Should Colleges Invest in Machine Learning? Comparing the Predictive Powers of Early Momentum Metrics and Machine Learning for Community College Credential Completion

Takeshi Yanagiura

April 2020

CCRC Working Paper No. 118

Address correspondence to:

Takeshi Yanagiura
Senior Research Assistant, Community College Research Center
Teachers College, Columbia University
525 W. 120th St., Box 174
New York, NY 10027
212-678-3091
Email: ty2317@tc.columbia.edu

Funding for this research was provided by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the author and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation. I thank Jordan Matsudaira, Judith Scott-Clayton, Thomas Brock, Davis Jenkins, John Fink, Doug Slater, Hayley Glatter, Yukikazu Hidaka, session participants at 2018 Association for Education Finance and Policy (AEFP) annual conference, and seminar participants at CCRC's brown bag seminar for valuable comments on earlier drafts. All errors are my own.

Abstract

Among community college leaders and others interested in reforms to improve student success, there is growing interest in adopting machine learning (ML) techniques to predict credential completion. However, ML algorithms are often complex and are not readily accessible to practitioners for whom a simpler set of near-term measures may serve as sufficient predictors. This study compares the out-of-sample predictive power of early momentum metrics (EMMs)—13 near-term success measures suggested by the literature—with that of metrics from ML-based models that employ approximately 500 predictors for community college credential completion. Using transcript data from approximately 50,000 students at more than 30 community colleges in two states, I find that the EMMs that were modeled by logistic regression accurately predict completion for approximately 80% of students. This classification performance is comparable to that of the ML-based models. The EMMs even outperform the ML-based models in probability estimation. These findings suggest that EMMs are useful predictors for credential completion and that the marginal gain from using an ML-based model over EMMs is small for credential completion prediction when additional predictors do not have strong rationales to be included in an ML-based model, no matter how large the number of those predictors may be.

Table of Contents

1. Introduction

As a part of efforts to improve community college graduation rates, there is growing interest among community college leaders and others in using machine learning (ML) techniques to predict credential completion due to their purported accuracy (e.g., Mathewson, 2015; Barshay & Aslanian, 2019). ML is attractive in part because anticipating who is not on track to graduate can help institutions become more effective in targeting students for interventions. ML results could also help colleges reflect on the progress of their current reforms. Community colleges often have multiple reforms underway but little knowledge about whether these efforts are making any difference until after students have had time to graduate (Jenkins & Bailey, 2017). But perhaps ML could offer more immediate insights. If ML predicts that graduation rates are likely to increase substantially, then that is a good indication that the current reforms are on the right path. If the prediction result indicates otherwise, then leaders may want to revisit the reforms and modify them. However, while several ML-based prediction tools for credential completion have been introduced, these models are too expensive for many resource-constrained community colleges to take advantage of. Additionally, the algorithms behind these tools are often too complex to be well-understood by practitioners who might want to develop similar in-house algorithms of their own.

Alternatively, based on a review of the relevant empirical literature, researchers have proposed a simpler set of near-term outcomes as leading predictors of long-term student success (Jenkins & Bailey, 2017; Jenkins, Brown, Fink, Lahr, & Yanagiura, 2018; Belfield, Jenkins, & Fink, 2019). While the proposed measures differ slightly among studies, in this paper, I identify 13 “early momentum metrics” (EMMs): 12 binary variables and one continuous variable that gauge a student’s academic progress during the first term or year in college (Table 1). Some studies using specific data samples show that the EMMs are highly correlated with credential completion at the student level (Belfield et al., 2019) and are readily available from the student transcript data that all colleges collect. Proponents argue that the EMMs are reasonable predictors of credential completion, and if aggregated at a higher level (e.g., institution, system, or state), that the EMMs could even serve as informative indicators in the evaluation of the effects of current institutional practices on long-term outcomes (Jenkins & Bailey, 2017; Jenkins et al., 2018; Belfield et al., 2019).

