schools. (2013). *Master's completion project: Pilot study on completion and attrition in STEM master's programs.* https://cgsnet.org/project/masters-completion-project

Ecton, W. G., Heinrich, C. J., & Carruthers, C. K. (2023). Earning to learn: Working while enrolled in Tennessee colleges and universities. *Aera Open, 9,* 1–16.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *International Journal of Intelligent Technologies and Applied Statistics, 11,* 105–111.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282). IEEE.

Huo, H., Cui, J., Hein, S., Padgett, Z., Ossolinski, M., Raim, R., & Zhang, J. (2023). Predicting dropout for nontraditional undergraduate students: A machine learning approach. *Journal of College Student Retention: Research, Theory & Practice, 24,* 1054–1077.

Johnson, I. Y. (2014). Female faculty role models and student outcomes: A caveat about aggregation. *Research in Higher Education, 55,* 686–709.

Kanter, R. M. (1977). Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American Journal of Sociology, 82,* 965–990.

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In *Data democracy* (pp. 83–106). Academic Press.

Lenio, J. A. (2021). Employer support as a predictor of online master's student retention. *Journal of College Student Retention: Research, Theory & Practice, 25,* 572–593.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the advances in neural information processing systems* (pp. 4765–4774).

Main, J. B. (2018). Kanter's theory of proportions: Organizational demography and PhD completion in science and engineering departments. *Research in Higher Education, 59,* 1059–1073.

Mayhew, M. J., Rockenbach, A. N., Bowman, N. A., Seifert, T. A., Wolniak, G. C., Pascarella, E. T., & Terenzini, P. T. (2016). *How college affects students: Vol. 3, 21st century evidence that higher education works.* Jossey-Bass.

Nahm, F. S. (2022). Receiver operating characteristic curve: Overview and practical use for clinicians. *Korean Journal Anesthesiology, 75,* 25–36.

National Center for Education Statistics. (2003). *Beginning postsecondary students (BPS 96/01)* [dataset]. https://nces.ed.gov/DataLab/

National Center for Education Statistics. (2024, May). *Condition of education: Graduate degree fields.* https://nces.ed.gov/programs/coe/indicator/ctb

Nicklin, J. M., Meachon, E. J., & Mcnall, L. A. (2019). Balancing work, school, and personal life among graduate students: A positive psychology approach. *Applied Research in Quality of Life, 14,* 1265–1286.

Price, D. V., & Davis, R. J. (2006). *Institutional grants and baccalaureate degree attainment.* National Association of Student Financial Aid Administrators.

Pyne, J., & Grodsky, E. (2020). Inequality and opportunity in a perfect storm of graduate student debt. *Sociology of Education, 93,* 20–39.

Rujimora, J., Campbell, L. O., & Demara, R. F. (2023). Exploring the student-to-faculty ratio and degree attainment in Florida. *Journal of Hispanic Higher Education, 23,* 215–227.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics, 21,* 660–674.

Saini, A. (2021). *Gradient boosting algorithm: A complete guide for beginners.* https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: A systematic literature review. *IEEE Access, 10,* 72480–72503.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 1,* 64–85.

Stout, R., Archie, C., Cross, D. & Carman, C. A. (2018). The relationship between faculty diversity and graduation rates in higher education. *Intercultural Education, 29,* 399–417.

Taniguchi, H., & Kaufman, G. (2005). Degree completion among nontraditional college students. *Social Science Quarterly, 86,* 912–927.

Texas Higher Education Coordinating Board. (2022). *Texas higher education accountability system interactive* [dataset]. http://www.txhigheredaccountability.org/AcctPublic/InteractiveReport/AddReport

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45,* 89–125.

Valletta, R. G. (2018). Recent flattening in the higher education wage premium: Polarization, skill downgrading, or both? In *Education, skills, and technical change: Implications for future US GDP growth* (pp. 313–334). University of Chicago Press.

William, A. (2021). A comprehensive mathematical approach to understand AdaBoost. *Medium.* https://towardsdatascience.com/a-comprehensive-mathematical-approach-to-understand-adaboost-f185104edced

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. *Proceedings of the 3rd international conference on learning analytics and knowledge* (pp. 145–149).

Woo, J., & Shaw, S. (2015). *Trends in graduate student financing: Selected years, 1995–96 to 2011–12.* National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2015026

Survival Analysis of Transfer Students

Shulin Zhou, Yihui Li, Margot Neverett, Beverly King,
Kyle Chapman, and Sharon McNair

About the Authors

Shulin Zhou, Yihui Li, Margot Neverett, Beverly King, and Kyle Chapman are with the Office of Institutional Planning, Assessment, and Research at East Carolina University. Sharon McNair is with the Department for People Operations, Success, and Opportunity at East Carolina University.

Abstract

Research conducted on transfer student outcomes consistently shows that there is a bachelor's degree completion gap between transfer students and nontransfer students. Researchers have explored several factors thought to impact bachelor's degree completion for transfer students, including demographic characteristics, number of credit hours transferred, transfer GPA, transfer institution type, and indicators of academic achievement. The findings from these studies have not always been consistent in whether (or how) these factors influence degree completion. The current study uses survival analysis to better understand college persistence for students transferring to a large, 4-year, public university located in the Southeast United States. Survival analysis, a statistical technique underutilized in higher education research, has several advantages over more traditional methods, such as regression. For example, survival analysis not only has the capacity to examine time-varying predictors, but also can include both uncensored and censored events (i.e., it can handle both students for whom the event of interest occurs during the time frame under investigation and students for whom it does not). In addition to variables explored in previous research, this study investigated aspects of students' majors (i.e., whether they changed majors after enrollment and whether majors were in STEM fields). Findings indicate that transfer students who are most likely to persist are generally younger, are full-time students, and are in STEM majors; and that they have higher prior academic achievement, a greater numbers of transfer hours, and at least one major change.

Keywords: survival analysis, transfer students, college students, persistence

The AIR Professional File, Winter 2025
Article 177

https://doi.org/10.34315/apf1772025

INTRODUCTION

Today's students often transition between different types of higher education institutions and pursue nontraditional pathways in their quest to complete a postsecondary degree. Not uncommonly, students enroll in 4-year universities after having attained an associate's degree, or, at least, some credit hours at a community college. This trend underscores the importance of examining the outcomes of transfer students, who face unique challenges and opportunities in their quest for bachelor's degrees. Despite the increasing prevalence of such nontraditional education pathways, research conducted on transfer student outcomes consistently shows that there is a bachelor's degree completion gap between transfer students and nontransfer students. For example, Shapiro et al. (2017) tracked a cohort of community college students for 6 years, beginning in 2010. Of the students who transferred to bachelor's degree–granting institutions, 42% attained bachelor's degrees. The American Council on Education's (2021) National Task Force on the Transfer and Award of Credit, citing the Shapiro et al. study, pointed out that "this rate of baccalaureate completion represents a roughly 17 percent gap for transfer students compared to students who receive a degree within the same institution of attendance (without transfer)" (p. 35). This discrepancy points to a critical need for higher education institutions to better understand and support the unique needs of transfer students.

The significance of studying transfer students extends beyond academic achievement to encompass broader implications for both students and institutions. From a student's perspective, successful completion of a bachelor's degree can significantly impact lifetime earnings, employment opportunities, and social mobility (Baum et al., 2013). For institutions, improving transfer student outcomes not only is a matter of academic responsibility, but also has financial implications, affecting enrollment management, resource allocation, and institutional reputation (Tinto, 1993). Furthermore, as higher education faces increasing scrutiny over costs and value, demonstrating success in facilitating transfer students' completion becomes paramount for institutional accountability and sustainability.

