2-2023

Balancing Inference and Prediction in Institutional Research: A Practical Comparison of Logistic Regression With Machine Learning Techniques in Modeling Student Persistence

Alison Weingarten
The Graduate Center, City University of New York

BALANCING INFERENCE AND PREDICTION IN INSTITUTIONAL RESEARCH: A PRACTICAL COMPARISON OF LOGISTIC REGRESSION WITH MACHINE LEARNING TECHNIQUES IN MODELING STUDENT PERSISTENCE

by

ALISON WEINGARTEN

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2023

APPROVAL

Balancing Inference and Prediction in Institutional Research: A Practical Comparison of Logistic Regression with Machine Learning Techniques in Modeling Student Persistence

by

Alison Weingarten

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Approved: December 2022

Jay Verkuilen, Chair of Examining Committee

Joan Lucariello, Executive Officer

Supervisory Committee:

David Rindskopf

Paul Attewell

Joan Lucariello

Nancy Floyd

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Balancing Inference and Prediction in Institutional Research: A Practical Comparison of Logistic Regression with Machine Learning Techniques in Modeling Student Persistence

by

Alison Weingarten

Advisor: Jay Verkuilen

In higher education research, causal inference has traditionally been the focus over predictive power, with statistical models designed to understand and explain the relationships between variables. In the field of institutional research in particular, there is a growing need to not only understand these causal relationships, but also predict what is likely to occur in the future (Which students are most likely to succeed at our college, and who should we admit? Which students are we most likely to lose to attrition, and how can we engage them? Which students are most likely to struggle academically, and what interventions can we provide?). While many institutional researchers are adept in statistical analysis, machine learning methods—widely touted as being more nimble and powerful at making predictions—are still relatively untapped in the field.

Using a standard institutional research dataset from a large public urban university system, this study compared the efficacy of conventional logistic regression with several machine learning classification techniques on predicting secondary educational outcomes. The analysis found that theory-based logistic regression performed similarly overall to the machine

learning methods, though the types of predictions made by each model varied. A discussion about the practical use of these methods for institutional researchers follows.

ACKNOWLEDGEMENTS

My greatest thanks to Jay for all your guidance, support, and patience over the years; I could not have asked for a better advisor to see me through this degree. Thank you as well to the dedicated members of my committee, whose valuable feedback and insights helped shape this dissertation.

CONTENTS

TABLES

For Dad, Tom, and Patrick

And for my mother, who would be so proud

Introduction

Institutional research is a function in higher education that deals broadly with the analysis and reporting of institutional data. Most postsecondary institutions in the US, and particularly ones that accept federal funding, have an institutional research office or function. The breadth of an IR office's responsibilities can vary widely among institutions, but often includes gathering and organizing internal data; administering surveys; conducting analyses to support administrative decision-making and inform policy development; assisting with efforts in planning, budgeting, and institutional effectiveness; and reporting institutional data to external agencies for compliance, funding, or benchmarking purposes.

As the field has matured over the last few decades, so has the role of institutional researchers. IR professionals are researchers skilled in wrangling large datasets and performing complex statistical analyses. They regularly need to generate outputs and communicate findings that are compelling to stakeholders who are not technically or statistically savvy. Institutional researchers may interact or collaborate with almost every area of an institution. In analyzing student performance, institutional researchers generally focus on measures of academic success, such as GPA, retention, and graduation, since maximizing these measures benefits both students and the institutions. As such, institutional researchers are often focused on which variables affect these outcomes.

Degree completion is arguably the most important outcome in higher education, to both students and the institution itself. Students attend college to broaden their horizons, grow intellectually, and prepare for careers. Earning a postsecondary degree can have a significant impact on their futures. For young adults (25-34 years old), the median earnings of those with a bachelor’s degree is 63% higher than those with a high school diploma or GED (Irwin et al.

2022), and for many career paths, a college education is considered essential. Most colleges and universities in the United States have a mission of providing students with a quality education, and most are also externally measured by their success in retaining and graduating their students. Several external ratings are published each year that affect the public's perception of how a school is performing, and these ratings can have a strong effect on the number and caliber of students who apply to a college each year.

As research on the topic of student persistence has grown, so too has the understanding that each institution must tailor its retention efforts to fit the needs of its students. Efforts to increase degree completion are typically measures of prediction or intervention. Prediction involves estimating which students are most at risk of dropping out, and intervention involves the ways in which an institution can improve students' chances of graduating. To address problems of prediction, logistic regression is the method of choice for many institutional researchers, due in large part to its perceived interpretability. For classification problems, logistic regression produces outputs that show the relationship between the predictor and outcome variables along with weights to give an indication of the importance of each variable, providing the researcher with results that are both explainable and relatively accessible to a wide audience.

Machine learning is a subset of computer science in which algorithms learn from prior experience to make future predictions. While statistical models are used to identify and understand relationships within data, machine learning was developed to discover patterns within very large or unstructured datasets. As data are amassed in ever increasing amounts, it becomes more difficult to explore a dataset and identify patterns using traditional statistics alone. For problems that involve massive datasets or data with extreme dimensionality, such as the data

common in fields like astronomy or biology, machine learning is often able to identify trends where the limits of human understanding have been exhausted.

Generally speaking, a researcher employing conventional statistical methods fits a model to a particular dataset and views the results as an explanation of that dataset, often with the goal of generalizing to a broader population. This process frequently begins with a set of a priori assumptions which can occasionally lead the researcher to disregard variables which could result in meaningful insights. Machine learning algorithms use a portion of a dataset to learn patterns, then test what they have learned against the rest of the data. The goal is to find a best-fitting function which is predictive of patterns in future datasets. Machine learning is traditionally used on massive datasets, because larger datasets are more conducive to pattern recognition and are more resistant to overfitting.

Much has been written about how machine learning may help improve on traditional statistical methods, particularly for problems focused on predicting future results or behaviors, and colleges could benefit greatly from the ability to make more accurate predictions about their students. Institutions that can identify which students are more likely to struggle academically are better positioned to provide those students with appropriate interventions, allowing those colleges to improve outcomes while more efficiently allocating their resources.

Colleges amass an ever-increasing wealth of data on their students—pre-college data are collected during the application process, measures of student satisfaction and experience are solicited through surveys, performance metrics are gathered through grades and test scores, level of engagement can be approximated through attendance and involvement in clubs, activities, and learning communities—but the extent to which they make use of this data varies. Quite often, it remains stored in student information systems without being utilized for analysis and

exploration. Machine learning, with its ability to identify trends in large, high-dimensional datasets, could help institutions make use of this untapped trove of information.

Additionally, machine learning methods may help to save time. Regression methods, considered standard practice by many institutional researchers, can require a time-consuming model-building process involving transforming predictors, creating interaction variables, testing for multicollinearity, and narrowing down the set of predictors for optimum model selection. In contrast, machine learning techniques, many of which are far more flexible and less susceptible to these issues, may be more efficient to implement.

Although the field of machine learning has been maturing for decades, it is still relatively new compared with traditional statistics, particularly in the social sciences. Academia, a field that is known for being resistant to change and slow to adapt, has unsurprisingly not widely explored the improvements machine learning may provide. Within institutional research, this may be due to several factors. For one, there is often a trade-off in interpretability that comes with more flexible methods, as will be discussed in more depth later. Institutional researchers must explain the results of their models to various stakeholders who are often concerned with understanding the relationships between variables, rather than just learning the outputs. Audiences who are accustomed to seeing model outputs and regression coefficients may find the lack of such results unsettling or untrustworthy. In some cases, there are even legal or compliance-related reasons to prefer models with interpretable outputs. In assessing pay equity, for example, it is necessary that the model outputs be auditable in order to examine potential sources of bias. Another reason machine learning has not caught on in institutional research may be due to its perceived impenetrability. While institutional research practitioners are often adept at conducting analyses using traditional statistical methods, for many machine learning may

appear too daunting or time-consuming to adopt. Among the reasons for this are the ostensible notion that all machine learning requires "big" data, the perceived lack of explainability of the outputs, and the relatively steep learning curve (in order to get started, one must learn new concepts and applications, learn how to code in new programming languages, understand the assumptions and mechanisms of the new methods, and learn how to interpret and use the results).

