Article

Journal of College Student Retention:
Research, Theory & Practice
2023, Vol. 24(4) 1054–1077

DOI: 10.1177/1521025120963821
journals.sagepub.com/home/csr
SAGE

Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach

Huade Huo[1], Jiashan Cui[1], Sarah Hein[1], Zoe Padgett[1], Mark Ossolinski[1], Ruth Raim[1] and Jijun Zhang[1]

Abstract

Student attrition represents one of the greatest challenges facing U.S. postsecondary institutions. Approximately 40 percent of students seeking a bachelor's degree do not graduate within 6 years; among nontraditional students, who make up half of the undergraduate population, dropout rates are even higher. In this study, we developed a machine learning classifier using the XGBoost model and data from the National Center for Education Statistics (NCES) Beginning Postsecondary Students (BPS) Longitudinal Study: 2012/14 to predict nontraditional student dropout. In comparison with baseline models, the XGBoost model and logistic regression model with features identified by the XGBoost model displayed superior performance in predicting dropout. The predictive ability of the model and the features it identified as being most important in predicting nontraditional student dropout can inform discussion among educators seeking ways to identify and support at-risk students early in their postsecondary careers.

Keywords

nontraditional students, dropout rates, machine learning, decision trees

[1]American Institutes for Research, Arlington, Virginia, United States

Corresponding Author:
Jijun Zhang, American Institutes for Research, 1400 Crystal Drive, 10th Floor, Arlington, VA 22202, United States.
Email: jizhang@air.org

While access to postsecondary education has increased substantially over the decades, student dropout remains a critical issue facing institutions today: approximately 40 percent of students seeking bachelor's degrees in the United States do not complete their degree within 6 years (Snyder et al., 2018). Dropout rates are significantly higher for the approximately 50 percent of undergraduate students who are classified as "nontraditional" (those who meet at least one of seven characteristics: delayed enrollment into postsecondary education, attended college part time, worked full time, was financially independent for financial aid purposes, had dependents other than a spouse, was a single parent, or did not have a high school diploma) (Choy, 2002; Radford et al., 2015). These nontraditional students may choose to start their postsecondary education at all sectors of education, including 4-year, 2-year, and less-than-2-year institutions.

Retention in postsecondary education is important because obtaining an undergraduate postsecondary degree can have a major impact on both individual and societal outcomes. These individual and societal benefits are particularly important for nontraditional students, who are more likely than traditional students to face socioeconomic barriers (Trenz et al., 2015). Because nontraditional undergraduates are retained at lower rates than traditional undergraduates, and attaining a degree is crucial for these students, it is important for researchers and universities to understand the factors that influence nontraditional student retention. Understanding how these factors differ from the factors that influence traditional student retention—and being able to predict which students are at a higher risk of dropping out—can allow postsecondary institutions to target programs to support these students, increasing the likelihood that they will complete their degree. Identifying and developing strategies to improve retention rates for these students is thus a top priority for postsecondary administrators, educators, and policymakers. One such promising approach lies in the use of machine learning (also known as data mining) to identify students who are likely to drop out.

Supervised machine learning models have been applied in a variety of prediction scenarios—including the prediction of postsecondary dropout and retention—with increasing frequency in recent years. In a supervised machine learning model, users "train" an algorithm by inputting independent variables—such as students' demographic characteristics and standardized testing scores—paired with an already-known target outcome, such as "dropout" or "persist." The model's parameters are adjusted according to the training data until it is capable of predicting as-yet unknown outcomes with maximum accuracy.

Applying machine learning models to the prediction of student dropout can produce a multitude of benefits to students and institutions alike, as educators can use these models to identify who among their student bodies is at risk of dropping out and develop strategies for providing those students with the support that may help them achieve a successful postsecondary career. In this study,

we identified important factors that predicted dropout using the recently developed XGBoost method and used both XGBoost and other models to predict dropout among nontraditional students. We evaluated the prediction accuracy across different models.

Literature Review

Undergraduate dropout and retention have long represented a major area of focus in postsecondary education research (Aulck et al., 2016; Tinto, 2006). At 4-year colleges alone, approximately 30 percent of first-year students do not return for a second year, a figure that represents billions of dollars of lost revenue for institutions and a sizable portion of the population that is less prepared to enter the U.S. workforce (Aulck et al., 2016).

Estimates of the number of nontraditional students vary, but researchers have agreed that nontraditional students are a substantive portion of all students in the United States. When defining "nontraditional" narrowly by age, nontraditional students make up 30 percent to 50 percent of the undergraduate student population in the United States (Forbus et al., 2011). However, the U.S. Department of Education found, using the same broader definition used in this paper, that in 1995–96, 1999–2000, 2003–04, 2007–08, and 2011–12, at least 70 percent of undergraduates had at least one of the seven defining characteristics of nontraditional students, and at least 50 percent had two or more characteristics (Radford et al., 2015).

Beyond their defining characteristics, nontraditional undergraduates tend to have other traits in common. Nontraditional students generally have longer commutes and feel less like a part of the college environment (Forbus et al., 2011). They also more often report being female or working class, and score higher on life stress, anxiety, and depression scales (Trenz et al., 2015). Overall, nontraditional undergraduates also tend to be less academically prepared for postsecondary education; lower percentages of nontraditional students than traditional students had a high school GPA of 3.50 or higher or took calculus as the highest level of mathematics in high school (Radford et al., 2015).

Although nontraditional students consist of a considerable portion of all undergraduate students, they have consistently been retained at lower rates than traditional students (Choy, 2002; Metzner & Bean, 1987; Skomsvold et al., 2011). Skomsvold et al. (2011, Table 3.0-D), who use the same definition of "nontraditional" used in this paper, have shown that, after 6 years, a lower percentage of first-time undergraduate students with zero of the seven defining nontraditional characteristics had exited postsecondary education without attaining a degree (14.6 percent) than their peers with one (28.6 percent), two or three (40.3 percent), or four or more nontraditional characteristics (48.8 percent). This pattern also held true when looking at each of the seven nontraditional characteristics individually. For example, 42.2 percent of

students who worked full time in their first year had exited postsecondary education without attaining a degree, compared with 25.6 percent of students who worked part time and 23.0 percent of students who did not work.

Previous studies on retention among nontraditional undergraduates suggest academic, financial, and social factors as possible explanations for the high levels of dropout seen among nontraditional students. In looking at academic factors, postsecondary grade point average and cognitive ability have been associated with dropout. Research has found that nontraditional undergraduates with a lower GPA are more likely to drop out of postsecondary education (Fortin et al., 2016; Metzner & Bean, 1987). Similarly, a study of nontraditional students found that those with higher measured cognitive ability were more likely to complete their degree (Taniguchi & Kaufman, 2005). These findings suggest the importance of academic performance in predicting retention.