The current study seeks to contribute to the literature on college persistence by using survival analysis, a statistical technique that, despite its potential, has been underutilized in higher education research. Survival analysis offers a nuanced approach to examining the time-dependent nature of student retention and completion, allowing for the inclusion of both censored and uncensored values (i.e., data for students for whom the event of interest occurs during the time frame under investigation and students for whom it does not) and the assessment of time-varying predictors (Ronco, 1995). This methodological choice is particularly relevant for studying transfer students, whose educational pathways may be more varied and complex than those of first-time-in-college students.

Unique to this study is the inclusion of two variables related to student major, addressing a gap in the existing literature. Previous research has often overlooked the role of academic discipline in influencing student outcomes, despite evidence suggesting that major choice can significantly impact persistence rates (Wright, 2018). By incorporating these variables, the study offers new insights into the factors that contribute to transfer students' success. The context of the institution in the study—a large, 4-year, public university—is especially pertinent, given the growing recognition of the role of public universities in providing accessible and affordable education to a diverse student body, including transfer students. This focus becomes even more significant because this university is increasingly attempting to understand and support the unique pathways to graduation for transfer students, moving beyond the traditional emphasis on first-time-in-college students.

LITERATURE REVIEW

Before describing the statistical analysis and results, a review of the literature will be presented, detailing key factors that have been found to impact student persistence. The literature is reviewed in two sections: The first section includes studies using analytic techniques other than survival analysis; the second section focuses specifically on studies that did use survival analysis. Note that, due to the low number of studies utilizing survival analysis to investigate transfer student success in college, some studies have been included in the latter section that focus on first-time-in-college students.

Non–Survival Analysis Studies

As noted by Barbera et al. (2017), research spanning several decades has attempted to understand predictors of student success in undergraduate degree programs. Studies have demonstrated that demographic variables, community college credentials, the number of credit hours transferred, and transfer institution type are all factors that influence transfer students' persistence rates.

DEMOGRAPHIC VARIABLES

The impact of student demographics on transfer student persistence rates is underscored in recent research by Marbouti et al. (2021) that focuses on students within San Jose State University's College of Engineering, a large percentage of whom were transfer students. Their study highlights significant disparities among students, particularly concerning ethnicity, gender, and financial aid patterns. Despite eligibility for financial assistance, Hispanic, first-generation, and low-income transfer students exhibit lower GPAs and experience delays in graduation. Wang (2009), utilizing data from the National Education Longitudinal Study of 1988 and the Postsecondary Transcript Study, examined graduation probabilities for community college transfers to 4-year institutions and revealed a gender discrepancy, with females demonstrating a higher likelihood of completing bachelor's degrees, even after accounting for other variables. Taplin's (2019) research at a large public university identified a significant association between family income (indicated by Pell grant eligibility) and both 1-year retention and 6-year graduation rates among transfer students.

COMMUNITY COLLEGE CREDENTIALS

The impact of earning an associate's degree on whether students go on to complete a bachelor's degree yields mixed findings (Zhang, 2022). On the one hand, some research suggests a positive influence of obtaining an associate's degree on community college students' likelihood of transferring to a 4-year university and achieving academic success post-transfer (Daddona et al., 2021). For example, Kopko and Crosta (2016) conducted logistic regression analysis on a statewide sample and discovered that transfer students entering a 4-year institution with an Associate of Arts or Associate of Science degree were approximately 50% more likely to graduate within 6 years compared to transfer students without an associate's degree. However, possession of an Associate of Applied Science degree did not enhance the likelihood of bachelor's degree attainment within 6 years.

Other research suggests that acquiring an associate's degree prior to transfer may not significantly affect students' academic success at 4-year institutions (Wang, Chuang, et al., 2017). Jenkins and Fink (2016), leveraging data from the National Student Clearinghouse, concluded that the relationship between community college credentials and bachelor's degree completion within 6 years was not universally observed across many states. Some authors suggest that the number of credits accepted by the receiving 4-year institution is more predictive of a transfer student's academic outcome than simply possessing an associate's degree (Monaghan & Attewell, 2015; Zhang, 2022).

NUMBER OF CREDIT HOURS TRANSFERRED

The existing literature presents a diverse range of findings regarding the relationship between transfer student success and the volume of transferred credit hours. Some studies suggest that transferring a greater number of credit hours correlates positively with transfer student success (Yang et al., 2018). These studies indicate that students who transfer more credits are more likely to fulfill degree requirements promptly and to achieve better academic performance (Daddona et al., 2021).

Conversely, other research offers nuanced perspectives, suggesting that, while transferring a substantial number of credit hours may seem advantageous initially, it can also pose challenges. For example, students transferring many credits may face difficulties assimilating into the new academic environment, meeting remaining degree requirements, or accessing necessary support services (Gardner et al., 2021).

If a relationship exists between number of transferred credit hours and transfer student success in college, it may not be a simple one. Luo et al. (2007) used an institutional sample of 1,713 transfer students from five cohorts, categorized by entering class level (freshman, sophomore, junior), as determined by credit hours transferred. Through sequential logistic regression, they found that different factors influenced retention for the three class levels. For entering freshmen, retention was predicted by gender and first-term GPA. Retention to the 2nd year for entering sophomores was predicted by hours transferred, but there were interactions with financial aid, age, and 1st-year GPA. One-year retention for entering juniors was predicted by a set of interacting factors: transfer credit hours, total credit hours, and GPA earned post-transfer.

TRANSFER INSTITUTION TYPE

Aulck and West (2017) performed a descriptive analysis of transcripts from nearly 70,000 entering students over an 8-year period at a large public institution to investigate persistence and attrition. They compared native freshmen, transfers from 2-year institutions, and transfers from 4-year institutions, and found that native freshmen and 2-year transfers had similar attrition rates and GPAs. Transfers from 4-year schools had higher GPAs than the other two groups but also had higher rates of attrition.

Survival Analysis Studies

Although not as common as statistical techniques such as regression analysis, some research studies in higher education have utilized survival analysis to examine factors that impact student persistence rates in the areas of student demographics, academic achievement, and college experience.

STUDENT DEMOGRAPHICS

Demeter et al. (2022) examined the role of student demographics in predicting time to degree completion. They found that factors such as gender, race/ethnicity, and first-generation status were significant predictors of graduation probability. Consistently, studies have confirmed the influence of gender on college persistence. For instance, Ronco (1995) reported that gender played a role, albeit a small one, in predicting exit, with female students slightly more likely to graduate than male students. Similarly, Wang, Wang, et al. (2017) revealed that female students tend to have a higher probability of degree completion. Hayward (2011) reported that gender (male) had a small negative effect on transfer. Chimka et al. (2007) found gender differences related to standardized test scores. Female students with better standardized math scores were more likely than similar male students to graduate. As far as race/ethnicity is concerned, Lin et al. (2020) noticed that a significant gap in the likelihood of bachelor's degree completion between Black and White students emerged more episodically, while the gap between Hispanic and White students developed earlier and remained more consistent over time. Wang, Wang, et al. (2017) also noted that the probability of completing a degree is higher for White students. In addition, in Murtaugh et al. (1999), univariate analysis suggested that Black, Hispanic, and American Indian students are at greater risk of withdrawing than are White students; the differences disappeared in a multivariate analysis, however, and Black students seemed to have reduced withdrawal risk, compared to White students. Lin et al. (2020) discovered that achieving academic milestones, such as credit momentum and the completion of pre-transfer associate's degrees, benefits all students, but benefits Black and Hispanic students disproportionately.