Purpose of Study

This paper aims to explore the utility of several machine learning methods for use by institutional researchers specifically for the problem of predicting student persistence. The reason for this focus is threefold. Most importantly, student persistence is a very common issue for institutional researchers to face, and it is one of considerable salience. Improving student retention and graduation is a goal for most colleges in the United States, and it is one that has a great impact on students and institutions alike. Second, this is a topic that appears to be particularly well-suited for machine learning. Currently, the most widely accepted method for examining student persistence among institutional researchers is logistic regression. There certainly remains a need to understand and explain student behavior, but on most campuses this is mainly an issue of prediction—to predict which students are at risk of attrition or to identify students in need of intervention—and it may be one for which newer methods have the potential to offer significant improvement. And finally, the effort to study and improve student persistence is not a new phenomenon to be investigated. This is an issue that is grounded in decades of established theory and research and is consequently relatively well-understood. Therefore, this paper does not intend to provide new explanations or contribute to the broader understanding of

student persistence. Rather, the purpose of this paper is to explore possible improvements to a widely accepted solution for a problem commonly encountered in institutional research.

For the purpose of this paper, *statistics*, *traditional statistics*, or *conventional statistics* will refer to the general concept of currently accepted statistical methods used most commonly in academia. Unless stated otherwise, *logistic regression* refers to multiple logistic regression using traditional statistical methods in particular. *Machine learning algorithm* refers to the program used to learn a model from a set of data. *Persistence* refers to the action of a student remaining within a particular college or school from their first year through degree completion. Throughout this paper, persistence may be used to refer to either retention (re-enrollment to the following year or semester of study) or graduation (successful completion of a degree or credential). *Attrition* refers to a student's failure to re-enroll at a particular college in consecutive semesters.

Literature Review

The subject of student persistence in higher education has been studied thoroughly for decades. The most widely recognized theory for this research is Vincent Tinto's integration model, which provides a theoretical framework for understanding student persistence. According to the theory, student retention relies on the combination of student characteristics and their academic and social integration within college. The theory argues that pre-college characteristics, academic success, and early engagement and commitment in college have the greatest impact on retention. Tinto's theory proposes that students must be able to integrate into all aspects of academic and social college life in order to persist in their education (Tinto 1993). William Spady and John Bean's theories also emphasize the importance of integration into both the academic and social systems students encounter at college. Bean's theory suggests that the factors which affect retention have more to do with the college itself, and that students may leave for similar reasons employees leave an organization. Spady's theory also addresses the importance of the interaction between the individual student and the college he or she attends and suggests that attrition is due in large part to poor academic performance and lack of social support (Burke 2019). Alexander Astin found that grades in high school, educational aspirations, study habits, and parental education are most predictive of retention and concluded that the more involved a student is in academic and social life at college, the more likely they are to persist (Astin 1985).

Colleges collect a myriad of information on their students while they are enrolled, but information about students before they arrive at college may also prove valuable to the overall understanding of their education pathways. Tinto (1993) has emphasized the importance of collecting information on students entering college to adequately assess retention programs.

Several pre-college characteristics have been shown to be significant predictors of retention at the college level, including parental education and income, students' educational goals and aspirations before entering college, ratings of self-confidence and self-efficacy, high school grades, and standardized test scores (Astin & Oseguera 2012). Institutional researchers can use pre-college student engagement data to help their colleagues better understand student backgrounds, experiences, and expectations, to design appropriate first-year programs to help ease the transition into college, to prepare academic advisors to better guide their students, and to help faculty adjust their curricula and teaching practices (Cole et al. 2009). Cole & Korkmaz (2010) demonstrated this by using data from the Beginning College Survey of Student Engagement, which asks students entering their first year of college about their high school engagement and experiences, as well as their expectations about college; data from the National Survey of Student Engagement, which asks students about their college experience in the spring of their first and senior years; pre-college data collected during the admissions process; and data collected while the students were enrolled in college. Using regression models and correlation studies, the authors showed several examples of how pre-college data can be used to enhance postsecondary educators' understanding of their incoming students' backgrounds and needs and increase their chances for success (Cole & Korkmaz 2010).

While machine learning techniques are not yet common among institutional researchers, they have been shown to be useful for educational data applications. In particular, several studies have compared the efficacy of various machine learning algorithms to logistic regression. In comparing a logistic regression model with three decision tree methods and three neural network methods, Herzog (2006) found that the optimal method depended on the complexity of the data and the outcome being predicted—for a well-understood outcome such as freshmen retention,

the decision trees and neural networks did not substantially outperform the logistic regression model, but for the more complicated outcome of time to degree completion, the data mining algorithms were markedly more effective than the logistic regression model. González and DesJardins (2002) compared two artificial neural network models (one trained using continuous variables and one using categorical dummy variables) with a logistic regression model and found that both ANNs were more effective at predicting which students would apply to college. However, the authors went on to lament the limitations of ANNs, including the relative lack of interpretability and the time consuming process of running the analysis (one ANN model took many hours to run—although it is worth noting that the training time required for neural network methods has decreased greatly since this paper was published), and suggested that using ANN and logistic regression in tandem may be a better approach for institutional researchers. Thomas & Galambos (2004) made a similar suggestion after comparing logistic regression to a decision tree method, the chi-squared automatic interaction detector (CHAID) algorithm, to predict measures of student satisfaction. They found that while the regression analysis identified variables that helped explain variance in student satisfaction, the decision tree analysis helped identify which elements of the college experience best differentiated between satisfied and dissatisfied students and revealed patterns in the data that were unidentifiable with regression analysis alone. Luan (2002) compared neural networks with two decision tree methods (C5.0 and CART) to predict which community college students would transfer to a four-year institution and found that CART was most effective at predicting which students would transfer and which would not. Lingjun et al. (2018) compared logistic regression to random forests and a standard CART model to predict both continuous and categorical higher education outcomes, and found that the machine learning algorithms, while comparable to one another in their performance, both

outperformed logistic regression in all aspects. The authors recommend random forest over CART and logistic regression for institutional researchers, based on its ease of implementation, flexibility, interpretability, and relatively low computational cost. Random forest may also be more appropriate than logistic regression for complicated models that include non-linear and interaction terms (Lingjun et al. 2018). Liu et al. (2021) compared logistic regression, support vector classification, decision tree, extreme gradient boosting, and random forest to predict measures of student learning in higher education. The authors found that logistic regression slightly outperformed the machine learning classifiers. Smirani et al. (2022) used a combination of three ensemble classifiers (light gradient boosting machine, extreme gradient boosting machine, and random forest) to predict student performance in higher education. The authors suggest that machine learning methods were particularly appropriate for fully online or hybrid learning environments because of the large amount of data captured by learning management systems.

Because much of machine learning is based on statistics, the two fields are intrinsically linked, and the boundary between them is hazy and porous. While it is difficult to draw a clear line of distinction, some practitioners in both fields argue that they diverge in their differing assumptions, applications, and goals. The true relationship between input and response variables (in nature and statistics alike) is an unknowable black box. Traditional statistical methods often rely on a data model and parameter estimates to attempt to explain what occurs within this black box. In contrast, machine learning uses optimization algorithms to determine an underlying function with many parameters that best maps the input to the response, essentially bypassing the interpretation of the internals of the black box and focusing solely on prediction (Breiman 2001).

Causal relationships have long been the chief purpose of statistical methods in academic research, with predictive power often considered an ancillary product. Some academics consider strictly prediction-focused methodologies unscientific or unacademic, while some machine learning practitioners point out that traditional statistical methods are rigid and rely on assumptions that are often not true to real life (Breiman 2001; James et al. 2013; Shmueli 2010). In the social sciences, statistical models are almost always used for causal inference, and it is often assumed that models that are high in explanatory power are also inherently useful for making predictions. In contrast, fields that make more use of machine learning techniques often focus almost exclusively on prediction, treating causal explanation as something of a byproduct (Kuhn & Johnson 2013; Shmueli 2010). Many statistical models are used to make predictions as well as provide explanations for phenomena, but in general predictive accuracy is not their strength. If a researcher's main interest is inference, then traditional statistical models may seem preferable, as they are far more interpretable. Conversely, machine learning models, which tend to sacrifice interpretability in favor of predictive power, can produce such complicated estimates of a model function that it can be difficult to comprehend any individual predictor's effect on the response (James et al. 2013).