Nontraditional undergraduates have been shown to experience higher financial stress than their traditional peers, relying on their personal income more frequently than traditional students (Forbus et al., 2011). Fortin et al. (2016) found that a lack of financial support from relatives was associated with the dropout of nontraditional students. As employment is one of the defining characteristics of nontraditional students, nontraditional students tend to work more hours than their traditional peers, and students who are employed while enrolled in college have been more likely to drop out compared with their peers (Forbus et al., 2011; Gilardi & Guglielmetti, 2011). In addition, factors such as being enrolled part time and having children have also been associated with lower retention rates for nontraditional undergraduates (Fortin et al., 2016; Taniguchi & Kaufman, 2005).

A number of studies have pointed to variables such as nontraditional undergraduates' typically busier lifestyles, responsibilities outside of schoolwork, and lower levels of involvement in campus social activities as potential sources of stress influencing their higher dropout rates (Forbus, et al., 2011; Hoyt et al., 2010; Willans & Seary, 2011). Given the additional responsibilities nontraditional undergraduates typically have, factors associated with campus involvement have also been affiliated with dropout. Gilardi and Guglielmetti (2011) found that higher use of support services offered by universities was correlated with persistence; however, nontraditional students were less likely to use these services. Additionally, Forbus et al. (2011) reported that nontraditional students were less likely to feel like they socially belong in a college environment. This feeling of belonging is important, as higher levels of perceived social integration on campus have been associated with persistence (Gilardi & Guglielmetti, 2011).

Researchers agree that further research is needed to gain a greater understanding of the challenges these students face on their path to completing an undergraduate degree and how to respond to them (Hoyt et al., 2010; Pontes & Pontes, 2012). We seek to explore further the factors that impact nontraditional student retention in this paper.

Ma et al. (2016) at The College Board report many benefits of higher education. These benefits include individual factors, such as better employment outcomes and higher earnings (both soon after completing a degree and in lifetime earnings), better access to health insurance and healthier lifestyles (such as lower smoking and obesity rates), and the ability for parents to spend more time with children. Benefits also include societal factors such as increased participation in volunteer organizations, higher voting rates among the more educated population, increased social mobility, and less reliance on social support such as food assistance programs.

There are many personal reasons for individuals to obtain postsecondary degrees. One such reason is the financial impact of higher education. According to McFarland et al. (2017), the unemployment rate in 2017 was lower for those with a bachelor's degree and higher for those with some college or no college education. Many researchers have studied the effect of education on earnings (Barrow & Rouse, 2005; Card, 1999; Hauser & Daymont, 1977; Murphy & Welch, 1992; Oreopoulos & Petronijevic, 2013; Perna, 2003; Tamborini et al., 2015) and employment outcomes (Mincer, 1991; Riddell & Song, 2011; Wolbers, 2000) and found that higher levels of education generally lead to higher earnings and fewer instances of unemployment.

Additional personal outcomes of higher education include parents having more time to spend with their children (Guryan et al., 2008), greater general life satisfaction (Salinas-Jiménez et al., 2013), and positive health outcomes (Grossman, 1976; Hasnain et al., 2007; Lleras-Muney, 2005; Painter et al., 2012). For example, Ross and Wu (1995) found that education improved health outcomes, both directly and through mechanisms of economic conditions (such as higher incomes and more fulfilling work), social conditions (such as higher levels of social support and psychological benefits of having more control over their lives), and lifestyle choices (such as not smoking, less drinking, and more exercise).

In addition to the individual benefits of completing a postsecondary degree, there are societal and economic benefits. Growth in the rates of higher education was correlated with lower unemployment rates, higher productivity, and greater economic growth and business creation (Howe, 1993; Jones & Vedlitz, 1993; Koropeckyj et al., 2017). A higher population of college graduates living in an area also increased the wages of non-graduates in the same area (Moretti, 2004). In contrast, college dropouts are associated with economic losses by way of lost income and lower tax receipts (Schneider & Yin, 2011). Additional societal factors influenced by greater rates of higher education have included decreased crime (Ehrlich, 1975) and increased political engagement (Hillygus, 2005).

Machine Learning Approaches

In recent years, an increasing number of studies have investigated machine learning as a potentially powerful tool for assessing students' likelihood of

dropping out of college before attaining a degree (Livieris et al., 2016). Being able to predict which of a school's students are most likely to drop out can allow institutions to allocate resources and target intervention programs in an informed manner, potentially producing a range of benefits for students and schools alike. Successfully targeted intervention programs for at-risk students are capable of improving schools' graduation rates and the precision of their tuition revenue forecasts. In turn, this can result in a higher number of graduates who are better prepared to enter the workforce as well as fewer resources spent on students who ultimately do not graduate (Herzog, 2006).

Data and Methods

This study uses data from the 2012–14 Beginning Postsecondary Students Longitudinal Study (BPS:12/14), which included about 24,770 participating students from 1,480 institutions (Hill et al., 2016). This dataset collected data from a cohort of students at 2 points in time: near the ends of their first year (2011–12) and third year (2013–14) following initial entry into a postsecondary institution. BPS:12/14 contains data on students' demographic characteristics, school and work experiences, persistence, transfer, and degree attainment over 3 academic years. Our analysis is limited by the modest amount of data in BPS:12/14 on students' undergraduate college readiness and background information, such as student transcripts, direct measures of social interactions, and detailed measures of high school performance. The inclusion of this information in another model would account for more student-level characteristics and potentially increase overall predictive accuracy.

In our analytical sample, 52 percent of students are classified as nontraditional. Among the seven characteristics this research uses to classify nontraditional students, about 29 percent of the student sample were financially independent, 17 percent had one or more dependents, and 11 percent were single caregivers (Table 1). Some 11 percent of these students did not receive a traditional high school diploma, and 37 percent delayed postsecondary enrollment for 1 year or more. About 12 percent were enrolled exclusively part time,[1] and 13 percent worked full time while enrolled in school.

XGBoost

XGBoost, the method examined in this paper, is a "boosted trees" model, or an "ensemble" version of the decision tree method (Chen & Guestrin, 2016). In a typical decision tree model, such as a C4.5 decision tree, a tree structure consisting of various nodes performs split operations on each node based on the information gain values for each feature (i.e., variables) of the dataset being applied. At each level, the attribute, or classification rules, with the highest information gain is chosen as the basis for the split criterion (Aguiar, 2015).

Table 1. Percentage Distribution of Undergraduates, by Selected Student Characteristics: 2011–12.