Fewer studies are available examining the relationship between socioeconomic status (SES) and college persistence. A notable exception is a study by Reynolds and Cruise (2020) who focused on the impact of SES on student retention. They found that students from lower-income backgrounds had a higher hazard rate of dropping out compared to their higher-income counterparts. Hutton (2015) suggested that financial aid had a small negative impact on graduation with, not surprisingly, the odds of departure lower when higher percentages of educational cost were covered by financial aid.

Not all researchers have found demographic variables to influence college persistence. Hutton (2015), for example, used discrete-time survival analysis to examine factors predicting community college students' completions at a public university, and concluded that college persistence and completion appeared to be unaffected by demographic variables. Finally, regarding age, results are mixed. As students age, according to research by Hayward (2011), they are generally less likely to transfer. If older students do transfer, however, according to Murtaugh et al. (1999), they are less likely to be retained. On the other hand, Wang, Wang, et al. (2017) found that older students tend to have a higher probability of degree completion.

ACADEMIC ACHIEVEMENT

Survival analysis has also been utilized in higher education research to examine the impact of academic achievement on student persistence rates. In general, researchers have found that persistence increases with better high school GPAs (Choudhury & Runco, 2020; Miller & Lesik, 2014; Murtaugh et al., 1999). For example, a study by Allensworth and Clark (2020) examined the relationship between high school GPA and time to graduation from college. They found that students with higher high school GPAs were less likely to drop out and took less time to graduate from college. However, Voelkle and Sander (2008) pointed out that the effect of high school GPA on dropping out of college may be completely mediated by university GPA, so there would be no additional predictive ability of high school GPA over university GPA. ACT and SAT scores are also important predictors: McNeish et al. (2020) explored the predictive power of standardized test scores on student retention. They found that students with higher test scores were less likely to drop out of college and graduated more quickly than students with lower scores. Looking at several academic achievement variables used in the college admissions process, Miller and Lesik (2014) found that retention was associated with higher entry-level academic preparation (ELAP) scores (categorized into high, medium, and low), and that the effect was consistent across time. ELAP was determined by a combination of ACT score, high school class rank, and the number of college prep units. Students with higher ELAP were found to be more likely to graduate in Years 4 and 5 compared to students with lower ELAP scores.

Achievement at a community college has also been shown to be related to likelihood of graduation with a bachelor's degree after transfer. In the study by Hutton (2015), the number of earned community college credit hours had a small but positive impact on graduation, while the attainment of an associate's degree had a larger positive impact. Hutton also noted that semester GPA had a strong impact on the odds that a student would eventually graduate or depart before graduation. The importance of college GPA as a predictor of outcomes is confirmed in many studies. For example, Murtaugh et al. (1999) found that retention increases with increasing first-quarter GPA. Similarly, Ronco (1995) found that students who exit through dropout or transfer are most likely to do so because of the immediate impact of a GPA below 2.0, with students having failing GPAs six and a half times more likely to drop out and eight and a half times more likely to transfer.

A full-time enrollment status is found to be positively related to graduation or credential completion and negatively related to dropout or transfer (Ronco, 1995; Wang, Wang, et al., 2017). Hutton's (2015) study confirmed that students who stop out (i.e., who leave college but eventually return) and were part-time students had significantly lower graduation rates and higher departure rates.

COLLEGE EXPERIENCE

Within the scope of understanding persistence rates for transfer students, using survival analysis offers a nuanced lens to explore the dynamic interplay between engagement in the college experience and these pivotal academic outcomes. For example, a study by Caruth (2018) explored the relationship between student engagement and time to graduation. Caruth found that students who were more engaged in campus activities and who had higher levels of social integration had a lower hazard rate of drop out or delay to graduation.

In the study by Miller and Lesik (2014), the effect of 1st-year experience participation on retention was found only for the 1st year, but the influence of 1st-year seminar participation reappeared for 4-year graduation, perhaps due to an indirect variable such as beginning college ability. In Choudhury and Runco's (2020) study, results suggested that a university course that focuses on time management, note-taking, test-taking, studying, and so on increases the retention rate by approximately 38%. Murtaugh et al. (1999) also found that students taking a freshman orientation course appeared to be at reduced risk of dropping out.

There is evidence that transfer shock, typically defined as a drop in GPA between pre- and post-transfer institutions, plays a role in student departure. In a study of North Carolina community college transfer students, the odds of departure were higher for students who experienced transfer shock; there was no statistically significant effect on graduation, however (Hutton, 2015).

Based on these findings, one can conclude that it is meaningful to conduct a multivariate survival analysis incorporating variables related to pre-academic preparation, college experience, and demographics. This conclusion can be supported by one of the findings from Miller and Lesik (2014), who noted that differing results were found in a survival analysis than in a descriptive analysis. For example, descriptive analysis showed a positive impact of 1st-year seminar across all ability levels, but survival models showed only initial effects. Murtaugh et al. (1999) also observed that the relationship between retention and race and/or ethnicity was different in the univariate versus multivariate views. Finally, Mourad and Hong (2008) emphasized the importance of considering the interaction effect of time and other variables. In their study, the effect of time resulted in changes from a statistically significant to a nonsignificant relationship, or from a nonsignificant to a statistically significant relationship for some variables. However, very few studies (Hutton, 2015; Lichtenberger & Dietrich, 2017) have used survival analysis to better understand transfer student persistence. To fill this gap, the current study used survival analysis to investigate the persistence of transfer students during their (initial) 4 years at the transfer institution. Specifically, the research addressed the following questions:

1| What is the estimated survival rate of transfer students within eight semesters after enrollment?

2| Are there significant differences between the survival rates of subgroups based on age, major, major change, transfer GPA, number of transfer credit hours, financial aid received, enrollment status, race and/or ethnicity, and/or gender?

3| How large are the effects of covariates on transfer students' persistence rates?

METHODS

Quantitative studies of college retention and completion have most often used regression models. Another statistical technique that has been gaining in popularity within higher education research is survival analysis. Survival analysis refers to "a set of statistical methods for investigating the time it takes for an event of interest to occur" (Statistical Tools for High-throughput Data Analysis [STHDA], n.d.). The origins of survival analysis can be traced back to early work on mortality in the 17th century (Lee & Go, 1997). Depending on the research focus and the academic field, survival analysis can also be referred to as event history analysis, duration analysis, hazard modeling, reliability analysis, or transition analysis (Box-Steffensmeier & Jones, 2004; Ronco, 1995). The meaning of the term "survival" is also context dependent. For example, in the medical field "surviving" means a patient does not experience a death event. In the education field, if the outcome of interest is attrition, "surviving" means a student does not drop out. Survival analysis has several advantages over traditional regression methods. First, the analysis can include both uncensored and censored events (i.e., include both students for whom the event of interest occurs during the timeframe under investigation and students for whom it does not); second, it has the capacity to examine time-varying predictors (e.g., students' term GPA); and third, the analysis can determine the relative importance of predictors on outcomes of interest (Ronco, 1995).

Data Set

In this study, transfer students are defined as students who started work toward a degree program in one postsecondary institution and then transferred to a different postsecondary institution with the intention of completing their degree. The study institution is a large public R2 (high research activity) university located in the Southeast United States. Transfer students account for about one-third of entering undergraduate students each year, with the largest proportion coming from North Carolina community colleges. The study population comprised all new transfer students entering the institution in the Summer or Fall terms of 2010 to 2017. For each student, the data set included indicators of enrollment (enrolled or not enrolled) in each term following entry up to eight semesters. Students who were not enrolled in a term were counted as having dropped out even if they subsequently reenrolled within the years under investigation. Data were retrieved from the university's data warehouse by the university's institutional research staff, with queries written in SQL and SAS. A total of 11,267 students were included in the final analysis, with an average age of 25, average transfer GPA of 3.10, and an average of 57 credit hours transferred. Table 1 presents additional characteristics of these students.