Table 1
First-Year Early Momentum Metrics (EMMs)

EMM	Data Type
1: Earned 6+ college credits in the first term	Binary
2: Earned 12+ college credits in the first term	Binary
3: Earned 15+ college credits in the first year	Binary
4: Earned 24+ college credits in the first year	Binary
5: Earned 30+ college credits in the first year	Binary
6: Completed college math in the first year	Binary
7: Completed college English in the first year	Binary
8: Completed college math & college English in the first year	Binary
9: Persisted from the first term to the second term	Binary
10: College credit pass rate	Continuous
11: Attempted 15+ credits (any level) in the first term	Binary
12: Attempted 30+ credits (any level) in the first year	Binary
13: Earned 9+ credits in major subjects in the first year	Binary

However, just because EMMs correlate with credential completion at a statistically significant level does not guarantee that EMMs are sufficient predictors of credential completion. This is partly because credential completion is unlikely to be a function of first-year academic performance alone. Many other factors could be associated with student success, including nonacademic factors such as financial aid receipt and parents' level of education (Fike & Fike, 2008). Whether a narrowly defined set of academic variables alone can accurately predict who will obtain a credential requires more examination. Additionally, EMMs might be statistically significant only within a specific sample. A model that performs well on one dataset does not necessarily perform in the same way on another dataset. To determine whether EMMs are satisfactory predictors of credential completion, it is important to examine the extent to which EMMs can accurately predict credential completers using *out-of-sample* data that were not used to develop the model, instead of *in-sample* data that were used to fit the model.

According to previous research, EMMs are correlated with credential completion; however, the extent of their predictive power remains unclear. To better understand their utility, this study compares the out-of-sample predictive performance of EMMs to the predictive performance of more complex ML-based models that utilize a substantially larger number of predictors extracted from student transcripts. This study uses transcript-level data

from approximately 50,000 students enrolled at more than 30 community colleges in two anonymous states to compare the predictive performance of EMMs modeled via logistic regression to the predictive performance of three ML-based models: regularized logistic regression, decision tree, and random forest. Each model has unique methodological strengths, and, unlike logistic regression, all three can deal with a large number of predictors through regularization, which optimally sorts out only important predictors. If EMMs demonstrate strong predictive power when compared to the ML-based models, then the additional benefits colleges could gain from using state-of-the-art ML techniques over EMMs is likely to be small. In contrast, if EMMs underperform in prediction when compared to the ML-based models, then EMMs may not be sufficient predictors, indicating there is room for improvement that may not have been identified in the current literature.

I examined each model's predictive performance based on its ability to: (1) correctly predict whether a student will earn a credential or not (hereafter referred to as "classification"), and 2) correctly estimate the probability of credential completion (hereafter referred to as "probability estimation").In summary of this study's results, I found that EMMs that were modeled via logistic regression with no interaction term accurately classified approximately 80% of students' actual credential completion outcomes using out-of-sample data. In comparison, the ML-based models which used 497 predictors outperformed the EMMs only slightly in classification. In terms of probability estimation, EMMs modeled by logit outperformed decision tree-based models and random forest-based models, while showing almost identical performance as the ML-based regularized logistic regression model. These results provide an additional layer of evidentiary support to the finding of Belfield et al. (2019) that EMMs are reasonable leading predictors for community college credential completion. Additionally, I found that EMMs' predictive performance is largely unchanged by the addition of demographic characteristics, major dummies, and institutional dummies or by the development of a model at the institutional level. Hence, EMMs' predictive power is comparable across student groups and can be used reliably across institutions.

Why does ML contribute little to the predictive performance of EMMs? In principle, the performance of an ML technique improves as the number of *relevant* predictors increases and not as the *total* number of predictors increases (Dash & Liu, 1997). My findings are

consistent with this principle, suggesting that EMMs may already contain a reasonable set of relevant predictors that can be captured from a transcript data file. However, it is worth emphasizing that I simply extracted from a transcript database the GPA and the numbers of attempted and earned courses—all of which are grouped by a two-digit classification of instructional programs (CIP) code and by academic level (remedial, first-year, or second-year)—in the first fall term and the first year, respectively. A more effective method for deriving more meaningful predictors from transcript data might have been available. Additionally, this paper's findings utilize students' transcript and demographic information as the only sources of predictors. Access to new types of novel data, such as those in a learning management system (LMS), measures of students' mindset or writings, or image data that capture how students interact with faculty in the classroom, might have enabled the use of more cutting-edge ML algorithms (involving, e.g., deep learning and natural language processing), which could have altered the predictive performance of the ML models. Whether and how the predictive performance of ML models may be improved by overcoming those limitations is left for future studies.