Selected student characteristics	Percent
Total	**100.0**
Being independent for financial aid purposes	
Yes	28.6
No	71.4
Has dependent(s)	
Yes	17.0
No	83.0
Single with dependents[a]	
Yes	11.3
No	88.7
High school completion status	
High school diploma, foreign high school, or homeschooled	88.7
GED or other equivalency, completion certificate, or no high school diploma, certificate, or other equivalency	11.3
Delayed postsecondary enrollment[b]	
Less than one year	63.0
One year or more	37.0
Attendance status[c]	
Any full-time	88.3
Exclusively part-time	11.7
Worked while enrolled[d]	
Worked full time	12.7
Worked part time	23.5
Did not work	63.8

Note. Estimates include students enrolled in Title IV-eligible postsecondary institutions in the 50 states and the District of Columbia. Details might not sum to total because of rounding.
Source. Authors' calculations from the 2012/14 Beginning Postsecondary Students Longitudinal Study (BPS:12/14) Restricted-Use Data File.
[a]Includes students who were single, never married; separated; widowed; or divorced.
[b]We define "delayed postsecondary enrollment" as not entering postsecondary education within the calendar year of completing high school. Delayed entry estimates exclude students who did not earn a high school diploma, certificate, or equivalency because these students did not have a high school completion date.
[c]Full-time status for the purposes of financial aid eligibility was based on 12 credit hours, unless the awarding institution employed a different standard.
[d]Estimates exclude students who worked in school-related jobs (e.g., work-study or assistantships) and jobs held while not enrolled, including summer break. Full-time status was defined as working 35 or more hours per week, and part-time status was defined as working less than 35 hours per week.

However, in a boosted trees model such as gradient boosted machines (GBM), a sequence of tree structures is constructed in which each successive tree is built with greater weight attached to misclassified records from the previous tree; thus, the final prediction is based on the aggregate results of each tree and minimizes misclassification (Delen, 2010; Herzog, 2006). Ensemble methods like boosting "have been called the most influential development in Data Mining and Machine Learning in the past decade" (Seni & Elder, 2010) and in recent years they have been proven to yield greater accuracy than older methods, such as those used by Aguiar (2015) and by Nakhkob and Khademi (2016). Aguiar (2015) found that boosting yielded higher predictive accuracy than decision tree, support vector machine, and logistic regression models in predicting delayed graduation from high school.

While many studies, especially in recent years, have explored the topic of machine learning and its application to the prediction of student dropout, this study, which applies XGBoost to dropout prediction and uses logistic regression as a baseline model for comparison, will contribute to this effort in a few key ways: First, through the exploration of XGBoost, a new boosting method that has shown great promise in other applications, and second, through the use of a dataset from the National Center for Education Statistics' Beginning Postsecondary Longitudinal Study (BPS). Numerous machine learning explorations of dropout prediction use data from a single school—or even a single course—and the conclusions drawn from those explorations are specific to those specialized groups of students. In contrast, BPS is a large and nationally representative sample of postsecondary students, and this study's results (and the factors it identifies as being most important in predicting nontraditional student dropout) may be more universally applicable to the broad subset of nontraditional students who tend to be at greater dropout risk than their traditional-student peers.

As shown in Figure 1, from BPS:12/14, we excluded the features specific to the 2014 collection except for "persistence" (i.e., retention), so features from 2011–12 and 2012–13 were used to predict retention as of the 2013–14 academic year. We included features from the 2014 collection that were updated to characteristics that would not be expected to change between years (race, sex, etc.). Persistence was converted into a binary feature; students were classified as "persisting" if they had attained a degree or certificate or were still enrolled anywhere during the 2013–14 academic year. 66 percent of students were classified as "persisting." All missing values were set to −9.

To create the logistic regression-based comparison models, we created an additional dataset with each categorical integer feature encoded as multiple binary features using a one-of-K scheme. Using the same random seeds for each dataset, we split both datasets into a training set (80 percent, or about 19,810 students) to build the model and a testing set (20 percent, or about 4,960 students) to evaluate and report results. In an attempt to balance the class

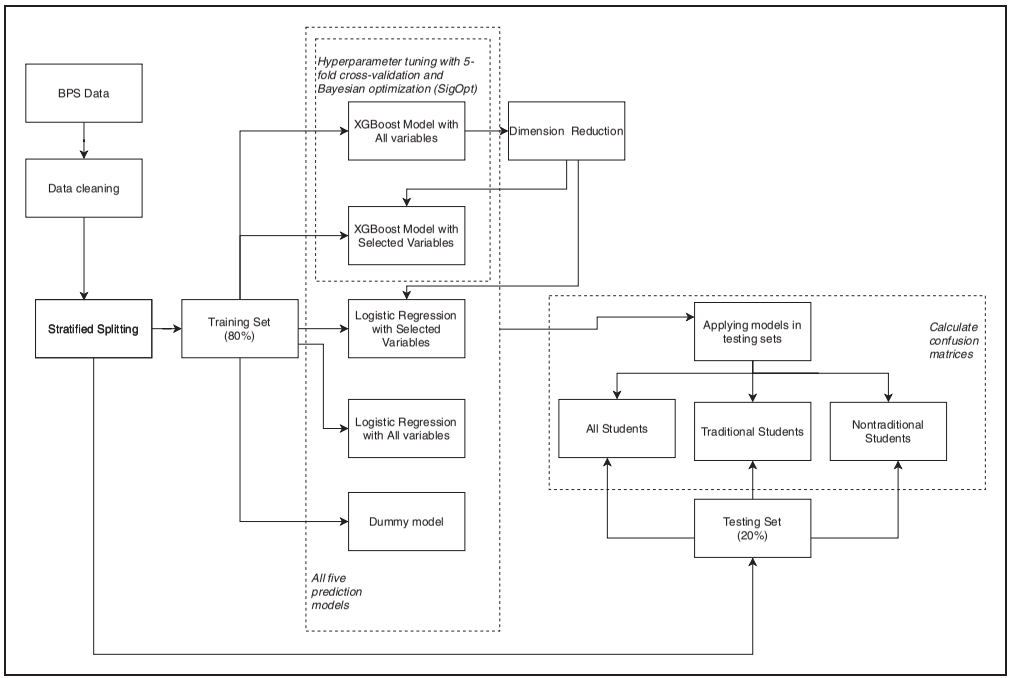

Figure 1. Study design.

distributions within the splits, the random sampling was done within the levels of the "persistence" feature. All students within the dataset, both nontraditional and traditional, were retained to develop the models.

Once we had training sets and testing sets, we fitted five models: XGBoost with all features, XGBoost with important features, logistic regression with all features, logistic regression with important features, and a dummy classifier model. Table 2 displays a comparison of these five models.

Model 1 is an XGBoost model with all 429 features. For both XGBoost models (Model 1 and Model 2), there is no conversion from categorical variable to multiple binary variables, as XGBoost is a tree-based algorithm, and tree-based algorithms can "easily handle qualitative predictors without the need to create dummy variables" (Hastie et al., 2017). By fitting Model 1, we obtained a set of parameters on each feature that we used to select important features (for Model 2 and Model 4) and a trained classifier that we could evaluate with testing data.