Table 1. Summary of Student Characteristics (N = 11,267)

Characteristic	N	%
Entered with an associate's degree	4,136	37%
Enrolled full time in their first semester	8,341	74%
Entered as declared or intended STEM majors*	2,406	21%
Changed major during 1st year	1,862	17%
Any financial aid received	8,388	74%
Pell grant received	5,117	45%
Need-based aid received	8,211	73%
Merit-based aid received	473	4%
Female	6,134	54%
Underrepresented minority**	2,681	24%

** Note: As identified by the U.S. Department of Homeland Security (DHS), n.d.*
*** Includes American Indian/Alaska Native, Black/African American, Hispanic, Native Hawaiian or Other Pacific Islander, and Two or More Races.*

A total of eleven variables were included in the analysis: nine covariates, one time variable (the number of semesters a student was enrolled), and one outcome variable. The outcome variable was a binary variable that indicated whether students left the university without graduating after the last semester in which they were enrolled. Students who graduated or were still enrolled within the period under investigation were in one category (Persisted, coded as 0) while students who left the university and did not return were in another (Departed, coded as 1). Several of the covariates could be considered time variant (i.e., changed major, STEM major, financial aid received, and enrollment status). All covariates are described in Table 2.

Table 2. Covariate Descriptions

Variable	Definition
AGE_AT_MATRIC	Age at first enrollment
CHANGED_MAJOR	Whether a student changed major in Year 1 (Y, N)
STEM_MAJOR	Major at end of Year 1 is a STEM major (Y, N)
TRANSFER_CREDIT_HOURS	Total hours transferred in
TRANSFER_UG_GPA	Transfer undergraduate GPA
FIN_AID_RECEIVED	Received any financial aid Year 1 (Y, N)
FT_PT_Flag	Full-time or Part-time status in first term (FT, PT)
GENDER	Female / Male
URM*	URM / Non-URM

* Note that race/ethnicity was represented by a dichotomous variable of underrepresented minority (URM) or non-URM. We define URM as race/ethnicity categories that are underrepresented in our student body relative to their representation in the region.

DATA ANALYSIS

Data were analyzed using the statistical software R. More detail on the basics of survival analysis can be found on the webpage for STHDA (n.d.) and in PowerPoint slides made available online as part of a workshop titled "Introduction to Survival Analysis in R" (UCLA Office of Advanced Research Computing, Statistical Methods and Data Analytics, n.d.). For the first research question, the Kaplan-Meier curve was used to estimate and visualize survival probability from Semester 1 to Semester 8. The Kaplin-Meier curve graphically represents the survival function and shows the probability of an event at a given time interval. The x axis represents time—in our case, the number of semesters elapsed since entry. The y-axis presents the estimated survival rate. Kaplin-Meier allows for the inclusion of censored data (i.e., data on cases for which the event has not yet occurred). As mentioned previously, the ability to utilize censored data is one of the major advantages of survival analysis over other statistical techniques such as logistic regression. For student data, regression examines only whether a student had or had not experienced the event of interest (e.g., retention) at a particular point in time. Survival analysis, however, allows for including data on "censored" students (i.e., students for whom we do not know an outcome by the end of a specific time period but about whom we have data from within a given timeframe). For the second question, the Stratified Kaplan-Meier plot was used to estimate and visualize survival curves, and the log-rank test was used to compare whether there was a difference between the survival curves of the seven selected groups. Finally, the Cox proportional hazard model was used to examine the relationships between the covariates and transfer students' persistence. The log-rank test also helps in variable selection in the Cox proportional hazard model.

The fundamental assumption in the Cox model is that the hazards are proportional, which means that the effect of a covariate is constant over time. Violation of this assumption suggests that the effect of this covariate is time varying. In this study, the examination of the proportional hazards assumption was performed through examining Schoenfeld residuals plots. Proportional hazard is indicated by a horizontal line. To fit a Cox model with time-varying coefficients, we used both a continuous function and a step function.

RESULTS

Research Topic 1: Estimated Survival Rate of Transfer Students within Eight Semesters after Enrollment.

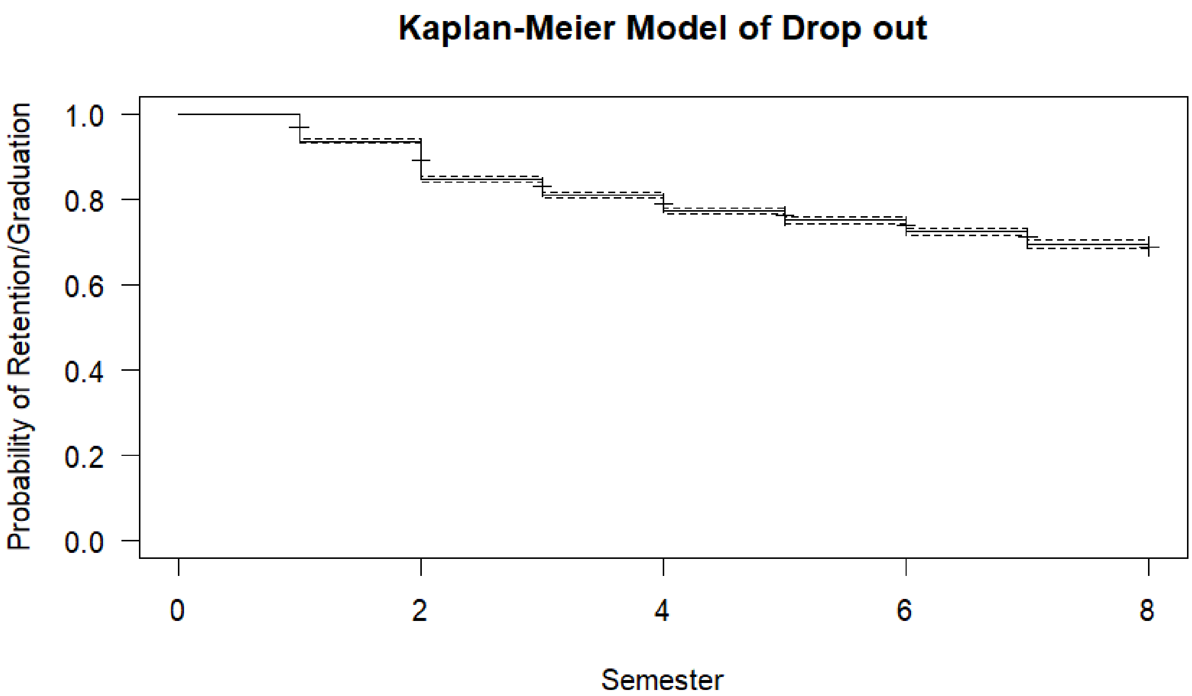
In survival analysis, the estimated survival probability represents the probability that a transfer student would persist after a given number of semesters. It was computed as the number of students who persisted after x semesters divided by the total number of students enrolled in the first semester. The scale is 0.00 to 1.00. Table 3 shows that the 95% confidence interval for the probability of a transfer student to persist after Semester 8 is between 67.1% and 69.2%. Figure 1 shows a Kaplan-Meier curve of the estimated persistence probabilities for all the transfer students in this study over eight semesters.

