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

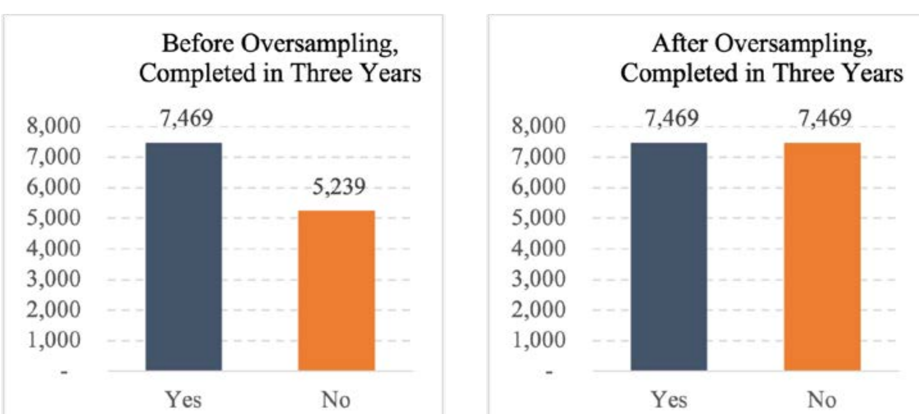
We applied five machine learning models (random forest, decision tree, extreme gradient boosting [XGBoost], gradient boosting, AdaBoost) and one traditional model (logistic regression) to identify the most appropriate model with the highest predictive power of a master's student degree completion. Decision 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?

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

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

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

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

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

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.

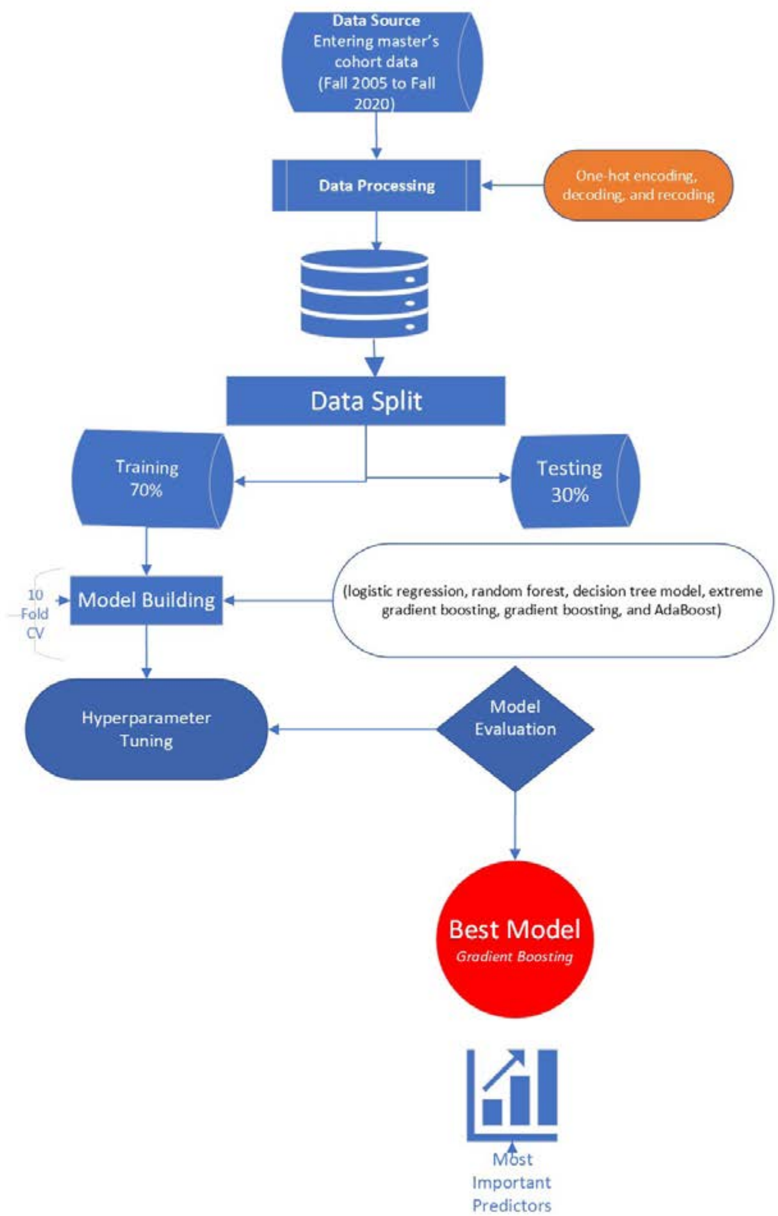
$$\textit{F1 score} = 2 \times \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