Model 2 is an XGBoost model with only important features. High-dimensional data could negatively affect the bias and variance of the model (a.k.a., the "curse of dimensionality") (Bellman, 1957; Hastie et al., 2017). After fitting Model 1, XGBoost returned a list of features along with their importance, measured by the improvement in accuracy a feature brings to the branches it is on (i.e., "gain"). To reduce the model's dimensions, we dropped all but the 50 most important features from Model 1, which contributed to nearly 90 percent of the gain, and fitted Model 2 with these top 50 features (see Table 5).

To optimize the performance of the XGBoost model, we used the Bayesian hyperparameter optimization provided by SigOpt. As an alternative to the grid search or random search approach, SigOpt increases model accuracy and accelerates model tuning (Dewancker et al., 2016). We fitted models with combinations of hyperparameters suggested by SigOpt, calculated the mean and standard deviation of binary error rates from five-fold cross-validation,

Table 2. Model Specifications.

	Model 1	Model 2	Model 3	Model 4	Model 5
Algorithm	XGBoost	XGBoost	Logistic Regression	Logistic Regression	Dummy
Encode categorical features as binary features	No	No	Yes	Yes	NA
Dimensionality reduction	No	Yes	No	Yes	NA
Number of features	429	50	1,624	219	1
Cross-validation	5-fold	5-fold	NA	NA	NA
Hyperparameter optimization	Bayesian	Bayesian	NA	NA	NA

submitted the training outcome to SigOpt for a new set of suggestions, and iterated this process 100 times to obtain the optimal set of hyperparameters. Table 3 shows the name, label, ranges of values explored, and optimal values of hyperparameters used in the prediction model.

To establish a benchmark for accuracy, three additional models were created: a logistic regression model with all features (Model 3), a logistic regression model with top XGBoost features (Model 4), and a dummy classifier model (Model 5). Logistic regression was fitted to the training data with all features (expanded from the original 429 features to 1,624 binary features) and then used to predict the possibility of dropout with testing data. Similarly, Model 4 fitted the training data with the top 50 features based on XGBoost's features of importance (the same 50 features used in Model 2, expanded to 219 binary features) and predicted the possibility of dropout with the testing data. A dummy classifier model makes predictions using simple rules. In this study, the dummy classifier model always predicted the most frequent label in the training set (i.e., "persistence").

Finally, we created confusion matrices using these five models and the testing set. The confusion matrix gives the number of true positives (the number of dropouts correctly predicted as dropouts), false positives (the number of non-dropouts incorrectly predicted as dropouts), true negatives (the number of non-dropouts correctly predicted as non-dropouts), and false negatives (the number of dropouts incorrectly predicted as non-dropouts) resulting from applying the model to the testing set. Quality metrics can then be calculated based on these values. Since nontraditional students have a higher risk of dropout than traditional students, we also tested our models, developed from the training dataset containing both traditional and nontraditional students, on the traditional and nontraditional student subsets of the testing set. We evaluated the quality of

Table 3. Hyperparameter Tuning Results by Model.

			Optimal value	
Hyperparameter	Label	Range	Model 1	Model 2
eta	Step size shrinkage used in update to prevent overfitting	[0,1]	0.25	0.25
gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree	$[0,\infty]$	6	9
max_depth	Maximum depth of a tree	$[0,\infty]$	3	3
nrounds	Number of boosting rounds in the current training	$[1,\infty]$	125	200
subsample	Subsample ratio of the training instance	(0,1]	1.0000	0.8291

each model's performance using a variety of performance measures, detailed below.

Since the test set is imbalanced, accuracy would be biased when making comparisons between different subsets. Therefore, we primarily use balanced accuracy and F_1 scores to evaluate our models. Additional measures included in our results, for finer detail, are accuracy, sensitivity/recall, specificity, precision/positive predictive value (PPV), and negative predictive value (NPV) (Exhibit 1).

Results

As dropout among nontraditional students presented as a challenge for educators, we focus on testing results that are from nontraditional students' subset in the test set, which were withheld from the training process. Among all models tested, the XGBoost models (Models 1 and 2) performed similarly well in terms of balanced accuracy (78.82 percent and 78.80 percent, respectively; see Table 4). These models also had comparably high F1 scores of 75.39 percent and 75.15 percent, respectively. In comparison, the XGBoost-informed logistic regression model with only the top variables had a balanced accuracy of 78.44 percent and an F1 score of 73.76 percent, higher than the logistic regression model without variable selection and the dummy model, suggesting variables selected by the XGBoost model can be generalized by using them in logistic regression models and improving its predictive power.

Sensitivity measures the percentage of actual dropouts who were correctly identified as being at risk of dropping out, and specificity measures the percentage of students who actually persisted 3 years after initial enrollment and had thus been correctly identified as not at risk of dropping out. Using data from the first 2 years following students' initial enrollment, the XGBoost model with

Exhibit 1. Equations defining evaluation measures used in this paper.

Evaluation Measure	Equation
Accuracy	$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$
Balanced Accuracy	$Balanced\ Accuracy = \left(\frac{1}{2}\right) * \left(\frac{TP}{P} + \frac{TN}{N}\right)$
Sensitivity/Recall	$Sensitivity\ or\ Recall = \frac{TP}{(TP+FN)}$
Specificity	$Specificity = \frac{TN}{(TN+FP)}$
Precision/PPV	$Precision\ or\ PPV = \frac{TP}{(TP+FP)}$
NPV	$NPV = \frac{TN}{(TN+FN)}$
F1 Score	$F1 = 2 * \left(\frac{Precision\ *\ Sensitivity}{Precision+Sensitivity}\right)$

Note. TP equals the number of true positive cases, FP equals the number of false positive cases, TN equals the number of true negative cases, and FN equals the number of false negative cases.
Source. Fawcett (2006) and Broderson et al. (2010).

Table 4. Evaluation Matrices for Nontraditional Students by Model.

	Model 1	Model 2	Model 3	Model 4	Model 5
Accuracy	0.7982	0.7978	0.5111	0.7844	0.5669
Balanced Accuracy	0.7882	0.7870	0.5041	0.7744	0.5000
Sensitivity/Recall	0.7140	0.7061	0.4519	0.6999	0.0000
Specificity	0.8624	0.8678	0.5563	0.8490	1.0000
Precision/PPV	0.7986	0.8032	0.4376	0.7797	N/A
NPV	0.7979	0.7944	0.5705	0.7874	0.5669
F1 Score	0.7539	0.7515	0.4446	0.7376	N/A

Source. Authors' calculations from the 2012/14 Beginning Postsecondary Students Longitudinal Study (BPS:12/14) Restricted-Use Data File.

selected variables (Model 2) had a sensitivity of 70.61 percent and a specificity of 86.78 percent. In other words, among nontraditional students, Model 2 can correctly detect 7 out of 10 dropouts while misclassifying less than 2 students as dropouts who did not have such risk. The XGBoost model with all variables (Model 1) and the XGBoost-informed logistic regression model with only the top variables (Model 4) also yield comparable satisfactory performance in comparison with the baseline models.