Table 3. Estimated Survival Probability and Hazard Rate by Number of Semesters Enrolled (N = 11,267)

Semester	Estimated survival probability	Lower 95% CI	Upper 95% CI	Hazard rate
1	0.936	0.932	0.941	0.064
2	0.847	0.840	0.854	0.095
3	0.810	0.802	0.817	0.044
4	0.772	0.764	0.780	0.047
5	0.751	0.743	0.759	0.027
6	0.724	0.716	0.733	0.036
7	0.695	0.686	0.705	0.040
8	0.681	0.671	0.692	0.020

Note: CI is confidence interval.

Figure 1. Kaplan-Meier Curve of the Estimated Persistence Probabilities

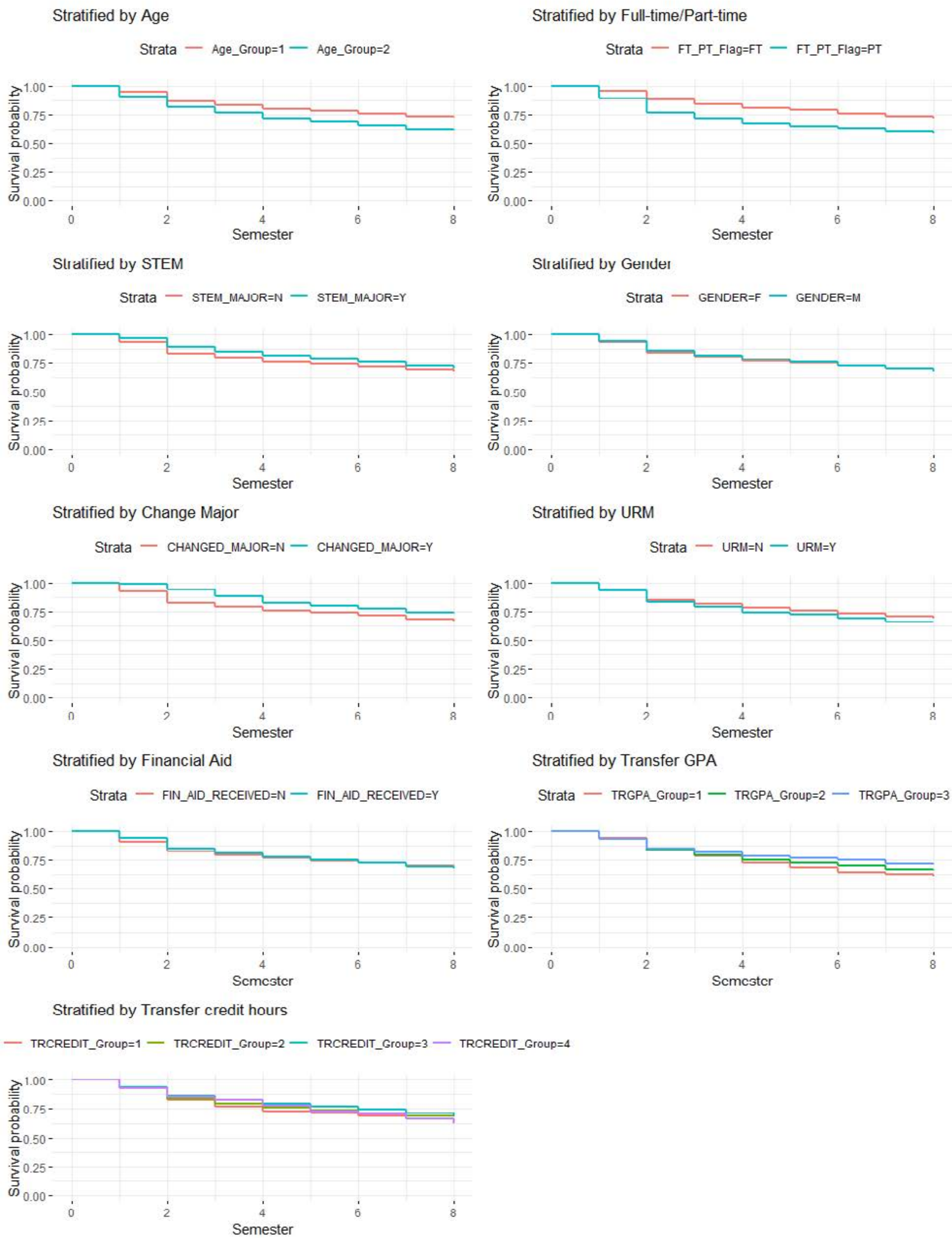


On the other hand, the hazard function is used to present the probability of an event occurrence at each period. In this study, we use the term "hazard rate" to refer to this probability. The hazard rates shown in Table 3 represent the probabilities that a transfer student would depart after x semesters. It was computed as the number of students who had departed after x semesters divided by the total number of students enrolled in x semester. The scale is thus 0.00 to 1.00. For example, for transfer students in this study, hazard rates were 0.064 and 0.095 after the first and second semesters of enrollment at the university, which are the two highest hazard rates among the eight semesters. After Semester 2 the hazard rates decreased and the changes in the rates were relatively small.

Research Topic 2: Survival Rates between Subgroups.

We used the stratified Kaplan-Meier method to estimate and visualize survival curves (see Figure 2) and the Gehan-Wilcoxon test to determine if there was a difference in the overall survival distributions between groups. The groups compared were based on age (24 and younger, over 24), majors (STEM, Non-STEM), change of major (Yes, No), transfer GPA (<2.5, ≥2.5 and <3, ≥3 and ≤4), transfer credit hours (<30, ≥30 and <60, ≥60 and <90, ≥90), financial aid (Yes, No), enrollment status (full time, part time), gender (female, male), and race/ethnicity (URM, Non-URM).

Figure 2. Stratified Kaplan-Meier Plot by Student Groups



The Kaplan-Meier plots for each group are presented in Figure 2. Gehan-Wilcoxon test results indicated that students who changed major were more likely to persist than those who did not change major (χ^2 (1)=37, *p*<.0001). Students who were in STEM majors were more likely to persist than those who were not in STEM majors (χ^2 (1)=19.1, *p*<.0001). URM students were more likely to drop out than non-URM students (χ^2 (1)=13, *p*=.0003). Students younger than 25 were more likely to persist than those above 25 years old (χ^2 (1)=124, *p*<.0001) and students with higher transfer GPAs were more likely to persist than those with lower transfer GPAs (χ^2 (2)=33.7, *p*<.0001). Full-time transfer students were more likely to persist than part-time students (χ^2 (1)=235, *p*<.0001). Survival distributions for the four transfer credit hours groups were significantly different (χ^2 (3)=18.6, *p*=.0003). Generally, the more transfer credits students brought in, the more likely they were to persist. However, after four semesters those students who had transferred in the highest number of credit hours were less, rather than more, likely to persist. Finally, no difference was found either between the survival rates of students who received financial aid and those who did not (χ^2 (1)=1.2, *p*=0.3), or between female students and male students (χ^2 (1)=0.7, *p*=0.4).

Research Topic 3: Size of the Effects of the Covariates on the Probability of Transfer Students' Retention and Graduation.

The Cox proportional hazard model was used to examine the effects of covariates on the probability of students' persistence. Covariates (see Table 2 for details) included age, major (STEM or not), major change, transfer GPA, transfer credit hours, financial aid, full-time or part-time status, gender, and race/ethnicity (coded as URM or Non-URM). The total sample size was 11,267 and the number of events (departures) was 3,080. A positive coefficient means lowered survival and a negative coefficient means increased survival.