The difficulties associated with uninterpretable models have not been lost on the machine learning community, and as a result there have been many recent developments to help improve these shortcomings. One approach is to train a surrogate model using a more interpretable method with the goal of approximating the predictions made by the uninterpretable model. For example, after training a random forest, one might train a decision tree on the same dataset but use the random forest's predictions as the outcome variable instead of the known outcome classes. This approach is adaptable and approachable, because it can easily be applied to various

methods and it is simple to implement, explain and understand (Molnar 2019). There are also a number of more advanced techniques that aim to explain the effect of individual predictors. Two common approaches are local interpretable model-agnostic explanations (LIME) and Shapely additive explanations (SHAP). Instead of creating a surrogate model for the entire uninterpretable machine learning model, LIME trains local surrogate models to approximate individual predictions, while SHAP estimates the average contribution each input variable has on the overall model (Lundberg & Lee 2017; Ribeiro et al. 2016).

In statistics and machine learning, error is determined by bias and variance. Bias refers to the tendency for a model to produce incorrect results in a consistent way (i.e., too high or too low). In contrast, variance refers to how greatly a model's parameters tend to deviate when applied to different datasets. Other things being equal, when we increase a model's bias, we decrease its variance (Yarkoni & Westfall 2017; James et al. 2013; Hastie et al. 2017). With explanation as the goal of social scientists utilizing traditional statistics, it is favorable to minimize bias over variance. However, minimizing bias at the expense of variance sacrifices generalizability to new datasets, and in many cases even removes the ability for the model to function when new samples are drawn from the same population. In machine learning, the preference is to find the ratio of bias to variance that best minimizes the prediction error (Yarkoni & Westfall 2017).

Machine learning can be supervised or unsupervised. Supervised machine learning is used when each input into a model is known and can be labeled and referred to beforehand. In supervised learning tasks, an algorithm is trained using data for which the inputs and outcomes are already known. The model learns by comparing what it produces to the correct outcomes to find errors and modifies itself accordingly (Aggarwal 2016; James et al. 2013). Supervised

machine learning algorithms are very useful for tasks that use historical data to predict future events, like which students are likely to drop out of school. For supervised learning tasks, the dataset is generally divided into two parts: a training set to train the algorithm, and a testing set to test the performance of the trained algorithm (James et al. 2013; Hastie et al. 2017). In unsupervised learning, the model is only given the input data, without any explicit labels or information about the data. In unsupervised tasks, the model must crawl through a dataset to find the structure and relationships hidden within (James et al. 2013). For the purposes of this paper, we will focus on supervised machine learning techniques, which are more relevant to the types of issues most frequently encountered by institutional research practitioners.

Overview of Common Classification Techniques

There are many machine learning techniques and algorithms available for use, and some are better suited for certain purposes than others. Below is a brief overview of several machine learning methods commonly used for classification problems.

K-Nearest Neighbors

K-Nearest Neighbors is a non-parametric method used for regression and classification problems. The researcher chooses *k*, and the algorithm looks within the "neighborhood" of each observation to determine the *k* nearest neighbors, then averages across them all to classify each data point (James et al. 2013; Kuhn & Johnson 2013). For classification problems, the algorithm uses majority rule to determine to which class an observation belongs. For regression problems, it averages the neighboring observations to calculate an output. It may appear that with a large enough training dataset, one could always approximate an optimal estimation using k-nearest

neighbors, because one would assume there should always be a large enough "neighborhood" of observations close to *x* to average across. However, in a concept known as the curse of dimensionality, this logic does not hold true in datasets with many predictors (Hastie et al. 2017).

Shrinkage Methods

Penalized regression techniques, like ridge, lasso, or elastic net regression, are useful for high-dimensional models where there are a large number of variables relative to the sample size. A model with too many predictors is penalized by shrinking the value of the coefficients of less important variables to (or towards) zero, which helps to reduce variance (Efron & Hastie 2016; James et al. 2013).

Ridge regression assigns a penalty to regression coefficients according to their size, which reduces the effect of some variables without removing them from the model. This leads to lower variance and therefore an overall lower error rate. Lasso regression shrinks the coefficients of some variables down to zero, eliminating some features entirely and resulting in a subset of optimal predictors. Elastic net regression shrinks some coefficients and sets others to zero, effectively combining both the ridge and lasso methods (Efron & Hastie 2016; James et al. 2013; Hastie et al. 2017).

Decision Trees

Decision trees, also referred to under the umbrella term Classification and Regression Trees (CART), are non-parametric methods used for classification or regression problems. The model predicts the value of the outcome variable by learning decision rules from the data and splits the dataset at a number of nodes, to create "branches" of the tree. In CART, the predictor space is divided into a number of non-overlapping regions, and for each observation within a

region, the prediction equals the mean of the response values for that region. Because it is not feasible to define every possible region, the algorithm uses a top-down, greedy approach. Top-down refers to starting with the entire dataset (the top of the tree), and greedy refers to the fact that the best split is made at each step of the process, regardless of whether there may be a future split that will lead to a better tree (James et al. 2013; Hastie et al. 2017). Because they can be prone to overfitting, decision trees are often pruned to remove non-critical sections of the tree and minimize error. CART-based methods are not restricted by the assumptions of linearity that regression methods are bound to, so they are less susceptible to issues of multicollinearity and can implicitly address interactions between variables. Additionally, because they select decision rules on individual predictors, tree-based methods do not struggle with large numbers of predictors the way logistic regression does (Lingjun et al. 2018). Decision trees are very easy to interpret and explain, can be displayed graphically and understood by non-experts, and can handle qualitative predictors, which are used very often in institutional research. However, they do not have as much predictive accuracy as more advanced tree-based methods and are not robust—a small change in the data can result in a large change in the final tree.

Tree-Based Ensemble Methods

Decision trees are simple to implement and interpret, but their prediction accuracy can not compete with some of the more advanced supervised machine learning methods. However, combining a large number of trees can result in considerable improvements to prediction accuracy at the expense of the interpretability of the simpler methods (James et al. 2013).

Bagging. Decision trees can suffer from high variance (e.g., if we split our dataset in half and trained each on a decision tree, the results could be quite different). Bagging, short for "bootstrap aggregating," is a procedure to help reduce the variance of a statistical learning

method, and it is often used with decision trees. Based on the idea that averaging a set of observations reduces variance, bagging takes many bootstrapped training sets from the population, builds separate prediction models using each set, and averages the resulting predictions. Bootstrapping here refers to the method of iteratively resampling the original dataset "with replacement," meaning an observation may appear more than once in each sample and the covariance between each sample is zero. The trees are allowed to "grow" deep and are not pruned, so each tree has high variance and low bias. Averaging across all of the trees then reduces the overall variance (Hastie et al. 2017; James et al. 2013; Kuhn & Johnson 2013).

Boosting and BART. Boosting is similar to bagging, however instead of creating trees that are independent from one another, boosting creates trees that are grown sequentially. Each new tree uses information from previously grown trees and is fit to the residuals from the previous model, rather than the outcome variable, as the response. This allows each tree to fit some data variation not explained by the trees grown before it (James et al. 2013; Chipman et al. 2010; Hastie et al. 2017). Extreme Gradient Boosting (XGBoost) is a boosting method that allows for additional hyperparameters which can help alleviate the concern of overfitting (Boehmke & Greenwell 2020).

Bayesian Additive Regression Trees (BART) is a similar method to boosting, but it relies on an underlying probability model rather than a pure algorithm. The underlying model consists of a set of prior probability assumptions that act as regulators to prevent any single tree from dominating the overall fit (Chipman et al. 2010).