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



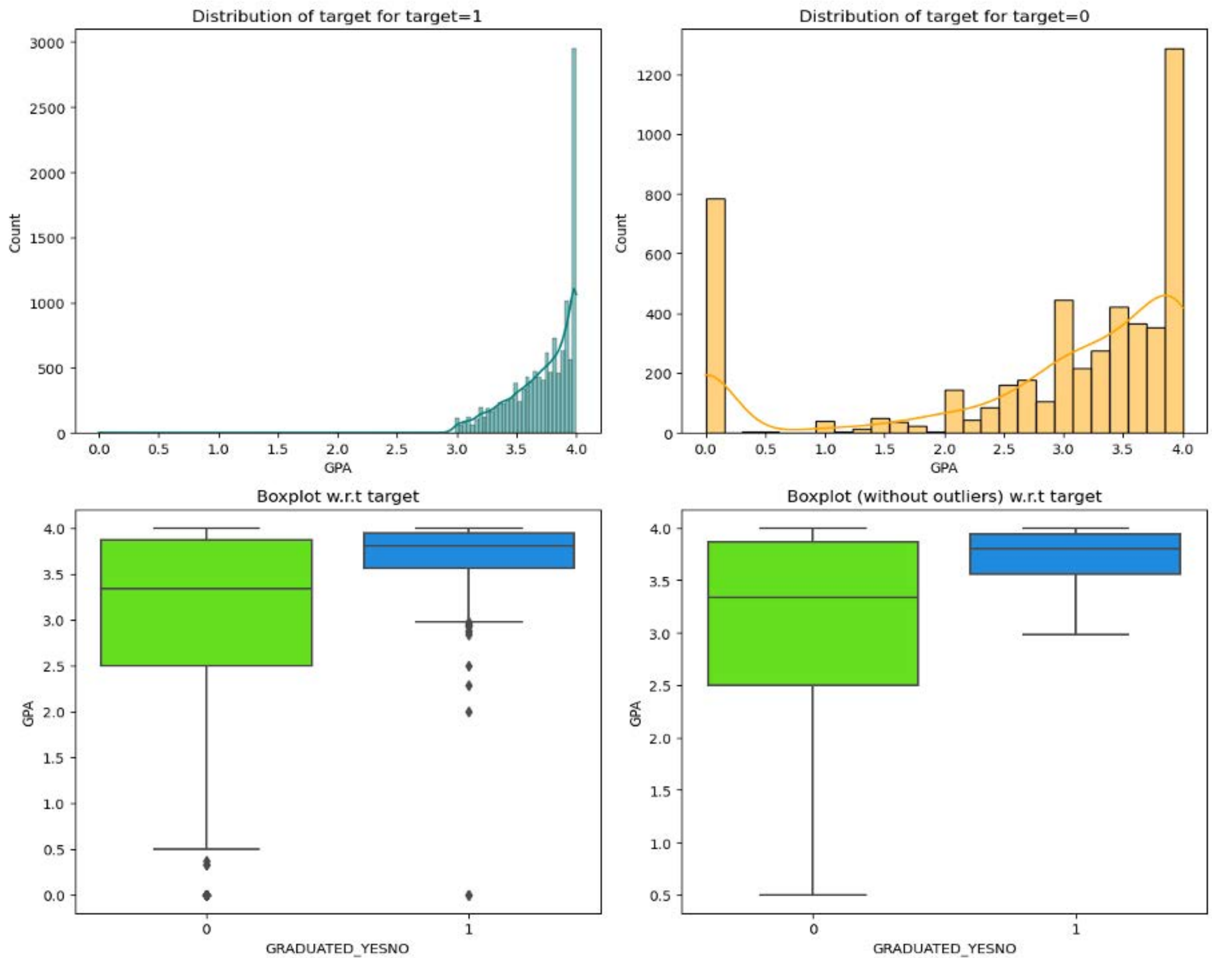
RESULTS

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.