Sensitivity and specificity will vary in tradeoff according to the prediction threshold specified. Our analysis (Table 4) uses a 0.5 prediction threshold for classifying the predicted probability of persisting into binary "persist" and "dropout" categories, but Figure 2 below compares the sensitivity and specificity of Models 2 and 4 as that prediction threshold varies. The receiver operating characteristic (ROC) curve for Model 2 is largely to the left and above the ROC curve for Model 4, indicating that Model 2 generally outperforms Model 4 in terms of sensitivity and specificity as the prediction threshold varies, and both Model 2 and Model 4 outperform the baseline model (e.g., Model 3). A comparison of the area under the curve (AUC) for each ROC curve confirms this visual analysis; the AUCs for Model 2 and Model 4 (0.8713 and 0.8498, respectively) are higher than the AUC for Model 3 (0.4959).

To provide interpretability of the XGBoost model with all features (Model 1) and its results, the model returns data on the importance of each feature using the "gain" metric, which measures the relative contribution of the corresponding feature to the model. The gain is calculated by considering each feature's contribution to each tree in this decision tree model. Features with a higher gain metric are more important in the model for generating a prediction than features with a lower gain metric. As shown in Table 5, first-year (academic year 2011–12) attendance pattern, total amount borrowed for direct subsidized and unsubsidized loans in the second year (academic year 2012–13), and first-year GPA were among the most important features for predicting dropout.

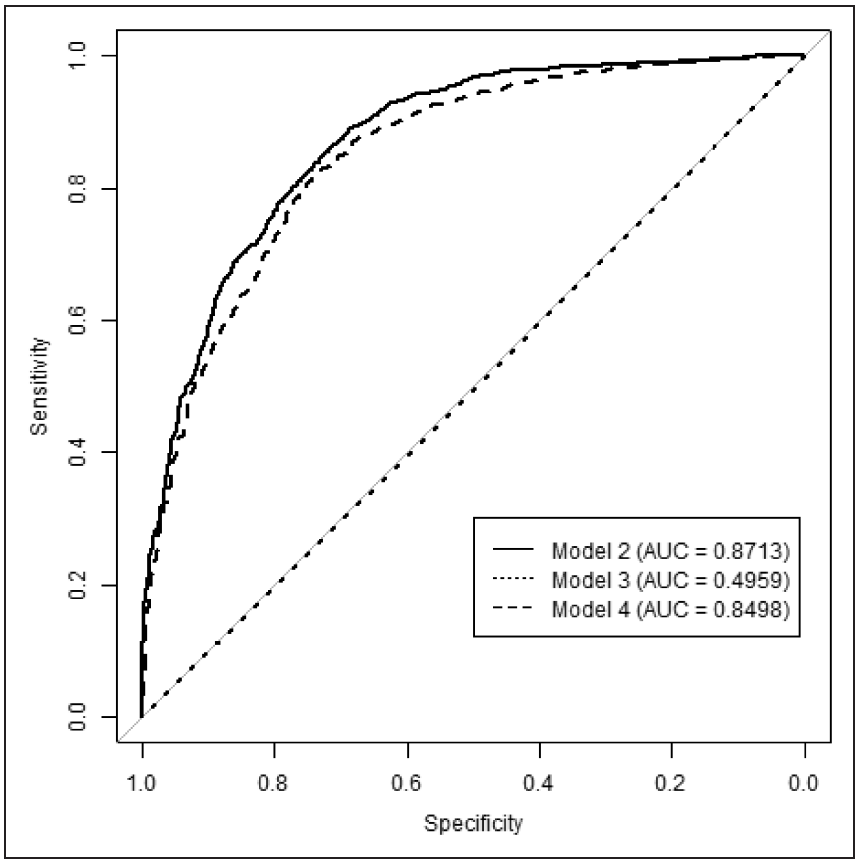

Figure 2. ROC curve and associated AUC for Models 2, 3, and 4.

Table 5. Features Used in XGBoost Model for Predicting Dropout 3 Years After First Enrollment.

	Name	Gain	Label	Variable type
1	attnstat	0.1185	Attendance pattern 2011–12	Categorical
2	stfy13	0.1064	Direct Subsidized and Unsubsidized Loans: Total borrowed 2012–13	Continuous
3	gpa	0.1001	Grade point average 2011–12	Continuous
4	tftype1	0.0888	Transfer status during 2011–12	Categorical
5	pell13	0.0709	Federal Pell Grant: Amount received 2012–13	Continuous

(continued)

Table 5. Continued.

	Name	Gain	Label	Variable type
6	degexp	0.0642	[Student's perceived] Likelihood of completing degree by expected date 2012	Continuous
7	hrswk13	0.0519	Jobs while enrolled: Hours worked 2012–13	Continuous
8	select3y	0.0410	Selectivity of last institution enrolled through June 2014	Categorical
9	instcat2	0.0348	Institutional category and control 2011–12 (degree-granting status, types of degrees or certificates awarded, and control of institution)	Categorical
10	degevr	0.0302	[Student's perceived] Likelihood of ever completing expected degree 2012	Continuous
11	efccps	0.0215	Expected Family Contribution (from Central Processing System) 2011–12	Continuous
12	ugdeg	0.0192	[Type of] Undergraduate degree program 2011–12	Categorical
13	tftype2	0.0176	Transfer status during 2012–13	Categorical
14	degexpdt	0.0174	Date (academic year) expected to complete degree requirements	Categorical
15	expba	0.0145	Bachelor's program intentions within 5 years 2012 (applies to students enrolled in an associate's degree program or non-degree undergraduate classes only)	Binomial
16	subloan	0.0132	Federal subsidized loans (Direct Subsidized & Perkins) 2011–12	Continuous
17	hrswk12	0.0131	Jobs while enrolled: Hours worked 2011–12	Continuous
18	numjbne13	0.0129	Jobs while not enrolled: Number of jobs 2012–13	Continuous
19	parhpamt	0.0104	Help from parents: Amount parents helped pay for expenses in 2011–12	Categorical
20	pellrat2	0.0098	Ratio of Pell grant to total grants 2011–12	Continuous
21	netcst2	0.0092	Student budget minus federal grants 2011–12	Continuous
22	tuition2	0.0089	Tuition and fees paid 2011–12	Continuous
23	riskindx	0.0088	Index of risk and nontraditional students 2012 (number of nontraditional student characteristics exhibited, scale of 0 to 7)	Categorical

(continued)

Table 5. Continued.