Using EMMs, one can aggregate each student's probability of credential completion to obtain the predicted graduation rate for an institution, system, or state. Anticipating future trends in the graduation rate can help decision-makers reflect on institutional practices and determine whether those practices should be altered. However, while EMMs can predict a student's likelihood of credential completion with approximately 80% accuracy, they still fail to correctly predict completion outcomes the remaining 20% of the time. This error rate may be less of a concern at the aggregate level but could be problematic at the individual level. Additionally, EMMs are not available until the end of a student's first year, at which point it is likely too late for an institution to intervene adequately based on EMMs. For these reasons, I do not recommend using EMMs for early alerts, as argued by the proponents of EMMs (Jenkins & Bailey, 2017; Belfield et al. 2019).

This paper's contributions are twofold: First, it examines the out-of-sample predictive performance of EMMs, which have previously been examined using only an in-sample framework. Second, this study offers insights into the potential usefulness of ML in the prediction of community college students' credential outcomes using transcript data as the primary data source. Regarding the first contribution, this study demonstrates that EMMs can

accurately predict completion for approximately 80% of students using out-of-sample data, thereby supporting the claim of Belfield et al. (2019) that these near-term indicators are reasonable leading predictors for credential completion. Regarding the second contribution, this study finds that the domain knowledge that has already accumulated in the academic momentum literature leaves little room for ML to further contribute if it mostly relies on variables without strong rationales, no matter how large the number of predictors may be. A broader implication is that ML may be most useful in less well-understood areas, where prior knowledge on meaningful variables is absent.

2. Conceptual Foundation for EMMs

EMMs' conceptual foundation comes from the literature on academic momentum, which considers the speed at which a student progresses toward graduation during an early stage in or before college (Wang, 2017). Several studies have concluded that students with momentum are more likely to graduate than those without it (Attewell, Heil, & Reisel, 2012; Wang, 2017). Researchers often consider the first semester or academic year of initial enrollment as a key period for gaining momentum (Attewell et al., 2012; Attewell & Monaghan, 2016; Belfield, Jenkins, & Lahr, 2016); however, momentum could begin in high school via enrollment in more academically rigorous courses (Adelman, 1999; Adelman, 2006; Wang, 2013), or it could be attained after the first year in college.[1]

Adelman (1999, 2006) was likely the first researcher to empirically examine the degree to which academic momentum affects the likelihood of credential completion. Later, Attewell and Monaghan (2016) used a propensity score matching (PSM) method and found that community college students who attempted 15 credit hours during their first semester of college were 9.1 percentage points more likely to earn a credential than students with similar

[1] Academic momentum is also discussed in terms of the type of institution that students attend: Some argue that attending a four-year university generates more momentum than attending a community college due to the diversion effect of community colleges, in which talented students are diverted from pursuing a bachelor's degree (Brint & Karabel, 1989; Dougherty, 1994; Long & Kurlaender, 2009; Reynolds & DesJardins, 2009; Doyle, 2011; Wang, 2013; Wang, 2015). This paper, however, focuses on academic momentum during the first year at a community college.

academic and sociodemographic backgrounds who attempted 12 credit hours in their first semester. Belfield et al. (2016) also examined the effect of first-year credit hour accumulation on credential completion using PSM. They found that students who attempted at least 27 credit hours in their first year (i.e., 15 credits in fall and at least 12 credits in spring and summer combined) were 18.8 percentage points more likely to earn a postsecondary credential than those who attempted 12 credit hours in their first semester and fewer than 15 credit hours in their second semester. Doyle (2011) also used PSM to find that there is a direct relationship between the number of credit hours students attempt in their first year and the likelihood that they will transfer to a four-year institution.

Research has also indicated that passing developmental, or remedial, courses is a key step in generating academic momentum. Using Florida's administrative data to track students at the state's 28 community colleges over six years, Calcagno, Crosta, Bailey, and Jenkins (2007) found that remedial students who passed college-level math during their first year had a higher likelihood of credential completion than remedial students who did not pass it during the same period of time. Leinbach and Jenkins (2008) reached a similar conclusion using data from community and technical college systems in Washington State. Researchers have also found that taking courses in specific subjects during the first year may generate momentum. Jenkins and Cho (2014) found that students who earned at least nine credits in subjects related to their major during their first year were more likely to earn a credential or transfer to a four-year institution than students who did not.