Random forests. Similar to bagging, random forests build a number of decision trees on bootstrapped training samples. However, while building trees for a random forest, a random

sample of predictors is chosen as candidates from the full set of predictors at each node of the tree. With bagging, the decision trees may all be highly correlated, which does not lead to as large a reduction in variance as if we were to average many uncorrelated trees. By only considering a subset of the predictors at each split, the average of the trees in a random forest is less variable and more reliable than the average of a set of bagged trees (James et al. 2013; Lingjun et al. 2018; Hastie et al. 2017).

Naive Bayes

Naive Bayes is a collection of algorithms used for binary or multi-class classification. It operates under the assumption that all features are conditionally independent, though this is rarely true in real-life circumstances (hence “naive”). Because naive Bayes assumes conditional independence among features, it is not very strong at handling correlated predictors. Still, naive Bayes classifiers tend to be highly effective for classification problems. The technique estimates the probability that each observation belongs to a particular class by aggregating the probabilities of each of its features belonging to that class. While individual probability estimates may be biased, the final estimates are able to withstand individual biases with the amount of variance saved by the “naive” assumptions the model makes. Naive Bayes is also well-suited for small training samples with high dimensionality, which is why it is very commonly used for text classification (Hastie et al. 2017).

Support Vector Machines

Support Vector Machines were developed for classification by the computer science community in the late 1990s and are often considered one of the best "out of the box" classifier techniques (James et al. 2013). The term "support vector machines" can refer to the maximal

margin classifier, the support vector classifier, and the support vector machine. The maximal margin classifier is the separating hyperplane that is farthest from the training observations. To find it, we compute the perpendicular distance from each observation in the training set to a given hyperplane that may separate the observation classes, and the smallest of these distances is the margin (Efron & Hastie 2016; James et al. 2013). The maximal margin hyperplane is the separating hyperplane that has the farthest minimum distance to the observations in the training set—or more simply, the center line of the widest slice of space that can be inserted between the two classes (James et al. 2013). Observations can then be classified based on which side of the maximal marginal hyperplane they fall. Any observations that lie along the margin itself are known as "support vectors" because they are vectors in p-dimensional space which "support" the maximal margin hyperplane, in the sense that if these points were to move at all, the maximal margin hyperplane would move as well (James et al. 2013; Kuhn & Johnson 2013). Of course, there may not always be a hyperplane that perfectly separates the classes, so there may not be a maximal margin classifier. When this occurs, we can create a hyperplane that uses a "soft margin" to separate the cases. This practice of generalizing the maximal margin classifier to the non-divisible case is known as the support vector classifier. Instead of attempting to find the largest possible margin between classes so that every observation is both on the correct side of the hyperplane and outside of the margin, we allow some observations to violate the boundaries and fall on the wrong side of the margin or even the hyperplane entirely (James et al. 2013; Wang & Lin 2016). The support vector classifier works well for two-class classification when the boundary between the two classes is linear, but for non-linear class boundaries it is not useful. We can address this by enlarging the feature space using kernels—the process of transforming linearly inseparable data into a higher-dimensional space in which it becomes

separable—and the resulting classifier is known as a support vector machine (James et al. 2013; Breiman 2001).

MARS

Linear models can be extended to represent non-linear relationships by including polynomial terms. An alternative to this is to use step functions, which break down the range of values a variable can take into bins, and then apply the mean of the variable as a constant across each group. The resulting shape of the function approximates a non-linear relationship. However, with a highly dimensional dataset, it can be difficult to identify where the cut points should be made for step functions to be effective (Boehmke & Greenwell 2020).

Multivariate Adaptive Regression Splines (MARS) helps alleviate this issue by assessing the appropriate cut points, referred to as knots, for each range of data (Friedman 1991; Boehmke & Greenwell 2020). The procedure examines each data point and identifies the location where the change in the linear relationship between X and Y shows the smallest error term, and more knots can be added to achieve a better-fitting function (Friedman 1991; Boehmke & Greenwell 2020; James et al. 2013).

Methodology

The following research questions guided this study:

1. Which of the tested methods provides the most accurate predictions of student persistence?
    a. How well do the machine learning methods perform “out of the box”? How does parameter tuning boost performance?
2. Which of these methods is most appropriate for use by institutional researchers?
    a. How interpretable or explainable are the results of each method? How well-aligned is each model to institutional research problems?

To address these questions, the present study used data from a large urban public university system. The dataset was provided by the university system’s institutional research office and included many variables commonly used and reported on by institutional research offices, such as student characteristics, demographics, pre-college academic achievements, and college-level performance and outcomes. The university system includes 25 campuses, 14 of which offer 4-year degrees. Of these 14 colleges, one was selected for inclusion in this study because of its size and demographic representation.

The population included 28,709 students entering the selected college as first-time freshmen between fall 2000 and fall 2014. The one-year retention rate for this population is 86% on average across all years, and the six-year bachelor’s graduation rate is 56%. The sample population is made up of 37% white, 30% Asian or Pacific Islander, 20% Hispanic, 13% Black, and 0.1% American Indian or Native Alaskan students.

The types of variables examined were demographics and student characteristics (Pell/TAP, veteran, first-gen, gender, ethnicity, college choice, age, major), pre-college

performance (HS GPA, standardized test scores), and college performance (GPA, remedial tests passed). The outcome variables were one-year retention and six-year graduation. The variables used to predict one-year retention included those related to pre-college academic preparation, student background and demographics, and first-semester college performance. The same variables were used to predict six-year graduation, with the addition of further college performance measures. The theory-based logistic regression model included fewer variables, selected using a combination of theory-based rationale and model fit metrics.

To avoid multicollinearity, which occurs when two variables measure the same construct, variance inflation factor (VIF) was assessed for all variables. Correlations were also examined, and strongly correlated variables or variables with a VIF over 5 were removed from the analysis. To assess how to handle potential missing data, descriptive statistics were calculated to determine the extent and type of the missingness. Because the percent of missing data were relatively small (<8% for any variable in both outcomes), and because the missing data were not concentrated in a particular outcome class, a KNN imputation was performed on the data during preprocessing. Additionally, all nonbinary variables were scaled. Scaling data is useful for many machine learning algorithms because it standardizes the data across a single range of values. The `MinMaxScaler()` function was used to scale all non-binary variables between 0-1.

To properly evaluate the models, each one must be trained, and then it must be tested on a set of previously unseen data. In preparing the dataset, minimal data manipulation or pre-processing was conducted, because some transformations can have an undesired impact on model learning. For example, centering variables or imputing missing data requires information from the entire dataset. If this is done prior to splitting the dataset into a training and testing set, the training set may inadvertently learn from data contained in the testing set. After basic data

cleaning and recoding were performed, the dataset for each outcome variable was randomly split into a 70% training set and a 30% testing set. To avoid contamination of the testing set, the split occurred prior to further preprocessing steps. For both outcome variables, the dataset was preprocessed using the same technique, and the same preprocessed dataset was used for each machine learning method in the study. A separate dataset, using the same preprocessing methods but fewer variables, was used for the manual logistic regression models.

While machine learning was designed for use with large, high-dimensional datasets, some amount of feature selection or dimensionality reduction may still be beneficial. Many machine learning algorithms can suffer from overfitting, which can be exacerbated by a large number of predictors. A very large number of predictors can also take a computational toll, requiring more time and resources to train a model. Nonetheless, there were several reasons feature selection was decided against for this particular study. Many machine learning algorithms, including the ones selected for comparison in this study, either have some amount of inherent feature selection included or are able to function fully even with many predictors. Additionally, the datasets used by institutional researchers are generally not considered “big data” and are not nearly as wide as the datasets common in many machine learning applications. Considering this, declaring feature selection as a prerequisite to using these methods was determined to be an unnecessary barrier to adoption.