	Name	Gain	Label	Variable type
24	enlen	0.0079	First institution: Months enrolled total 2011–12	Continuous
25	tfedaid2	0.0072	Total federal aid (includes Veterans'/ DOD) 2011–12	Continuous
26	loanpct2	0.0065	Ratio of loans to total aid (including Direct PLUS Loans to parents) 2011–12	Continuous
27	cincome	0.0065	Total income (continuous) 2012	Continuous
28	parborn	0.0060	Parent born in US, PR, or US Territory	Categorical
29	pctenrnr	0.0056	Percent enrolled [at first institution]: Nonresident alien 2011–12	Continuous
30	curconf	0.0055	Academic confidence: 2011–12	Categorical
31	pctenrbk	0.0049	Percent enrolled [at first institution]: Black, non-Hispanic 2011–12	Continuous
32	t4lnamt2	0.0047	Title IV loans (includes Direct PLUS Loans to parents) 2011–12	Continuous
33	pctenrap	0.0047	Percent enrolled [at first institution]: Asian/Pacific Islander 2011–12	Continuous
34	efcaid	0.0045	Aid amount subject to federal EFC (expected financial contribution) limitation 2011–12	Continuous
35	budgetbk	0.0044	Budgeted cost of books and supplies	Continuous
36	cc2010s	0.0043	Carnegie Classification 2010: Size and setting 2011–12	Categorical
37	hcmathhi	0.0043	Highest level of high school mathematics	Categorical
38	trio	0.0042	TRIO program eligibility criteria 2011–12	Categorical
39	selectv2	0.0042	Selectivity of first institution (4-year institutions) 2011–12	Categorical
40	netcst10	0.0038	Tuition and fees minus federal grants 2011–12	Continuous
41	distance	0.0035	Distance from student's home (in miles) to first institution 2011–12	Continuous
42	inststat	0.0035	Location of first institution: State 2011–12	Categorical
43	plus13	0.0033	Direct PLUS Loans to parents: Amount borrowed 2012–13	Continuous
44	enrlsize	0.0032	First institution: Fall enrollment 2011–12	Continuous

(continued)

Table 5. Continued.

	Name	Gain	Label	Variable type
45	fedapp	0.0032	Applied for federal aid 2011–12	Binomial
46	stafct1	0.0031	Direct Subsidized Loan maximum 2011–12	Categorical
47	fedpct	0.0029	Ratio of federal aid to total aid 2011–12	Continuous
48	cc2010p	0.0025	Carnegie Classification 2010: Undergraduate instructional program 2011–12	Categorical
49	budnonaj	0.0024	Non-tuition expense budget (attendance adjusted) 2011–12	Continuous
50	Attendmr	0.0000	Purpose (first): Main reason for taking just classes 2011–12 (applies only to students who were not in a degree program)	Categorical

Source. Authors' calculations from the 2012/14 Beginning Postsecondary Students Longitudinal Study (BPS:12/14) Restricted-Use Data File.

Discussion and Conclusion

Our study adds to the growing body of literature that has found that machine learning models predict student dropout with higher accuracy than logistic regression modeling. Instead of limiting analysis to a relatively small dataset of students in a particular course or institution, this study expands upon the existing machine learning literature by using a large, national dataset. Given the prevalence of dropout at the postsecondary level, particularly among nontraditional undergraduate students, it has become increasingly important to develop a method to accurately identify individuals with a higher risk of dropping out. Our results suggest that a machine learning approach has the potential to help institutions address this growing issue. More accurate identification of students at risk of dropping out would allow universities to target their spending more effectively toward the retention of these students. Using a machine learning model, such as XGBoost, to predict dropout could result in cost savings for institutions, which may have limited resources available to dedicate to student retention.

In addition to predicting nontraditional student dropout, our model also identified the variables that produced the greatest gains in the model's performance. These variables were those that XGBoost found to predict the retention of nontraditional students best. The top variables in our model included (1) whether a student was enrolled full or part time, full year or part year, and the number of institutions attended; (2) total amount borrowed in loans; (3) grade point average in the first year; (4) transfer status; (5) amount received in federal Pell Grants; (6) student's perceived likelihood of completing degree on

time; (7) hours worked per week while enrolled; (8) selectivity of institution; (9) category and control of institution; and (10) student's perceived likelihood of ever completing expected degree. Those top variables, along with other variables identified by the model, can be used with other modeling methods to predict dropout.

Many of our top variables for predicting nontraditional student dropout reflect much of what has been found in the literature. Traditional methods, such as logistic regression, have identified variables including grade point average, employment status, and enrollment status (i.e., part- vs. full-time student) as factors that can be used to predict dropout (Fortin et al., 2016; Gilardi & Guglielmetti, 2011; Taniguchi & Kaufman, 2005). These three factors appeared in the top 10 variables generated by our model, suggesting their importance in predicting nontraditional student dropout. Additionally, previous research has identified the amount of financial support received from relatives as a factor associated with dropout (Fortin et al., 2016). Similar variables, including expected family financial contribution and financial help from parents, also appeared in the top 20 variables of our model.

While there was an overlap between the variables identified by our machine learning model and by the existing literature, there were also some differences in what was found to predict nontraditional student dropout. Traditional methods indicated that a nontraditional student's employment prior to entering undergraduate education, having or not having dependents, distance of commute, and perceived sense of belonging on campus were factors that could be used to predict attrition (Fortin et al., 2016; Gilardi & Guglielmetti, 2011; Taniguchi & Kaufman, 2005). Although our dataset contained these variables, they did not appear in the top 50 variables identified by our model as producing the greatest predictive gains. Our results suggest that other variables might instead have more predictive power. For example, our model identified several variables associated with financing postsecondary education, such as total amount borrowed in loans and amount received in federal Pell Grant, as strong predictors of dropout. Additionally, we found a student's perceived likelihood of both completing their degree on time and of ever completing their expected degree to be top predictors of dropout. Other key predictors identified by XGBoost and not by traditional methods were transfer status, the selectivity of an institution, and category and control of an institution. Our results suggest that retention policies should take these factors into account, in addition to the factors identified in studies that use traditional methods.

In the future, we plan to further increase predictive accuracy by exploring ensemble machine learning methods and involving additional data, including student records and transcripts data, which offers more details on undergraduate students' course-taking patterns. Additional research is needed to determine the optimal machine learning models and approaches for predicting dropout among nontraditional undergraduate students, especially at the national level. Further studies should look at ways to introduce survey weights into machine learning models and the consequences of not utilizing these weights, to obtain nationally representative results.

Authors' Note

Ruth Raim is also affiliated to District of Columbia Public Schools, Washington, District of Columbia, United States and Mark Ossolinski is also affiliated to University of Missouri, Columbia, Missouri, United States.