Based on these previous correlational studies, Jenkins and Bailey (2017) developed a set of first-year academic momentum indicators, organized as easily measurable variables: *credit momentum*, earning a minimum number of credits in the first term or year; *gateway momentum*, passing college-level math and/or English courses in the first year; and *program momentum*, earning a minimum number of credits in a student's broadly defined major field during the first year. Jenkins et al. (2018) added the *fall-to-spring retention rate* and the *course completion rate*—or the ratio of the number of completed credits relative to the number of attempted credits—to the mix. The indicators—which all appear in Table 1—are referred to as early momentum metrics (EMMs); they comprise one continuous and 12 binary variables that can be observed by the end of a student's first year. Using administrative data

on community college students from three statewide systems, Belfield et al. (2019) found that EMMs are strongly associated with the likelihood of obtaining a credential.

However, just because EMMs positively correlate with credential completion does not necessarily guarantee that they can serve as satisfactory leading indicators of credential completion. EMMs contain only a narrow set of academic variables and exclude other factors that could potentially predict credential completion. For example, EMMs are not able to identify community college students' academic objectives, which may be relevant to their likelihood of credential completion. Some students seek an associate degree, whereas others may attend a college to obtain a vocational certificate or to take just a few courses for skill development. Additionally, community college students' ability to gain academic momentum is often impeded by family responsibilities; multiple part-time jobs; or a lack of support from family, peers, and/or institutions. These nonacademic factors also may affect how likely a student is to graduate. In the absence of measures that capture these potentially influential factors, the predictive power of EMMs alone for credential completion must be evaluated before they are used confidently in practice. It is also critical to use out-of-sample data, rather than in-sample data, for model validation, because a model that performs well on one dataset does not necessarily perform well on another dataset.

3. Data

This paper uses transcript-level data for fall 2011 first-time community college entrants at two statewide higher education systems over six years. For both states, I tracked students' transfer enrollment and completion patterns across states using the National Student Clearinghouse's data. Table 2 describes the predictors that I used to construct the prediction model. The total sample size is 47,956; 34,395 students (or 71% of the total) attended community colleges in state A, and 13,561 students (or 29% of the total) attended community colleges in state B. In both states, the students enrolled in college for the first time in the fall of 2011; the sample excludes high school students who took courses through dual enrollment

Table 2
Descriptive Statistics of the Dataset, Fall 2011 Cohort, States A and B

	State A	State B	Total
Panel A: Demographics and Completion Rate			
N	34,395	13,561	47,956
Female	52%	56%	53%
Nonwhite	34%	30%	33%
Adult	31%	17%	27%
Remedial ever	68%	71%	69%
Resident	97%	97%	97%
Number of institutions			> 30
6-year completion rate	21.8%	23.9%	22.4%
Panel B: Early Momentum Metrics (EMMs)			
EMM 1: Earned 6+ college credits in the first term	41.6%	51.5%	44.4%
EMM 2: Earned 12+ college credits in the first term	13.7%	17.7%	14.8%
EMM 3: Earned 15+ college credits in the first year	29.6%	39.2%	32.3%
EMM 4: Earned 24+ college credits in the first year	12.2%	16.0%	13.3%
EMM 5: Earned 30+ college credits in the first year	4.5%	4.2%	4.4%
EMM 6: Completed college math in the first year	17.0%	24.5%	19.1%
EMM 7: Completed college English in the first year	40.3%	51.1%	43.4%
EMM 8: Completed college math and English in the first year	12.5%	21.3%	15.0%
EMM 9: Persisted from the first term to the second term	64.8%	73.5%	67.3%
EMM 10: College credit pass rate	60.5%	61.6%	60.8%
EMM 11: Attempted 15+ credits (any level) in the first term	13.9%	15.6%	14.4%
EMM 12: Attempted 30+ credits (any level) in the first year	13.6%	12.5%	13.3%
EMM 13: Earned 9+ credits in subjects that are related to the major in the first year	14.2%	20.4%	16.0%
Average number of EMMs satisfied	3.23	3.86	3.35
Panel C: Other Academic Performance Variables			
College-level cumulative first-year GPA	2.02	1.95	2.00
Remedial credits attempted/total attempted credits in 1st yr.	29.0%	26.0%	28.2%
College math & Eng. credits attempted/total attempted credits in 1st yr.	15.6%	18.9%	16.5%
100-level credits attempted/total attempted credits in 1st yr.	58.2%	57.9%	58.1%
200-level credits attempted/total attempted credits. in 1st yr.	10.6%	14.9%	11.8%
STEM credits attempted/total attempted credits in 1st yr.	13.4%	18.1%	14.7%
CTE credits attempted/total attempted credits in 1st yr.	13.0%	9.9%	12.1%
College-level credits earned/total attempted credits in 1st yr.	48.2%	48.4%	48.3%
Credits earned within the same major/total attempted credits in 1st yr.	15.9%	15.7%	15.8%

programs. The demographic distributions are comparable between the two states, with one exception: The proportion of adult students is 31% in state A and 17% in state B. The average six-year completion rate is 23.9% for state B and 21.8% for state A. Hence, the class distribution of the credential outcome leans toward noncompleters in both states. I define completers as students who earned a bachelor's degree, associate degree, long-term certificate, or short-term certificate within six years of initial matriculation.