Cross-validation is a method which can be used to determine how well a model will generalize to an unseen dataset. The most simple form of cross-validation involves randomly splitting a dataset into a training and testing set, allowing the model to learn patterns using the training set and to test its predictions on a brand new set of data. This method works well for large sample sizes, but on smaller datasets there may not be enough data to simply split the

existing data into two sets. In these cases, there are several methods of cross-validation that can be used to augment a small dataset by holding portions of the original dataset aside as a testing set and using the remaining data to train the model. One well-known example is k-fold cross-validation, which splits the dataset into *k* subsets of approximately equal size, performs the training process on *k-1* subsets, and tests on the remaining subset; the process is repeated *k* times, until each subset has been used for testing. In the present study, the sample size was sufficiently large as to not require these additional cross-validation methods. However, they are still recommended for smaller datasets, such as those which are common in institutional research.

Cross-validation measures can also be used to identify optimal hyperparameters for machine learning models. When paired with a grid or random search across hyperparameter values, cross-validation methods can be used to select the hyperparameters which will perform best for the task at hand. In the present study, cross-validation for hyperparameter tuning was performed for two of the algorithms, but was ultimately discarded as the procedure was determined to be prohibitively computationally expensive. One procedure regularly took upwards of 200 minutes to complete, and the other took an average of 88 minutes. Importantly, the improvements to the models with tuned hyperparameters compared with out of the box models was minimal. As a result, cross-validation for hyperparameter tuning was omitted for the remainder of this study, and all results shown are for out of the box models.

Of the machine learning methods described in the previous section, elastic net regularization, decision trees, random forests, XGBoost, MARS, and support vector machines were used in this study, as well as a comparison with a logistic regression machine learning algorithm. The remaining methods were discarded for various reasons: K-nearest neighbors tends to perform poorly on datasets with much dimensionality, and naive Bayes is most often used for

text classification or for multi-class classification problems. Of the penalized regression methods, elastic net was selected because it combines features of both lasso and ridge regression methods. Decision trees were selected because, while not known to be the most robust predictors, they have some of the most interpretable outputs of any machine learning technique. XGBoost and random forests were selected because they both show improvements over the predictive accuracy of bagging alone.

To implement each method, the following algorithms in Python were used: machine learning logistic regression using `LogisticRegression` alone and in combination with the elastic net parameter, both from the scikit-learn library; decision trees using `DecisionTreeClassifier` from scikit-learn, an optimized version of the CART (Classification and Regression Trees) algorithm which constructs binary trees using the predictors and thresholds that provide the most information at each node; random forests using `RandomForestClassifier` from scikit-learn, an estimator that fits decision trees on random samples of the dataset and averages them to improve their predictive accuracy; extreme gradient boosting using `XGBoost`; MARS using `PyEarth`; and support vector machines using `SVC` from scikit-learn.

Results

A quantitative study using logistic regression and several machine learning classification algorithms was conducted to assess the strength of each method in predicting student outcomes (one-year retention and six-year graduation). Because the purpose of this study is to compare the efficacies of various methods and not to contribute to the broader body of research about factors impacting higher education outcomes, the discussion of results is limited to the evaluation metrics for the models used.

There are a number of methods for assessing model fit and predictive power, and several were examined for this study. When performing binary classification predictions, there are four types of outcomes that can occur. Cases correctly identified as belonging to a class (true positives), cases correctly identified as not belonging to a class (true negatives), cases incorrectly identified as belonging to a class (false positives), and cases incorrectly identified as not belonging to a class (false negatives). Classification accuracy is the ratio of correct predictions to all predictions made. Precision is the percent of true positives out of all cases identified as belonging to a class (true positives + false positives). Recall is the percent of true positives out of all cases which truly belong to the class (true positives + false negatives). Precision and recall can be combined into a more balanced metric known as the F1-score (Sokolova & Lapalme 2009). The models were also assessed using graphs of the Reception Operating Characteristic (ROC) curve and the area under the ROC curve (AUC), which show an aggregated measure of how well the classifier performed by plotting the true positive rate against the false positive rate at all thresholds.

As stated previously, only 14% of students are not retained each year on average, creating an imbalance in outcome classes. With imbalanced classes, prediction models are apt to

misclassify the minority outcome class in favor of an overall low misclassification rate. In situations of imbalanced classes, it is important to focus on particular evaluation metrics. For this study, we will compare all methods using the AUC, as it measures the true positive rate against the false positive rate at all thresholds and can therefore provide a more unbiased aggregate evaluation metric. However, as AUC is an aggregate measure without regard for probability threshold, it is important to consider other metrics as well, including specificity (true negative rate) and sensitivity (true positive rate) in particular. For the models in this study to be most useful, they must identify the maximum number of true negatives (students who are predicted to be and who actually are at risk of dropping out), while minimizing the number of false positives (students who are incorrectly predicted to be retained or graduate, and who may have benefitted from intervention). However, to avoid wasting resources, the models should also ideally not inflate the number of false negatives (students who are identified as needing extra support but do not require it).

For classification problems with imbalanced classes, it can be more difficult to assess the model using the selected evaluation metrics, which tend to favor the majority class and in some cases are calculated in reference to the positive class. For example, precision and recall are both calculated with the number of true positives as the numerator, and only specificity is dominantly concerned with negative cases. Because the positive class so highly outweighs the negative class in the first outcome variable of this study, most of the evaluation metrics appeared to perform very well simply by predicting the majority of cases as retained. One method of adjusting a model's ability to correctly classify cases is to use an optimized threshold value for classification probability. By default, the classification threshold is set to 50%, meaning a model that predicts a student to have a 51% chance of graduating would assign it to the positive (graduated) class. To

help the models perform more accurately, an optimized threshold was identified by using each model's ROC curve to ensure the most appropriate cutoff point. The optimized threshold was defined as the point on the ROC curve where sensitivity and specificity are closest to the value of AUC and where the difference between sensitivity and specificity is smallest. This adjustment provided more balanced predictions across all models, raising the specificity (which is of chief concern to this application) at the expense of other measures. This was particularly beneficial for the highly imbalanced first outcome variable, which saw specificity increase from an average of 0.198 to 0.580, and saw a reasonable decrease in sensitivity from 0.986 to 0.805. Because AUC is an aggregate measure across the entire ROC curve, it was not impacted by this change.

Outcome 1: One-year retention

Table 1 shows the model metrics for each of the methods tested for the first outcome variable. It is clear that the machine learning algorithms do not wildly outperform the theory-based model; in fact, the model metrics for the theory-based model outperform several of the machine learning models on AUC, and the results are highly similar across the board. Comparing the theory-based model to the logistic regression algorithm model—two essentially identical models save for the number of variables used—suggests that the presence of more variables may alone account for much of the model improvement in the machine learning algorithms.

Table 1. One-Year Retention Outcome Summary								
Evaluation Measure	"Manual" Logistic Regression	Algorithm Logistic Regression	Algorithm Logistic Regression w/ Elastic Net	Decision Tree	Random Forest	XGBoost	MARS	Support Vector Machine
Number of Variables	20	59	59	59	59	59	59	59
Accuracy	0.719	0.746	0.745	0.840	0.793	0.744	0.749	0.823
Precision	0.925	0.929	0.928	0.914	0.917	0.916	0.927	0.916
Recall/Sensitivity	0.732	0.762	0.762	0.898	0.833	0.772	0.767	0.874
Specificity	0.641	0.649	0.646	0.492	0.549	0.572	0.639	0.516
F1 Score	0.817	0.838	0.837	0.906	0.873	0.838	0.840	0.895
AUC:	0.746	0.766	0.766	0.749	0.749	0.730	0.759	0.733

Of the methods tested, machine learning logistic regression, theory-based logistic regression, and MARS had the highest measures of specificity, allowing them to correctly predict students who would not be retained to their second year. Decision tree, random forest, XGBoost and SVM had much lower measures of specificity. Using machine learning logistic regression versus a decision tree in this case would result in an additional 193 students being correctly identified as in need of intervention. However, the decision tree method had a much higher measure of sensitivity, which compared with traditional logistic regression resulted in 1,225 fewer students incorrectly identified as needing intervention. Of the methods tested, the two with the highest F1 scores are Decision tree and SVM. These two methods correctly predicted the most students who will retain to the following year while incorrectly predicting the

fewest students who are not in need of intervention. However, both methods overlook the most students who are in need of intervention.