Acknowledgments

We would like to thank Ryan Baker, Markus Broer, Enis Dogan, and Xiaying Zheng for valuable analytical advice. We gratefully acknowledge the assistance of: Clark Elliott, Colleen Gaffney, Martin Hahn, David Katzman, Yuting Li, and Charles Yeomans. The AERA 2018 discussant Taylor Acee offered many constructive suggestions.

Data availability statement

The data that support the findings of this study are available from the National Center for Education Statistics, but restrictions apply to the availability of these data, which were used under license for the current study and so are not publicly available. Information about obtaining a restricted-use data license can be found at https://nces.ed.gov/pubsearch/licenses.asp.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Note

1. Estimates exclude students who worked in school-related jobs (e.g., work-study or assistantships) and jobs held while not enrolled, including summer break. Full-time status was defined as working 35 or more hours per week, and part-time status was defined as working less than 35 hours per week.

ORCID iD

Zoe Padgett https://orcid.org/0000-0001-6748-4060

References

Aguiar, E. (2015). Identifying students at risk and beyond: A machine learning approach. University of *Notre Dame*, Notre Dame, IN. https://curate.nd.edu/show/1v53jw8435h

American Institutes for Research. (2013). *Predictors of postsecondary success*. College and Career Readiness and Success Center. https://ccrscenter.org/sites/default/files/CCRS%20Center_Predictors%20of%20Postsecondary%20Success_final_0.pdf

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). *Predicting student dropout in higher education*. University of Washington. Retrieved from https://arxiv.org/abs/1606.06364

Barrow, L., & Rouse, C. E. (2005). Does college still pay? *Economist's Voice*, *2*(4), 1–8. http://econ.ucsb.edu/~tedb/Courses/Ec100C/Readings/barrow-rouse.pdf

Bellman, R. E. (1957). *Dynamic programming*. Princeton University Press. https://doi.org/10.1016/S0305-0548(99)00088-X

Broderson, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 International conference on pattern recognition* (pp. 3121–3124). https://doi.org/10.1109/ICPR.2010.764

Card, D. (1999). The causal effect of education on earnings. *Handbook of Labor Economics*, *3*, 1801–1863. https://doi.org/10.1016/S1573-4463(99)03011-4

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Choy, S. (2002). *Nontraditional undergraduates* (NCES 2002-012). https://nces.ed.gov/pubs2002/2002012.pdf

Delen, D. (2010). A Comparative Analysis of Machine Learning Techniques for Student Retention Management. *Decision Support Systems*, *49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Dewancker, I., McCourt, M., Clark, S., Hayes, P., Johnson, A., & Ke, G. (2016, December). A strategy for ranking optimization methods using multiple criteria. In *Workshop on Automatic Machine Learning* (pp. 11–20). http://proceedings.mlr.press/v64/dewancker_strategy_2016.pdf

Ehrlich, I. (1975). On the relation between education and crime. In F. T. Juster (Ed.), *Education, income, and human behavior* (pp. 313–338. http://www.nber.org/chapters/c3702.pdf

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Forbus, P., Newbold, J. J., & Mehta, S. S. (2011). A study of non-traditional and traditional students in terms of their time management behaviors, stress factors, and coping strategies. *Academy of Educational Leadership Journal*, *15*, 109–125. https://www.questia.com/library/journal/1G1-273616190/a-study-of-non-traditional-and-traditional-students

Fortin, A., Sauvé, L., Viger, C., & Landry, F. (2016). Nontraditional student withdrawal from undergraduate accounting programmes: A holistic perspective. *Accounting Education*, *25*(5), 437–478. https://doi.org/10.1080/09639284.2016.1193034

Gilardi, S., & Guglielmetti, C. (2011). University life of non-traditional students: Engagement styles and impact on attrition. *The Journal of Higher Education*, *82*(1), 33–53. https://doi.org/10.1080/00221546.2011.11779084

Grossman, M. (1976). The correlation between health and schooling. In N. E. Terleckyj (Author), *Studies in income and wealth: Household production and consumption* (pp. 147–224). http://www.nber.org/chapters/c3962.pdf

Guryan, J., Hurst, E., & Kearney, M. (2008). Parental education and parental time with children. *Journal of Economic Perspectives*, *22*(3), 23–46. https://doi.org/10.1257/jep.22.3.23

Hasnain, M., Levy, J. A., Mensah, E. K., & Sinacore, J. M. (2007). Association of educational attainment with HIV risk in African American active injection drug users. *AIDS Care*, *19*, 87–91. https://doi.org/10.1080/09540120600872075

Hastie, T., Tibshirani, R., & Friedman, J. (2017). The elements of statistical learning: Data mining, inference, and prediction. https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf

Hauser, R. M., & Daymont, T. N. (1977). Schooling, ability, and earnings: Cross-sectional findings 8 to 14 years after high school graduation. *Sociology of Education*, *50*(3), 182–206. https://doi.org/10.2307/2112649

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/10.1002/ir.185

Hill, J., Smith, N., Wilson, D., & Wine, J. (2016). *2012/14 Beginning postsecondary students longitudinal study (BPS:12/14): Data file documentation*. (NCES 2016-062). https://nces.ed.gov/pubs2016/2016062.pdf

Hillygus, D. S. (2005). The missing link: Exploring the relationship between higher education and political engagement. *Political Behavior*, *27*, 22–47. https://doi.org/https://doi.org/10.1007/s11109-005-3075-8

Howe, W. J. (1993). The effects of higher education on unemployment rates. In: Becker W.E. & Lewis D.R. (Eds.), *Higher education and economic growth*. Springer. https://doi.org/10.1007/978-94-015-8167-7_6

Hoyt, J., Howell, S., Touchet, J., Young, S., & Wygant, S. (2010). Enhancing nontraditional student learning outcomes in higher education. *PAACE Journal of Lifelong Learning*, *19*, 23–31. https://www.iup.edu/ace/paace/v19-2010/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. Springer Science + Business Media. https://doi.org/10.1007/978-1-4614-7138-7

Jones, B. D., & Vedlitz, A. (1993). Higher education, business creation, and economic growth in the American states. In: Becker W. E. & Lewis D. R. (Eds.), *Higher education and economic growth*. Springer. https://doi.org/10.1007/978-94-015-8167-7_8

Koropeckyj, S., Lafakis, C., & Ozimek, A. (2017). *The economic impact of increasing college completion*. American Academy of Arts & Sciences. https://www.amacad.org/multimedia/pdfs/publications/researchpapersmonographs/CFUE_Economic-Impact/CFUE_Economic-Impact.pdf

Livieris, I. E., Mikropoulos, T. A., & Pintelas, P. (2016). A decision support system for predicting students' performance. *Themes in Scienc es & Technology Education*, *9*, 43–47. http://earthlab.uoi.gr/theste/index.php/theste/article/view/209