Panel B in Table 2 presents the percentage of students who attained each of the EMMs, which are all binary variables except for EMM 10, the college credit pass rate. Students in state B outperformed students in state A on all EMMs except EMM 5, earned 30+ college credits in the first year, and EMM 12, attempted 30+ credits (any level) in the first year. The average number of EMMs attained, excluding the continuous variable EMM 10, is 3.23 for students in state A and 3.86 for students in state B. At a quick glance, state B's higher credential completion rate is expected due to students' higher performance on EMMs.

Panel C presents the averages of the academic variables that are not EMMs but are used as predictors in ML-based models. They include the first-year GPA for college-level courses; the fraction of attempted remedial credit hours relative to all attempted credit hours; the fraction of attempted college-level English and math credits;[2] and the fractions of attempted first year-level credits, second year-level credits, STEM credits, and career and technical education (CTE) credits attempted during the first year. The predictors also include the average total number of college-level credits that were earned as a percentage of the total number of attempted credits, as well as the average number of credits that were earned in subjects that are related to the major as a percentage of the total number of attempted credits. One substantial difference reflected in Panel C is that state A students were more likely to take CTE courses than state B students; namely, I found that CTE courses accounted for 13% of the total attempted credits in state A versus 6.3% in state B. The remaining variables are comparable between the two states.

Meanwhile, variables employed in the ML-based prediction models also include the numbers of attempted and earned courses, as well as students' average grades during their

[2] This is different from EMMs 6, 7, and 8, which are binary variables indicating whether a student completed college-level English and/or math. This variable, on the other hand, is a continuous variable concerning attempted credit hours in college-level math and English relative to all attempted credit hours.

first year in college. In the ML-based models, these variables are grouped according to the two-digit CIP code and course level (remedial, first-year, or second-year) in the first term of enrollment and the entire first year. In addition, the ML models replaced EMMs with the continuous credit-hour variables used to derive EMMs, such as credit hours attempted and completed for both college-level and remedial-level courses in the first semester and year, respectively, and the number of credit hours earned in a major-related subject during the first year. This data arrangement releases an algorithm from the constraints of EMMs, which are human-made metrics that may not necessarily optimize the predictive power of the ML-based models. Combining these variables with the demographic variables (Panel A), the other academic variables (Panel C), and the major and institutional dummy variables results in a fully saturated ML model with a total of 497 predictors.

Multicollinearity is one concern in modeling using these variables because some of the variables are likely to capture duplicate information. In practice, the *p*-values for correlated variables tend to become too large and sometimes lose their statistical significance even if they are meaningful predictors (Taddy, 2019). While coefficient estimation is not an objective of this study, multicollinearity may adversely affect the prediction results since they rely on estimated coefficients (Taddy, 2019). This could be problematic for a traditional regression model, but it is less problematic for ML because the ML algorithms that are used in this study can mitigate its harm via regularization, which is a technique for automatically removing duplicate variables to maximize the out-of-sample predictive performance (Hastie, Tibshirani, & Friedman, 2009). This regularization function is one reason why ML is an attractive option for a model that involves many predictors.

4. Analytical Strategy

4.1 Algorithms

Higher education researchers began to use supervised ML techniques to predict student outcomes in the early 2000s (e.g., González & DesJardins, 2002; Luan, 2002; Herzog, 2005; Delen, 2010; Delen, 2011; Nandeshwar, Menzies, & Nelson, 2011; Raju & Schumacker, 2015; Aulck, Velagapudi, Blumenstock, & West, 2016). However, despite the time invested in

studying ML techniques, finding consistent evidence regarding which algorithms yield better results is difficult. Virtually all of the previous studies come to different conclusions about which algorithm performs best, thereby suggesting that methodological superiority is likely to be context-dependent and to vary by data type. Additionally, because most studies examine data from a single institution, the generalizability of the findings on methodological superiority is weak. To account for these challenges, this study employs multiple ML algorithms so that the findings are not driven by an arbitrarily selected algorithm. Specifically, I use the following four algorithms for modeling: (1) logistic regression, (2) regularized logistic regression, (3) decision tree, and (4) random forest.