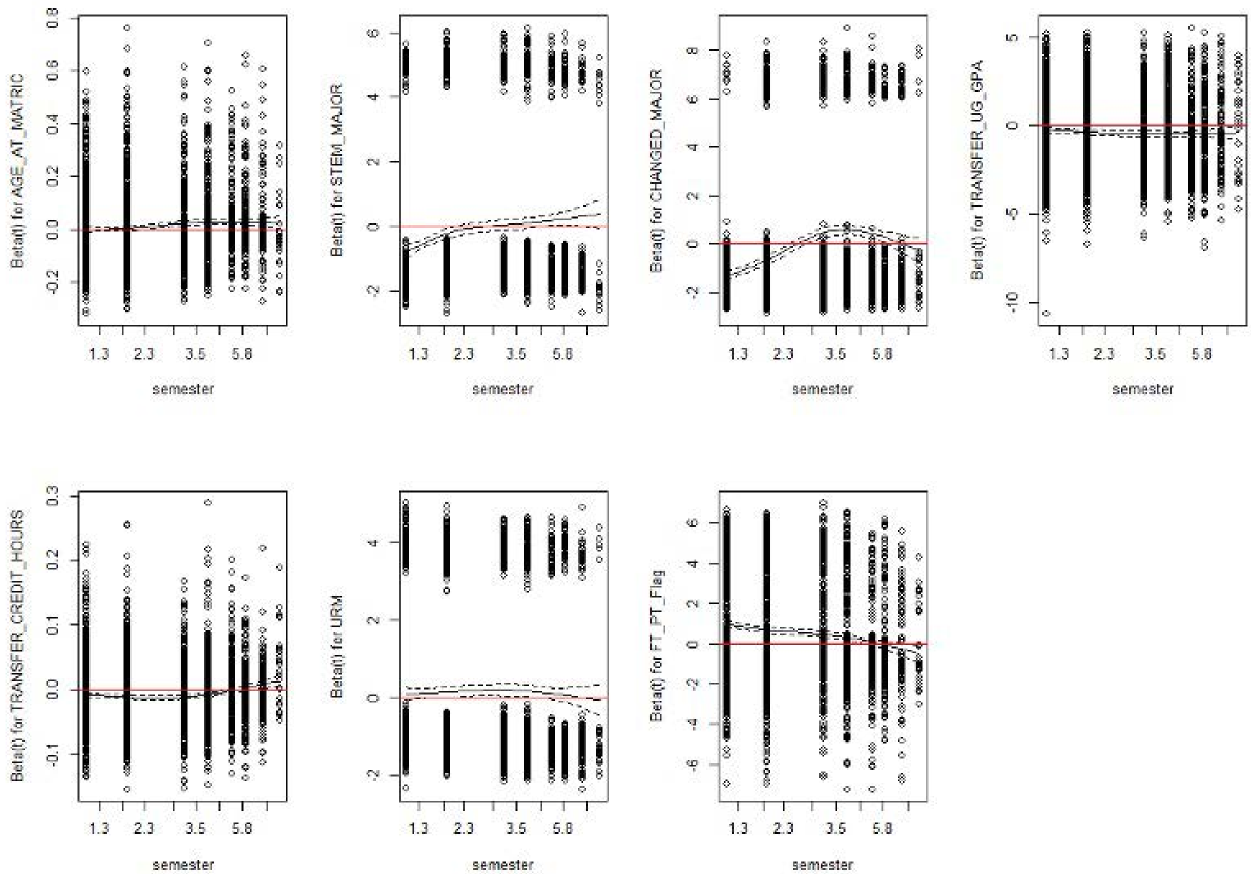
To optimize variable selection and check the fundamental assumption of the Cox model that the hazards were proportional, a Cox model with all covariables was run. We examined the p value and the Schoenfeld residuals plot of each variable. All covariates were significant at a .95 confidence level. Because the financial aid variable was not significant in the log-rank test of the stratified Kaplan-Meier model and had a much higher p value in the Cox model than all other variables, this variable was removed to optimize the model. The same was true for the gender variable. The results of the optimized model are presented in Table 4.

Table 4. Estimated Effects of Selected Variables on the Probability of Departure in the Optimized Cox Model

Term	Estimate	Std. Error	Statistic	p. Value	Exp (Estimate)
Age at Matriculation	0.014	0.003	5.641	<.001	1.014
STEM Major = Yes	–0.199	0.046	–4.290	<.001	0.820
Changed Major = Yes	–0.355	0.053	–6.646	<.001	0.701
UG GPA	–0.376	0.042	–8.917	<.001	0.686
Transfer Credit Hours	–0.008	0.001	–8.408	<.001	0.992
Part time = Yes	0.543	0.045	11.959	<.001	1.721
URM = Yes	0.134	0.042	3.212	<.01	1.143

The Schoenfeld residuals plots can help determine whether covariates are time varying. The plot of Schoenfeld residuals against time should not show a pattern of changing residuals for the covariate; that is, the smoothed plot should be flat and close to zero. If there is a pattern, that covariate is time dependent. Generally, a nonzero slope is an indication of a violation of the proportional hazard assumption. Figure 3 shows that the major (STEM vs. non-STEM) effect increased the first two semesters, then went flat in the 2nd year, and increased again after the 2nd year; the full-time/part-time effect constantly decreased over time; and the major change effect increased from matriculation to Semester 4 and then decreased. These variables obviously violate the proportional hazards assumption of the Cox proportional hazard regression since they are time varying. Thus, they should be investigated further.

Figure 3. Schoenfeld Residuals Plots



To fit a Cox model with time-varying coefficients, we used a stratified Cox proportional hazard model. The timeline was cut into three strata: Semesters 1 and 2, Semesters 3 and 4, and Semesters 5 through 8. We applied the strata function on STEM_MAJOR, CHANGED_MAJOR, and FT_PT_Flag covariates.
The estimated effects of selected variables on the probability of departure in the final Cox model are presented in Table 5. Note that the estimate of AGE_AT_MATRIC, TRANSFER_UG_GPA, TRANSFER_CREDIT_HOURS, and URM covariates are averaged over the strata, while the STEM_MAJOR, CHANGED_MAJOR, and FT_PT_Flag covariates have estimates on each stratum.

Table 5. Estimated Effects of Selected Variables on the Probability of Departure in the Final Cox Model

Term	Estimate	Std. Error	Statistic	p. Value	Exp (Estimate)
AGE_AT_MATRIC	0.014	0.003	5.685	<.001	1.014
TRANSFER_UG_GPA	–0.373	0.042	–8.822	<.001	0.689
TRANSFER_CREDIT_HOURS	–0.008	0.001	–8.371	<.001	0.992
URM = Yes	0.132	0.042	3.185	<.01	1.142
STEM_MAJOR: strata(tgroup) tgroup=1	–0.501	0.068	–7.404	<.001	0.606
STEM_MAJOR: strata(tgroup) tgroup=2	0.021	0.085	0.241	*	1.021
STEM_MAJOR: strata(tgroup) tgroup=3	0.295	0.099	2.994	<.05	1.344
CHANGED_MAJOR: strata(tgroup) tgroup=1	–1.154	0.099	–11.711	<.001	0.315
CHANGED_MAJOR: strata(tgroup) tgroup=2	0.363	0.082	4.399	<.001	1.437
CHANGED_MAJOR: strata(tgroup) tgroup=3	0.009	0.111	0.083	*	1.009
FT_PT_Flag: strata(tgroup) tgroup=1	0.673	0.055	12.172	<.001	1.961
FT_PT_Flag: strata(tgroup) tgroup=2	0.506	0.079	6.413	<.001	1.659
FT_PT_Flag: strata(tgroup) tgroup=3	0.1206	0.102	1.183	*	1.128

* Indicates the variable is not significant at $p < .05$.