Lleras-Muney, A. (2005). The relationship between education and adult mortality in the United States. *Review of Economic Studies*, *72*(1), 189–221. http://www.jstor.org/stable/3700689

Ma, J., Pender, M., & Welch, M. (2016, December). *Trends in higher education: Education pays* 2016. https://trends.collegeboard.org/sites/default/files/education-pays-2016-full-report.pdf

McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., Gebrekristos, S., Zhang, J., Rathbun, A., Barmer, A., Bullock Mann, F., & Hinz, S. (2017). *The condition of education 2017* (NCES 2017-144). https://nces.ed.gov/pubs2017/2017144.pdf

Metzner, B., & Bean, J. (1987). The estimation of a conceptual model of nontraditional undergraduate student attrition. *Research in Higher Education*, *27*(1), 15–38. https://www.jstor.org/stable/40195801

Mincer, J. (1991). *Education and unemployment* (NBER working paper no. 3838). http://www.nber.org/papers/w3838.pdf

Moretti, E. (2004). Estimating the social return to higher education: Evidence from longitudinal and repeated cross-sectional data. *Journal of Econometrics*, *121*(1-2), 175–212. https://doi.org/10.1016/j.jeconom.2003.10.015

Murphy, K. M., & Welch, F. (1992). Wages of college graduates. In: Becker W. E. & Lewis D. R. (Eds.), *The economics of American higher education*. Springer. https://doi.org/10.1007/978-94-011-2950-3_5

Nakhkob, B., & Khademi, M. (2016). Predicted increase enrollment in higher education using neural networks and data mining techniques. *Journal of Advances in Computer Research*, *7*, 125–140. https://www.researchgate.net/publication/312916057_Predicted_increase_enrollment_in_higher_education_using_Neural_Networks_and_Data_Mining_techniques

Oreopoulos, P., & Petronijevic, U. (2013). Making college worth it: A review of the returns to higher education. *The Future of Children*, *23*, 41–65. https://eric.ed.gov/?id = EJ1015240

Painter, J. E., Wingood, G. M., DiClemente, R. J., DePadilla, L. M., & Simpson-Robinson, L. (2012). College graduation reduces vulnerability to STIs/HIV among African-American young adult women. *Women's Health Issues*, *22*, e303–e310. http://doi.org/10.1016/j.whi.2012.03.001

Perna, L. (2003). The private benefits of higher education: An examination of the earnings premium. *Research in Higher Education*, *44*(4), 451–472. http://www.jstor.org/stable/40197315

Pontes, M. F., & Pontes, N. H. (2012). Enrollment in distance education classes is associated with fewer enrollment gaps among nontraditional undergraduate students in the US. *Journal of Asynchronous Learning Networks*, *16*, 79–89. https://eric.ed.gov/?id = EJ971041

Radford, A. W., Cominole, M., & Skomsvold, P. (2015). *Demographic and enrollment characteristics of nontraditional undergraduates: 2011–12* (NCES 2015-025). https://nces.ed.gov/pubs2015/2015025.pdf

Riddell, W. C., & Song, X. (2011). *The impact of education on unemployment incidence and re-employment success: Evidence from the U.S. labour market* (IZA discussion paper no. 5572). https://ssrn.com/abstract = 1790683

Ross, C., & Wu, C. (1995). The links between education and health. *American Sociological Review*, *60*(5), 719–745. https://doi.org/10.2307/2096319

Salinas-Jiménez, M. M., Artés, J., & Salinas-Jiménez, J. (2013). How do educational attainment and occupational and wage-earner statuses affect life satisfaction? A gender perspective study. *Journal of Happiness Studies*, *14*(2), 367–388. https://doi.org/10.1007/s10902-012-9334-6

Schneider, M., & Yin, L. (2011). *The high cost of low graduation rates: How much does dropping out of college really cost*? American Institutes for Research. https://www.collegeincolorado.org/Images/CiC/pdfs/Press_Room/high_cost_of%20low_graduation.pdf

Seni, G., & Elder, J. F. (2010). Ensemble methods in data mining: Improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *2*(1), 1–126. https://doi.org/10.2200/S00240ED1V01Y200912DMK002

Skomsvold, P., Radford, A. W., & Berkner, L. (2011). *Six-year attainment, persistence, transfer, retention, and withdrawal rates of students who began postsecondary education in 2003–04* (NCES 2011-152). https://nces.ed.gov/pubs2011/2011152.pdf

Snyder, T. D., de Brey, C., & Dillow, S. A. (2018). *Digest of education statistics, 2016* (NCES 2017–2094). https://nces.ed.gov/pubs2017/2017094.pdf

Tamborini, C. R., Kim, C., & Sakamoto, A. (2015). Education and lifetime earnings in the United States. *Demography*, *52*(4), 1383–1407. https://doi.org/10.1007/s13524-015-0407-0

Taniguchi, H., & Kaufman, G. (2005). Degree completion among nontraditional college students. *Social Science Quarterly*, *86*(4), 912–927. https://doi.org/10.1111/j.0038-4941.2005.00363.x

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

Trenz, R. C., Ecklund-Flores, L., & Rapoza, K. (2015). A comparison of mental health and alcohol use between traditional and nontraditional students. *Journal of American College Health*, *63*(8), 584–588. https://doi.org/10.1080/07448481.2015.1040409.

Willans, J., & Seary, K. (2011). I feel like I'm being hit from all directions: Enduring the bombardment as a mature-age learner returning to formal learning. *Australian Journal of Adult Learning*, *51*, 119–142. https://files.eric.ed.gov/fulltext/EJ951989.pdf

Wolbers, M. H. J. (2000). The Effects of Level of Education on Mobility between Employment and Unemployment in the Netherlands. *European Sociological Review*, *16*(2), 185–200. https://doi.org/10.1093/esr/16.2.185

Author Biographies

Huade Huo is a researcher at American Institutes for Research. His interests include the intersection of data science and quantitative public policy analysis.

Jiashan Cui is a researcher at the American Institutes for Research. Her research interests are in the area of family engagement and early childhood education.

Sarah Hein is a research associate at American Institutes for Research. She holds a B.A. in Economics from the University of Wisconsin-Madison and is currently an M.P.P. student at the George Washington University.

Zoe Padgett M.S., is a research associate at the American Institutes for Research. She holds a master's degree in Survey Methodology from the University of Maryland.

Mark Ossolinski is a former editorial assistant at American Institutes for Research. He holds a B.A. in English and Spanish from the University of Michigan and is currently a master's student at the University of Missouri School of Journalism.

Ruth Raim is a former research assistant at the American Institutes for Research and a current teacher with D.C. Public Schools. She holds a M.Ed in Reading Education from Vanderbilt University and a B.S. in Psychology from Davidson College.

Jijun Zhang PhD, PMP, is a principal researcher at the American Institutes for Research. Her research focuses on the analysis of large-scale complex survey data.