Outcome 2: Six-year graduation

Table 2 shows the model metrics for each of the methods tested for the second outcome variable. Like the first tested outcome, the machine learning algorithms perform similarly to the theory-based model across all measures. Though all of the models performed within range of each other, random forest had the highest AUC measure.

Table 2. Six-Year Graduation Outcome Summary

Evaluation Measure	"Manual" Logistic Regression	Algorithm Logistic Regression	Algorithm Logistic Regression w/ Elastic Net	Decision Tree	Random Forest	XGBoost	MARS	Support Vector Machine
Number of Variables	18	71	71	71	71	71	71	71
Accuracy	0.718	0.711	0.711	0.707	0.726	0.719	0.711	0.719
Precision	0.758	0.772	0.772	0.716	0.747	0.751	0.765	0.758
Recall/Sensitivity	0.733	0.689	0.690	0.794	0.776	0.749	0.702	0.735
Specificity	0.699	0.739	0.738	0.596	0.662	0.681	0.722	0.698
F1 Score	0.745	0.729	0.729	0.753	0.761	0.750	0.732	0.746
AUC:	0.793	0.794	0.794	0.768	0.800	0.794	0.793	0.791

Of the methods tested, machine learning logistic regression and MARS had the highest measures of specificity, allowing them to correctly predict students who would not graduate

within six years. Decision tree and random forest had much lower measures of specificity. Using machine learning logistic regression versus a decision tree in this case would result in an additional 468 students being correctly identified as in need of intervention. However, as was the case in the first outcome, the decision tree method had a much higher measure of sensitivity, which compared with machine learning logistic regression resulted in 441 fewer students incorrectly identified as needing intervention. Of the methods tested, the two with the highest F1 scores are decision tree and random forest. These two methods correctly predicted the most students who will graduate within six years while incorrectly predicting the fewest students who are not in need of intervention. However, both methods overlook the most students who are in need of intervention.

Discussion

In the field of institutional research, logistic regression is currently the most well-established method used for predicting student persistence. In recent years, studies have been conducted on the benefits of using machine learning in academia, and researchers have demonstrated that machine learning techniques may offer improvements over logistic regression for issues of prediction. The purpose of this study was to explore the utility and efficacy of several machine learning methods for predicting student persistence and to provide a practical comparison between these methods and traditional logistic regression.

The present paper has made several references to providing students who are in need of assistance with "interventions," and as such it is necessary to expand on what those interventions may entail. The basic concept of intervention is as simple as providing some amount of additional resources to a student to support their effort to persist in college. These interventions can take many forms and can vary greatly in terms of cost and time commitment. Low-cost interventions can include publicizing information about resources available to all students, holding information sessions about how to improve study habits or time management, or having academic advisors send emails to their students to check in on progress. More costly interventions may include having the student receive additional tutoring sessions or enrolling them in a learning community (providing them with a cohort of students who all follow the same curriculum to help enable engagement in a high-touch environment). Higher-cost interventions can include providing financial assistance (such as subsidizing the cost of textbooks or transportation to ease financial burdens), developing individualized support programs, or having academic advisors scheduling regular touchpoints with at-risk students.

In any approach to providing at-risk students with necessary additional support, there is an inherent tradeoff between available resources and the ability to identify and reach the students most in need. If universities had unlimited funds, time, information, and human resources, they could potentially provide all of their students with individualized support to enable their success. Needless to say, this is unrealistic. Choosing the appropriate method for predicting student outcomes is not a one-size-fits-all application and will depend heavily on the particular university and its goals.

The present study utilized optimized threshold selection for each model, which adjusted each model's ability to correctly classify cases based on the model's ROC curve. The optimized threshold was chosen to maximize the value of both sensitivity and specificity, providing a more balanced prediction particularly for the imbalanced class outcome of one-year retention. Choosing an appropriate cutoff point for this application is crucial because it can determine the difference between overlooking students who truly need intervention and wasting valuable resources by providing intervention to students who are not at risk of dropping out. To illustrate the importance of selecting an appropriate threshold value, consider the following three examples for predicting one-year retention. In the first example, the threshold for the algorithm logistic regression model was kept at a standard 50% (Table 3.1). In this case, the model's specificity is only 0.227, meaning 23% of the students were correctly identified as in need of intervention, and 77% of at-risk students were overlooked. However, this threshold does manage to identify many at-risk students without overloading resources. An institutional researcher may still hesitate to recommend this threshold to avoid the majority of at-risk students being left without intervention. In the second example, the threshold was optimized at 86% (Table 3.2). In this case, the model's specificity has increased to 0.649, meaning 65% of students who were at

risk of dropping out were correctly identified. When compared to the previous example, this model correctly identified an additional 518 students who were in need of intervention. However, it also highly overestimated the number of students who were at risk of dropping out, incorrectly identifying an additional 1,637 students as at-risk. An institutional researcher may choose not to recommend this threshold to avoid placing undue strain on the university's resources. In the third example, the threshold is lowered to 25% (Table 3.3). In this case, the specificity has dropped to 0.076, and far fewer students are predicted to be at risk of dropping out. However, an institutional researcher may question the utility of this threshold if it can only identify a small number of students in need, preventing the model from performing its ostensible purpose.

Table 3.1 Algorithm Logistic Regression: 50% Threshold		Actual	
		Not Retained	Retained
Predicted	Not Retained	279	117
	Retained	949	7268

Table 3.2 Algorithm Logistic Regression: 86% Threshold (Optimized)		Actual	
		Not Retained	Retained
Predicted	Not Retained	797	1754
	Retained	431	5631

Table 3.3 Algorithm Logistic Regression: 25% Threshold		Actual	
		Not Retained	Retained
Predicted	Not Retained	93	26
	Retained	1135	7359

Lest an institutional researcher end up feeling like Goldilocks, all of the examples above may be useful in approaching the issue of predicting student retention without overtaxing university resources by being applied together. Using this approach, one can identify tiers of students identified with different thresholds and apply interventions to each group based on available resources. For example, an institutional researcher may recommend that all 2,551 students predicted to be at risk using the optimized threshold above receive a low-cost intervention of receiving an email with information about the university's tutoring center and study tips. Subsequently, they may recommend that all 396 students predicted to be at risk using the 50% threshold be enrolled in a learning community or schedule extra time at the tutoring center. Finally, they can recommend that the 119 students predicted to be at risk using the 25% threshold meet with an academic advisor to receive individualized support. This tiered approach can help save resources by targeting the most at-risk students with more costly interventions while ensuring the fewest at-risk students be left without any additional support.

Another issue to consider when using these methods for real-world application is that of interpretability. While this study intended to identify the best methods for predicting student outcomes, in practice it is often necessary to understand the inner workings of a model, particularly for influencing policy changes. Many higher education stakeholders have come to expect auditable, interpretable results like those that come out of logistic regression and would have difficulty trusting a black box model as the basis for implementing important changes. And such hesitance would be reasonable—there is much evidence to suggest that machine learning algorithms are prone to demographic biases. If the main goal for predicting student outcomes is simply to flag students who may be in need of support, this is less of a concern. But it is

important to consider the overarching purpose and desired outcomes when choosing a method for this particular application.

One interpretation of the results may suggest that traditional logistic regression be used in favor of the tested machine learning methods, and thus a null result for this particular study. However, it is worth considering how unique a case this study is. In the field of higher education, very few topics have been studied as thoroughly and for as long as the topic of predicting student persistence. In light of that, the present study provided an excellent chance to demonstrate how much value machine learning can provide over what is currently considered a gold standard methodology for predicting student outcomes. When stacked up against such a well-studied problem, machine learning does not wildly outperform theory-based logistic regression. However, it is important to note that all of the tested methods performed only middlingly. The highest AUC values were 0.766 for predicting one-year retention and 0.800 for predicting six-year graduation. If these levels of predictive power are the best that can be achieved via the tested methods, it is possible the limitation lies within the datasets themselves, and that the next obstacle to overcome is the amount and quality of data collected. It is unknown how much improvement is realistically available beyond the measures achieved in this study, but it is likely that machine learning will be integral to informing progress in the future. Considering how much learning activity has begun to take place online in the last several years, colleges will soon be sitting on vastly unexplored troves of data that would be impenetrable without the use of machine learning. With online learning, every movement and measure of engagement can be tracked. Detailed academic records, interactions with course materials, time spent on schoolwork, touchpoints with instructors, classmates, and advisors, and scores of other predictors

will be available to study. In order to analyze these datasets and predict future student outcomes, machine learning will likely become necessary standard practice.