The Exp. (Estimate) column in Table 5 is the back-transformed coefficient of the covariates of focus. It is similar to the odds ratio concept in logistic regression. If the value is greater than one, the chance of an event occurring increases; if the value is less than one, the chance of the event decreases. Results show that, assuming equality of other hazard factors, all factors in the model are statistically significant. Specifically, for each additional year of age at matriculation (at baseline), departure hazard increases by 1% on average. For each one credit hour brought in, departure hazard decreases by 0.7% on average. For each one point of transfer GPA increase at baseline, departure hazard decreases by 31% on average. Additionally, from matriculation to Semester 2, departure hazard for STEM students is 60% of that for non-STEM students; the departure probability of STEM students increases over time, however, as shown by the coefficient changing from negative to positive. After four semesters, STEM majors are 34% more likely to drop out than are non-STEM majors. In their first two semesters, part-time students are 96% more likely to drop out than are full-time students, but this effect constantly decreases over time as indicated by the coefficient of FT_PT_Flag: strata(tgroup) changing from 0.67 to 0.12; that is, the departure probability of part-time students gradually decreases. From Semester 3 to 4, they are 66% more likely to drop out. After four semesters, there is no difference in the drop-out probability of part- and full-time students. Finally, the effect of changing major is significant from matriculation to Semester 4: Students who change their major in Year 1 are 68% less likely to depart within Year 1 than are students who do not change majors; from Semester 3 to 4, however, students who change majors are 44% more likely to drop out than those who did not change majors. After four semesters, there is no difference in the drop-out probability of those who change majors and those who do not because the *p*-value is much greater than .05.

DISCUSSION AND CONCLUSIONS

Few studies have used survival analysis to study transfer student persistence. To fill the gap, this study used survival analysis to investigate the persistence of transfer students during their (initial) 4 years at a transfer university. Our findings reveal critical insights into the factors affecting the persistence of these students. The probability of students persisting stands at 0.681 with a 95% confidence interval (0.671, 0.692). When analyzing specific subgroups, we observed several significant trends. For instance, transfer students who changed majors after transferring, those majoring in STEM fields, individuals below 25 years, those with a higher transfer GPA, and those transferring more credit hours all displayed higher persistence rates. Moreover, full-time students and non-URM students also showed a higher likelihood of persistence.

These findings replicate those of several prior researchers but conflict with others. The impact on transfer student persistence of credit hours transferred (Hutton, 2015; Luo et al., 2007), enrollment status (Hutton, 2015; Ronco, 1995; Wang, Wang, et al., 2017), and age (Murtaugh et al., 1999) were replicated. However, our study showed no distinction between the survival rates of students based on gender or financial aid received, which diverges from previous studies. For instance, Wang (2009) found that females had a higher probability of completing a bachelor's degree, echoing Ronco (1995) and Wang, Wang, et al. (2017) in the context of survival analysis. The nonsignificance of financial aid in our study also contrasts with Hutton (2015) who suggested that financial aid could impact persistence, albeit marginally.

The study is limited in a few ways. For example, additional demographic and academic variables could have been considered in the model such as first-term or 1st-year GPA. Perhaps more importantly, because the focus of a survival analysis is the passage of time, other time-varying covariates could have been included, such as GPA per semester or credit hours earned per semester. There are two kinds of time-varying covariates, and future studies should include both. One type of covariate changes value over time, and also changes over time in its impact on the outcome variable; the other does not change value over time but its effect on the outcome variable changes over time. The current study included no covariates of the first type. Finally, using a simplistic financial aid variable (FIN_AID_RECEIVED,Y, N) may have accounted for this variable not reaching significance and being omitted from the model. An alternative way of coding financial aid (e.g., as a continuous variable) or having more than one financial aid variable might have altered the results. Further research is warranted to help identify reasons for conflicting findings and to provide additional support for our assertion that survival analysis is a useful tool in understanding what factors help or hinder transfer student success.

There are implications in the current study for fostering transfer student success, defined as persistence in college. Providing academic and other types of support targeted specifically to transfer students and adult learners may prove beneficial. Advising students to maximize the number of transfer hours applied toward their intended major would likely improve graduation rates for these students. Advisors can work with transfer students to create clear academic pathways that consider their specific profiles, such as age, transfer credits, and ability to take a full-time course load. Our findings indicate that the impact of some variables changes over time, so interventions could be targeted at specific semesters to optimize impact. It is also the case that 4-year institutions might benefit from closer collaborations with community colleges, ensuring smoother academic transitions, aligning curricula, and providing shared resources and support for students.

With the declining number of high school graduates (Bransberger et al., 2020) and, thus declining number of first-time-in-college students, the recommendations above and other best practices in transfer student success (see, e.g., Smith et al., 2021) are important focal points for institutions wishing to maintain enrollment in the coming years. Fall 2023 data from the National Student Clearinghouse Research Center (NSCRC) show the number of new students who transferred into a new institution grew 5.3% compared to Fall 2022, with transfers representing 13.2% of all continuing and returning undergraduates (NSCRC, 2024). As the landscape of higher education continues to evolve, with shifting demographics and enrollment patterns, it becomes increasingly important to address the challenges faced by transfer students. The research presented in this study highlights the need for proactive measures to bridge the completion gap and to ensure that transfer students have equitable opportunities to attain their educational goals. Institutions that prioritize the success of transfer students will be better positioned to adapt to the changing educational landscape and maintain robust enrollment in the years to come.

REFERENCES

Allensworth, E. M., & Clark, K. (2020). High school GPAs and ACT scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher, 49*(3), 198–211.

American Council on Education. (2021). *Reimagining transfer for student success.* The National Task Force on the Transfer and Award of Credit. https://www.acenet.edu/Documents/Reimagining-Transfer-for-Student-Success.pdf

Aulck, L., & West, J. (2017). Attrition and performance of community college transfers. *PLoS ONE 12*(4): e0174683. https://doi.org/10.1371/journal.pone.0174683

Barbera, S. A., Berkshire, S. D., Boronat, C. B., & Kennedy, M. H. (2017). Review of undergraduate student retention and graduation since 2010: Patterns, predictions, and recommendations for 2020. *Journal of College Student Retention: Research, Theory, & Practice, 22*(2), 227–250.

Baum, S., Ma, J., & Payea, K. (2013). *Education pays: The benefits of higher education for individuals and society.* College Board. https://research.collegeboard.org/trends/education-pays

Box-Steffensmeier, J. M., & Jones, B. S. (2004). *Event history modeling: A guide for social scientists.* Cambridge University Press.

Bransberger, P., Falkenstern, C., & Lane, P. (2020). *Knocking at the college door* (10th ed.). Western Interstate Commission for Higher Education. https://www.wiche.edu/resources/knocking-at-the-college-door-10th-edition/

Caruth, G. D. (2018). Student engagement, retention, and motivation: Assessing academic success in today's college students. *Journal of Higher Education Theory and Practice, 18*(1), 41–63.

Chimka, J. R., Reed-Rhoads, T., & Barker, K. (2007). Proportional hazards models of graduation. *Journal of College Student Retention: Research, Theory & Practice, 9*(2), 221–232.

Choudhury, A. R., & Runco, M. (2020). Testing the effect of UNIV1000 on retention in a regional university in the US. *Journal of Education and Learning, 9*(5), 198–204.