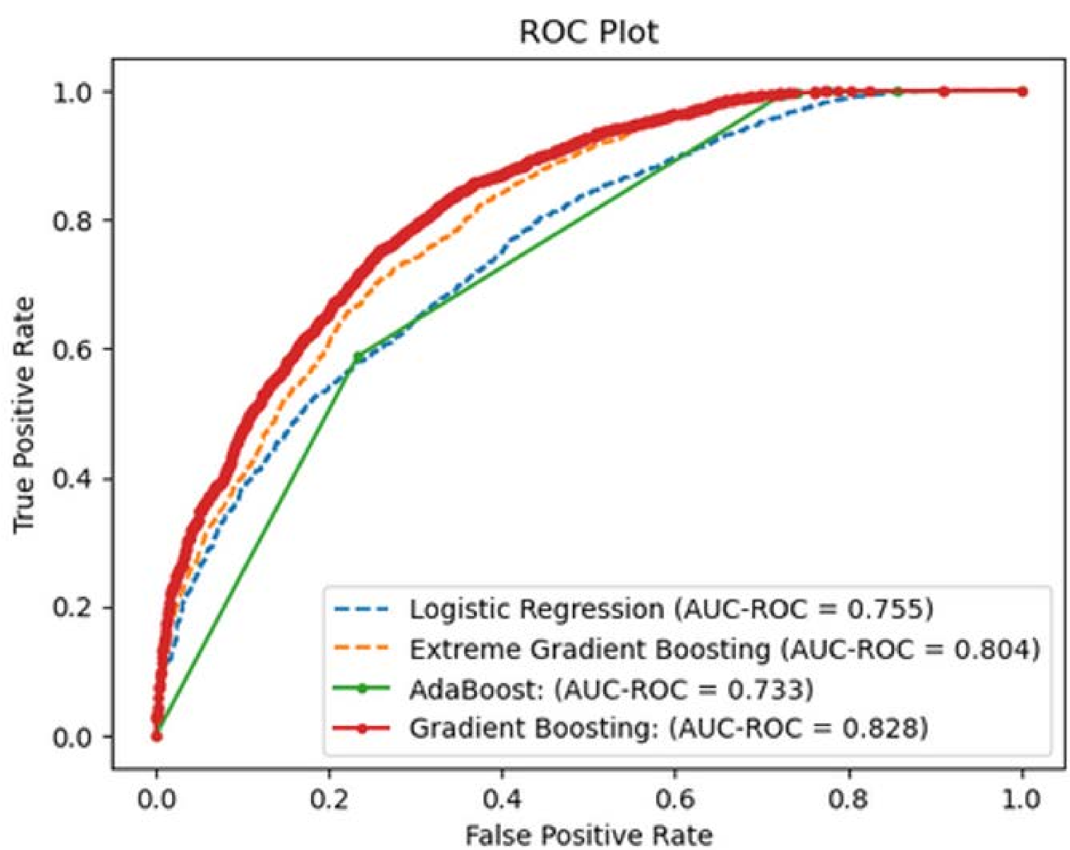
There are several evaluation metrics we can use to adjudicate between the remaining models. The accuracy metric is best used when we are interested in correctly predicting both completions and non-completions. For example, the gradient boosting model correctly predicted student degree completion outcomes 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