The results of this analysis show that a theory-based logistic regression performs similarly to the machine learning methods tested for overall predictive value, though there is some important nuance within the types of predictions made and how each model achieved them. While each model may be used with some success, it is imperative to consider how each will be used in practice. Each method has varying levels of interpretability or auditability, which may be more or less important to practitioners depending on the particular application or audience. Additionally, each of the tested models was relatively simple to implement "out of the box" and performed similarly to logistic regression even without hyperparameter tuning, which was deemed unnecessary for this study. Although additional cross-validation methods and hyperparameter tuning exercises were discarded for this study due to high computational cost, with a smaller dataset these additions are still recommended and would likely be far less burdensome to execute. In any case, none of the tested methods delivered any obstacles that should hinder implementation in regular processes for any institutional research office that is already equipped to perform traditional logistic regression or other statistical analyses.

Limitations

The data analyzed for this study are unique to the university system that provided them, and as such the generalizability of the results may be limited. In addition, the variables provided were somewhat limited to what may be considered a "traditional" institutional research dataset. This sort of dataset contains much of the information needed to report to external parties, and to conduct a traditional, theory-based logistic regression to predict or explain student outcomes. However, some potentially informative data were missing from the dataset, namely FAFSA

(Free Application for Federal Student Aid) data (which would include information about the student's household income and parental education), courses taken, and student engagement measures (participation in clubs, involvement in campus activities, use of tutoring, advising, or other academic resources, etc.). Beyond this, there is much more data that could be collected on students, particularly with the proliferation of distance learning in recent years. It is probable that the real value of machine learning for this application may lie in identifying significant prediction patterns within these yet to be examined areas, such as engagement with course content, detailed attendance and assignment tracking, utilization of academic resources, timing of registration, and so on. Therefore, it is recommended to conduct a similar study once this sort of data have been consistently collected for several years.

Appendix A: Variables Used

Outcome 1 - One-year retention

The predictors included for the theory-based logistic regression to predict one-year retention were:

Academic preparation:

- College Preparatory Initiative (CPI) units - the number of units taken in courses designated as college-preparatory by the university
- College Admissions Average (CAA) - weighted high school GPA calculated by the university
- Total SAT score

Background and demographics:

- City resident
- Disabled
- Veteran
- Economically disadvantaged (computed by the university system to indicate whether a student participated in any of several economic assistance programs)
- Educationally disadvantaged (computed by the university system to indicate whether a student has ever failed a skills test)
- Foreign born
- Female
- Asian or Pacific Islander
- Black
- Hispanic

College performance:

- Pell or Tuition Assistance Program (TAP) financial award flag
- First semester GPA
- First semester credits achieved
- First semester credits withdrawn
- Needed any remediation
- Passed any remediation

The predictors included for the machine learning methods to predict one-year retention were:

Academic preparation:

- College Preparatory Initiative (CPI) units (the number of units taken in courses designated as college-preparatory by the university) for English, foreign language, math, science, social studies, and arts
- College Admissions Average (CAA) (weighted high school GPA calculated by the university) for English, foreign language, math, science, and social studies
- Math and verbal SAT scores
- Type of high school: within the city and public, within the city and private, within the state

Background and demographics:

- Residency: City resident, state resident, foreign resident
- Foreign born
- Disabled

- Veteran
- Single parent flag
- Economically disadvantaged (computed by the university system to indicate whether a student participated in any of several economic assistance programs)
- Educationally disadvantaged (computed by the university system to indicate whether a student has ever failed a skills test)
- English as a second language (ESL) flag
- Limited English flag (indicates whether the student self-reported as speaking a language other than English at home, and/or being equally or more comfortable with a language other than English)
- Resident tuition flag (whether the student is paying discounted tuition rate for being a resident of the city)
- Pell or Tuition Assistance Program (TAP) financial award flag
- Financial aid awarded during year one
- No delayed entry (whether a student enrolled within 15 months of high school graduation)
- College was in student's top three choices on the university system's application
- Fall entry
- Female
- Dependent
- Age at entry
- Asian or Pacific Islander
- Black

- Hispanic
- Race/gender interactions

College performance:

- Needed any remediation
- Passed any remediation
- Full-time in semester one
- First semester GPA
- First semester credits attempted
- First semester credits achieved
- First semester credits withdrawn
- Declared major in first semester: STEM or non-STEM/undeclared
- Enrollment in the university system's higher education opportunity program in year one

Outcome 2 - Six-year graduation

The predictors included for the theory-based logistic regression to predict six-year graduation were:

Academic preparation:

- College Preparatory Initiative (CPI) units - the number of units taken in courses designated as college-preparatory by the university
- College Admissions Average (CAA) - weighted high school GPA calculated by the university
- Total SAT score

Background and demographics:

- Entry age
- City resident
- Female
- Asian or Pacific Islander
- Black
- Hispanic

College performance:

- Fourth semester GPA
- First year credits attempted
- Second year credits attempted
- First year credits withdrawn
- Second year credits withdrawn
- Declared major in second year: STEM or non-STEM/undeclared
- Enrollment in the university system's higher education opportunity program in year two

The predictors included for the machine learning methods to predict six-year graduation were:

Academic preparation:

- College Preparatory Initiative (CPI) units (the number of units taken in courses designated as college-preparatory by the university) for English, foreign language, math, science, social studies, and arts

- College Admissions Average (CAA) (weighted high school GPA calculated by the university) for English, foreign language, math, science, and social studies
- Math and verbal SAT scores
- Type of high school: within the city and public, within the city and private, within the state

Background and demographics:

- Residency: City resident, state resident, foreign resident
- Foreign born
- Disabled
- Veteran
- Single parent flag
- Economically disadvantaged (computed by the university system to indicate whether a student participated in any of several economic assistance programs)
- Educationally disadvantaged (computed by the university system to indicate whether a student has ever failed a skills test)
- English as a second language (ESL) flag
- Limited English flag (indicates whether the student self-reported as speaking a language other than English at home, and/or being equally or more comfortable with a language other than English)
- Resident tuition flag (whether the student is paying discounted tuition rate for being a resident of the city)
- Pell or Tuition Assistance Program (TAP) financial award flag
- Financial aid awarded during year one

- No delayed entry (whether a student enrolled within 15 months of high school graduation)
- College was in student's top three choices on the university system's application
- Fall entry
- Dependent
- Age at entry
- Female
- Asian or Pacific Islander
- Black
- Hispanic
- Race/gender interactions

College performance:

- Needed any remediation
- Passed any remediation
- Full-time in year one and two
- First semester GPA
- First and second year credits attempted
- First and second year credits withdrawn
- Declared major in first semester: STEM or non-STEM/undeclared
- Enrollment in the university system's higher education opportunity program in year one

Appendix B: Model Results

Outcome 1 - One-year retention

Confusion Matrices

Manual LR		Actual		Number of Variables: 20 Accuracy: 0.719 Precision: 0.925 Recall/Sensitivity: 0.732 Specificity: 0.641 F1 Score: 0.817 AUC: 0.746
		Not Retained	Retained	
Predicted	Not Retained	787	1979	
	Retained	441	5406	

Algorithm LR		Actual		Number of Variables: 59 Accuracy: 0.746 Precision: 0.929 Recall/Sensitivity: 0.762 Specificity: 0.649 F1 Score: 0.838 AUC: 0.766
		Not Retained	Retained	
Predicted	Not Retained	797	1754	
	Retained	431	5631	