Demeter, E., Dorodchi, M., Al-Hossami, E., Benedict, A., Slattery Walker, L., & Smail, J. (2022). Predicting first-time-in-college students' degree completion outcomes. *Higher Education, 84*(3), 589–609.

Daddona, M. F., Mondie-Milner, C., & Goodson, J. (2021). Transfer student resources: Keeping students once they enroll. *Journal of College Student Retention: Research, Theory, & Practice, 23*(3), 487–506.

Gardner, J. N., Rosenberg, M. J., & Koch, A. K. (Eds.). (2021). *The transfer experience: A handbook for creating a more equitable and successful postsecondary system.* Taylor & Francis Group.

Hayward, C. (2011). The transfer velocity project: A comprehensive look at the transfer function. *Journal of Applied Research in the Community College, 18*(2), 17–28.

Hutton, R. (2015). *Modeling bachelor's degree completion rates of North Carolina community college transfer students using a discrete-time survival analysis.* [Unpublished doctoral dissertation]. University of North Carolina at Chapel Hill.

Jenkins, D., & Fink, J. (2016). *Tracking transfer: New measures of institutional and state effectiveness in helping community college students attain bachelor's degrees.* [Report]. Community College Research Center (CCRC), Columbia University. https://ccrc.tc.columbia.edu/media/k2/attachments/tracking-transfer-institutional-state-effectiveness.pdf

Kopko, E. M., & Crosta, P. M. (2016). Should community college students earn an associate degree before transferring to a 4-year institution? *Research in Higher Education, 57*, 190–222.

Lee, E. T., & Go, O. T. (1997). Survival analysis in public health research. *Annual Review of Public Health, 18*(1), 105–134.

Lichtenberger, E., & Dietrich, C. (2017). The community college penalty? Examining the bachelor's completion rates of community college transfer students as a function of time. *Community College Review, 45*(1), 3–32.

Lin, Y., Fay, M. P., & Fink, J. (2020). *Stratified trajectories: Charting equity gaps in program pathways among community college students.* [Working Paper No. 126]. Community College Research Center (CCRC), Columbia University. https://ccrc.tc.columbia.edu/publications/stratified-trajectories-program-pathways.html

Luo, M., Williams, J. E., & Vieweg, B. (2007). Transitioning transfer students: Interactive factors that influence first-year retention. *College and University, 83*(2), 8.

Marbouti, F., Ulas, J., & Wang, C.-H. (2021). Academic and demographic cluster analysis of engineering student success. *IEEE Transactions on Education, 64*(3), 261–266.

McNeish, D., Dumas, D. G., & Grimm, K. J. (2020). Estimating new quantities from longitudinal test scores to improve forecasts of future performance. *Multivariate Behavioral Research, 55*(6), 894–909.

Miller, J. W., & Lesik, S. S. (2014). College persistence over time and participation in a first-year seminar. *Journal of College Student Retention: Research, Theory & Practice, 16*(3), 373–390.

Monaghan, D. B., & Attewell, P. (2015). The community college route to the bachelor's degree. *Educational Evaluation and Policy Analysis, 37*(1), 70–91.

Mourad, R., & Hong, J.-H. (2008). Application of survival analysis to study timing and probability of outcome attainment by a community college student cohort. *Journal of Applied Research in the Community College, 16*(1), 19–28.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education, 40*, 355–371.

National Student Clearinghouse Research Center (NSCRC). (2024, February). *Transfer and progress: Fall 2023 report.* Transfer and Progress, National Student Clearinghouse Research Center. https://nscresearchcenter.org/transfer-and-progress/

Reynolds, J., & Cruise, S. (2020). Factors that influence persistence among undergraduate students: An analysis of the impact of socioeconomic status and first-generation students. *Interchange, 51*, 199–206.

Ronco, S. L. (1995, May). *How enrollment ends: Analyzing the correlates of student graduation, transfer and dropout with a competing risks model.* [Conference session]. 35th Annual Forum of the Association for Institutional Research, Boston, MA.

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Yuan, X., Nathan, A., & Hwang, Y. (2017, September). *Tracking transfer: Measures of effectiveness in helping community college students to complete bachelor's degrees.* [Signature Report No. 13]. National Student Clearinghouse.

Smith, A., Sturtevant, K., Nango, C., & Bullough, E. (2021). *Transfer students: Retention and persistence.* Fulton Library, Utah Valley University. https://www.uvu.edu/library/docs/transfer_students_retention_and_persistence.pdf

Statistical Tools for High-Throughput Data Analysis (STHDA). (n.d.). *Survival analysis.* http://www.sthda.com/english/wiki/wiki.php?title=survival-analysis

Taplin, C. (2019). *Examining the factors that influence the success of students who transfer to a four-year public research university.* [Unpublished doctoral dissertation]. State University of New York at Buffalo.

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

UCLA Office of Advanced Research Computing, Statistical Methods and Data Analytics. (n.d.). *Introduction to survival analysis in R.* https://stats.oarc.ucla.edu/r/seminars/introduction-to-survival-analysis-in-r/

U.S. Department of Homeland Security (DHS). (n.d.). DHS STEM designated degree program list. https://www.ice.gov/sites/default/files/documents/stem-list.pdf

Voelkle, M. C., & Sander, N. (2008). University dropout: A structural equation approach to discrete-time survival analysis. *Journal of Individual Differences, 29*(3), 134.

Wang, X. (2009). Baccalaureate attainment and college persistence of community college transfer students at four-year institutions. *Research in Higher Education, 50*(6), 570–588.

Wang, X., Chuang, Y., & McCready, B. (2017). The effect of earning an associate degree on community college transfer students' performance and success at four-year institutions. *Teachers College Record, 119*(2), 1–30.

Wang, X., Wang, Y., Wickersham, K., Sun, N., & Chan, H. (2017). Math requirement fulfillment and educational success of community college students: A matter of when. *Community College Review, 45*(2), 99–118.

Wright, C. (2018). *Choose wisely: A study of college major choice and major switching* (Publication No. 10987539). [Doctoral dissertation, Pardee RAND Graduate School]. ProQuest Dissertations & Theses Global.

Yang, Y., Briggs, K., Avalos, S., & Anderson, C.M. (2018). Examining incoming credit differences between first-year and transfer students. *NACADA Journal, 38*(2), 31–44.

Zhang, Y. L. (2022). Early academic momentum: Factors contributing to community college transfer students' STEM degree attainment. *Journal of College Student Retention: Research, Theory, & Practice, 23*(4), 873–902.

ABOUT THE AIR PROFESSIONAL FILE

The AIR Professional File features journal-length articles grounded in relevant literature that synthesize current issues, present new processes or models, or share practical applications related to institutional research. All submissions are peer-reviewed. For more information about AIR Publications, including *The AIR Professional File* and instructions for authors, visit www.airweb.org/publications.

ABOUT AIR

The Association for Institutional Research (AIR) empowers higher education professionals to leverage data, analytics, information, and evidence to make decisions and take actions that benefit students and institutions and improve higher education. For over 50 years, AIR has been committed to building and sustaining strong data-informed decision cultures within higher education institutions.

Association for Institutional Research
Christine M. Keller, Executive Director and CEO
www.airweb.org
ISSN 2155-7535

THANK YOU

Authors who submit materials to AIR Publications receive feedback from members of the AIR community. AIR expresses sincere appreciation for the individuals who serve as peer reviewers. The following individuals reviewed manuscripts recently submitted to *The AIR Professional File*.

Tracey Baham

Stefano Fiorini

Ivana Fredotovic

Robynn Shannon

Bob Wilkinson

Henry Zheng

Linli Zhou