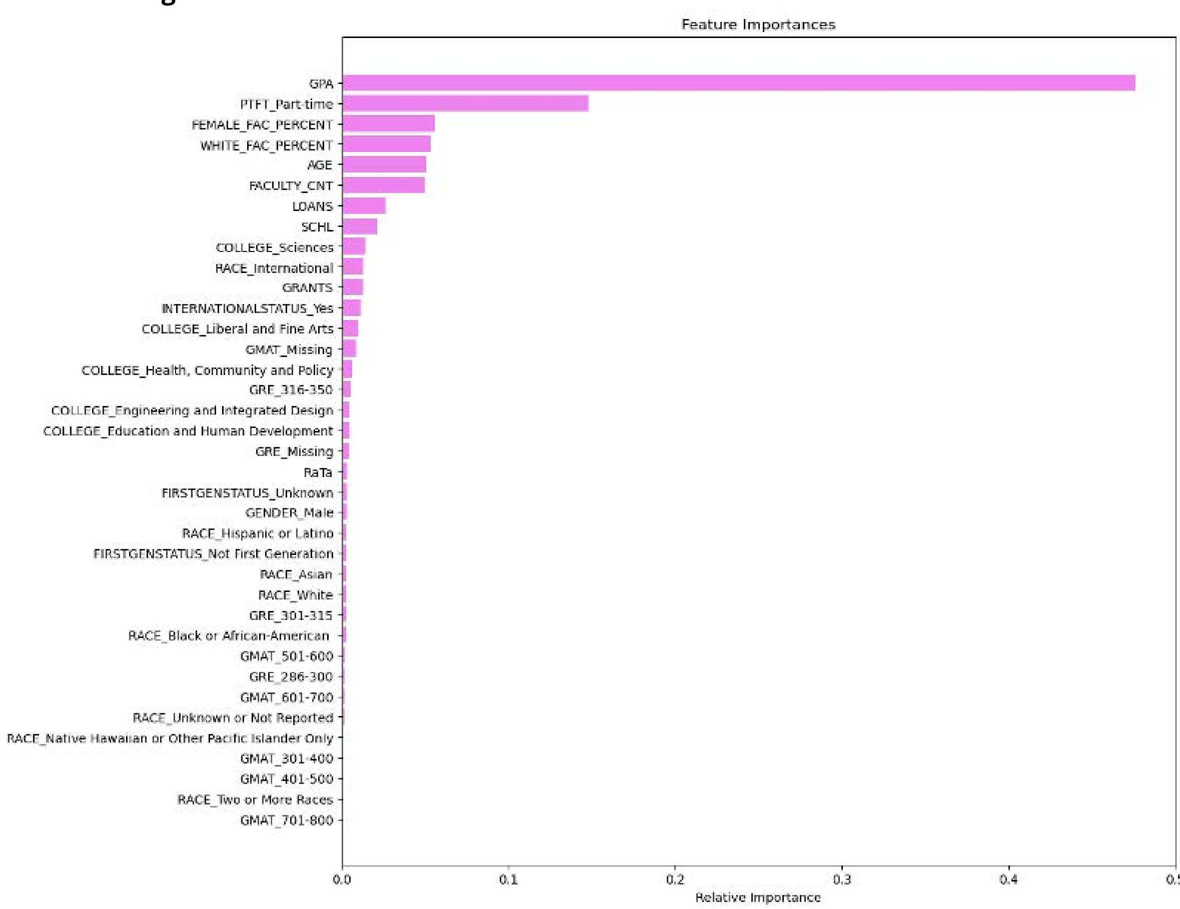


Source: 15 years of entering master's cohort data from UTSA.

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

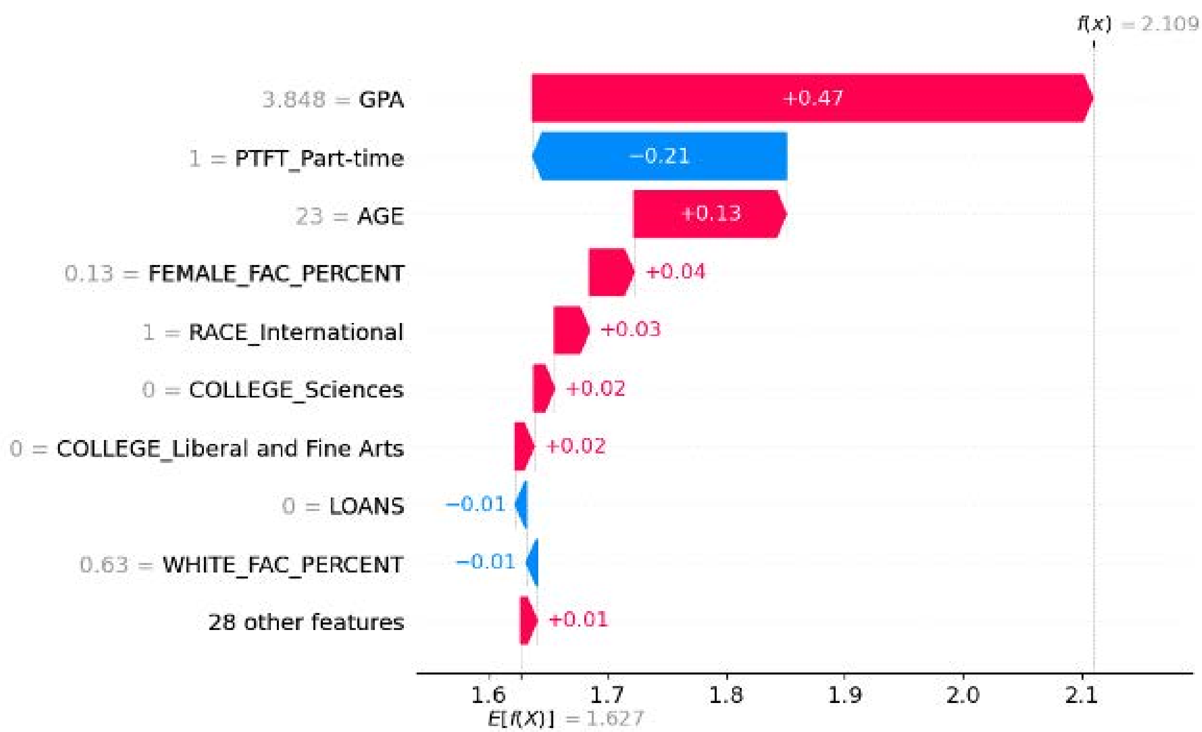


Source: 15 years of entering master's cohort data from UTSA.

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