Decision Tree		Actual		Number of Variables: 59 Accuracy: 0.840 Precision: 0.914 Recall/Sensitivity: 0.898 Specificity: 0.492 F1 Score: 0.906 AUC: 0.749
		Not Retained	Retained	
Predicted	Not Retained	604	754	
	Retained	624	6631	

Random Forest		Actual		Number of Variables: 59 Accuracy: 0.793 Precision: 0.917 Recall/Sensitivity: 0.833 Specificity: 0.549 F1 Score: 0.873 AUC: 0.749
		Not Retained	Retained	
Predicted	Not Retained	674	1230	
	Retained	554	6155	

XGBoost		Actual		Number of Variables: 59 Accuracy: 0.744 Precision: 0.916 Recall/Sensitivity: 0.772 Specificity: 0.572 F1 Score: 0.838 AUC: 0.730
		Not Retained	Retained	
Predicted	Not Retained	702	1681	
	Retained	526	5704	

MARS		Actual		Number of Variables: 59 Accuracy: 0.749 Precision: 0.927 Recall/Sensitivity: 0.767 Specificity: 0.639 F1 Score: 0.840 AUC: 0.759
		Not Retained	Retained	
Predicted	Not Retained	785	1719	
	Retained	443	5666	

SVM		Actual		Number of Variables: 59 Accuracy: 0.823 Precision: 0.916 Recall/Sensitivity: 0.874 Specificity: 0.516 F1 Score: 0.895 AUC: 0.733
		Not Retained	Retained	
Predicted	Not Retained	634	928	
	Retained	594	6457	

Outcome 2 - Six-year graduation

Confusion Matrices

Manual LR		Actual		Number of Variables: 18 Accuracy: 0.718 Precision: 0.758 Recall/Sensitivity: 0.733 Specificity: 0.699 F1 Score: 0.745 AUC: 0.793
		Not Graduated	Graduated	
Predicted	Not Graduated	2290	1127	
	Graduated	988	3088	

Algorithm LR		Actual		Number of Variables: 71 Accuracy: 0.711 Precision: 0.772 Recall/Sensitivity: 0.689 Specificity: 0.739 F1 Score: 0.729 AUC: 0.794
		Not Graduated	Graduated	
Predicted	Not Graduated	2422	1311	
	Graduated	856	2904	

Decision Tree		Actual		Number of Variables: 71 Accuracy: 0.707 Precision: 0.716 Recall/Sensitivity: 0.794 Specificity: 0.596 F1 Score: 0.753 AUC: 0.768
		Not Graduated	Graduated	
Predicted	Not Graduated	1954	870	
	Graduated	1324	3345	

Random Forest		Actual		Number of Variables: 71 Accuracy: 0.726 Precision: 0.747 Recall/Sensitivity: 0.776 Specificity: 0.662 F1 Score: 0.761 AUC: 0.800
		Not Graduated	Graduated	
Predicted	Not Graduated	2169	945	
	Graduated	1109	3270	

XGBoost		Actual		Number of Variables: 71 Accuracy: 0.719 Precision: 0.751 Recall/Sensitivity: 0.749 Specificity: 0.681 F1 Score: 0.750 AUC: 0.794
		Not Graduated	Graduated	
Predicted	Not Graduated	2231	1056	
	Graduated	1047	3159	

MARS		Actual		Number of Variables: 71 Accuracy: 0.711 Precision: 0.765 Recall/Sensitivity: 0.702 Specificity: 0.722 F1 Score: 0.732 AUC: 0.793
		Not Graduated	Graduated	
Predicted	Not Graduated	2367	1254	
	Graduated	911	2961	

SVM		Actual		Number of Variables: 71 Accuracy: 0.719 Precision: 0.758 Recall/Sensitivity: 0.735 Specificity: 0.698 F1 Score: 0.746 AUC: 0.791
		Not Graduated	Graduated	
Predicted	Not Graduated	2289	1119	
	Graduated	989	3096	

References

Aggarwal, C. (2016). An Introduction to Data Classification. In Aggarwal, C., Data Classification: Algorithms and Applications (1st ed., ch. 1). CRC Press, Taylor & Francis Group.

Astin, A. (1985). Achieving Educational Excellence. San Francisco, CA: Jossey-Bass Publishers.

Astin, A., & Oseguera, L. (2012). Pre-College and Institutional Influences on Degree Attainment. In Seidman, A., College student retention formula for student success (2nd ed., ch. 6). Rowman & Littlefield Publishers.

Boehmke, B. & Greenwell, B. (2020). Hands-on machine learning with R. CRC Press.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. Statistical Science, 16(3), 199-215.

Burke, A. (2019). Student Retention Models in Higher Education: A Literature Review. College and University, 94(2), 12–21.

Chipman, H., George, E., & McCulloch, R. (2010). BART: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1), 266–298. https://doi.org/10.1214/09-aoas285

Cole, K., Kennedy, M., Ben-Avie, M. (2009). The role of precollege data in assessing and understanding student engagement in college. New Directions for Institutional Research, 2009(141), 55–69. https://doi.org/10.1002/ir.286

Cole, J., & Korkmaz, A. (2010). Using longitudinal data to improve the experiences and engagement of first-year students. New Directions for Institutional Research, 2010, 43–51. https://doi-org.ezproxy.gc.cuny.edu/10.1002/ir.371

Efron, B., & Hastie, T. (2016). Computer Age Statistical Inference. In Computer Age Statistical Inference. Cambridge University Press.

Friedman (1991). Multivariate Adaptive Regression Splines. The Annals of Statistics, 19(1), 1–67. https://doi.org/10.1214/aos/1176347963

González, J. M. B., & DesJardins, S. L. (2002). Artificial Neural Networks: A New Approach to Predicting Application Behavior. Research in Higher Education, 43(2), 235–258. https://doi-org.ezproxy.gc.cuny.edu/10.1023/a:1014423925000

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). The Elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer.

Lingjun, H., Levine, R. A., Fan, J., Beemer, J., & Stronach, J. (2018). Random Forest as a Predictive Analytics Alternative to Regression in Institutional Research. Practical Assessment, Research & Evaluation, 23(1), 1–16.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. New Directions for Institutional Research, 2006(131), 17–33. https://doi-org.ezproxy.gc.cuny.edu/10.1002/ir.185

Irwin, V., De La Rosa, J., NCES; Wang, K., Hein, S., Zhang, J., Burr, R., Roberts, A., AIR; Barmer, A., Bullock Mann, F., Dilig, R., and Parker, S., RTI (2022). The Condition of Education 2022. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R (1st ed. 2013.). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling (1st ed. 2013.). Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Liu, Huang, J., & Xie, Y. (2021). Educational Visualization Application Based on Machine Learning Algorithm to Predict Student Learning. 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), 2, 236–241. https://doi.org/10.1109/ICIBA52610.2021.9688161

Luan, J. (2002). Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 2002(113), 17–36. https://doi.org/10.1002/ir.35

Lundberg, S. & Lee, S. (2017). A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017).

Molnar, C. (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book.

Ribeiro, M. T., Singh, S., Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. http://dx.doi.org/10.1145/2939672.2939778

Shmueli, G. (2010). To Explain or to Predict? Statistical Science, 25(3), 289-310. doi:10.1214/10-STS330.

Smirani, Yamani, H. A., Menzli, L. J., & Boulahia, J. A. (2022). Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths. Scientific Programming, 2022, 1–15. https://doi.org/10.1155/2022/3805235

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Thomas, E. H., & Galambos, N. (2004). What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analysis. Research in Higher Education, 45(3), 251–269. https://doi-org.ezproxy.gc.cuny.edu/10.1023/B:RIHE.0000019589.79439.6e

Tinto, V. (1993). Leaving College: Rethinking the Causes and Cures of Student Attrition (2nd ed.). Chicago: University of Chicago Press.

Wang, P.W., Lin, C.J. (2016). Support Vector Machines. In Aggarwal, C., Data Classification: Algorithms and Applications (1st ed., ch. 7). CRC Press, Taylor & Francis Group.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons from Machine Learning. Perspectives on Psychological Science, 12(6), 1100–1122. https://doi.org/10.1177/1745691617693393