Equation (1) is the mathematical expression for the logistic regression and regularized logistic regression models.

$$arg\left\{\sum_{i=1}^{n}\left[y_i(X\beta) - log\ (1 + \ e^{X\beta})\right] + \alpha\lambda_1 \sum_{j=1}^{p}\left|\beta_j\right|^2 + (1-\alpha)\lambda_2 \sum_{j=1}^{p}\beta_j^2\right\} \quad (1)$$

where $\lambda_1\ and\ \lambda_2$ are the regularization terms, X includes all covariates without interactive terms, and α signifies the extent to which the model uses LASSO over Ridge, where LASSO is used if α = 1 and Ridge is used if α = 0. LASSO penalizes the unimportant variables by setting the associated coefficients to zero, whereas Ridge assigns those predictors lighter but nonzero weights. Elastic Net essentially takes a middle approach between LASSO and Ridge by simultaneously using the two regularization terms in a single equation. Additionally, the equation reduces to a conventional logistic regression if the penalty terms are removed, namely, if both $\lambda_1\ and\ \lambda_2$ are set to 0. I trained the model by tuning $\lambda_1\ and\ \lambda_2$ and α to optimize the predictive performance using the glmnet package in R.

The decision tree algorithm relies on a classification and regression tree (CART), a nonparametric decision tree algorithm that selects a set of predictors according to the degree to which they contribute to the model. With CART, the most important predictor is at the top of tree, which branches out to the next most important predictor. The tree-splitting process is repeated until the variables that contribute significantly to the entire model have been

exhausted, at which point a Gini impurity is used as a threshold for variable addition, which is defined as follows:

$$Gini\ (t) = \sum_{i=1}^{K} p(t)\big(1 - p(t)\big) \tag{2}$$

where $p(t)$ refers to the proportion of observations that belong to class i at node t. In the case of a binary variable, the Gini impurity is maximal if the data are equally distributed between the classes and declines to zero as the distribution becomes skewed toward one of the classes.

Researchers often "prune" the tree to avoid overfitting by setting a threshold value of the complexity parameter (CP), which measures the degree of model complexity (Hastie et al., 2009). A lower CP threshold value corresponds to higher complexity, which corresponds to a higher likelihood of overfitting. I tuned CP at five levels: 0.01, 0.005, 0.001, 0.0001, and 0.00001. The model continues the addition of variables to the tree until the marginal improvement in the model falls below each of the thresholds. I used the rpart package in R to choose the CP that yields the best predictive performance.

One well-known problem with the decision tree algorithm is its tendency to overfit the data by including too many predictors even after pruning the tree. The problem of having too many predictors is often referred to as the "curse of dimensionality," which degrades the model's predictive performance (Hastie et al., 2009). To address this problem, I utilized a random forest algorithm (Breiman, 2001) known for its ability to outperform the decision tree algorithm if the sample size is sufficiently large (Ali, Khan, Ahmad, & Maqsood, 2012). I set the total number of iterations to 500, and for each iteration, the random forest algorithm selects a bootstrapped sample to create a tree. At each node, the algorithm selects the variable that best contributes to the model from a randomly selected pool of predictors. Last, at each leaf, the algorithm votes for the best class to which the data point belongs. The algorithm computes the average voting results for each data point at the very end of the final iteration.

One tuning variable is the number of predictors that are randomly chosen at each node. The random forest algorithm in R selects $\sqrt{q}$ by default, where q is the total number of variables in the model. An ideal tuning strategy when using the random forest algorithm is to conduct a grid search by trying all possible combinations of the number of trees and the

number of choice variables. However, this is computationally expensive in practice. For this reason, I tuned the predictive performance of the random forest by testing three values:$\sqrt{q}$, half of the total number of predictors in the model, and three quarters of the total number of predictors in the model. Then, I selected the value that yielded the best predictive performance in terms of the area under the curve (AUC).

4.2 Models

Using the four algorithms discussed above (i.e., a conventional logistic regression, regularized logistic regression, decision tree, and random forest), I created a total of seven models and examined how the predictive powers vary by model. The first prediction model, referred to as the demographic model, employs a conventional logistic regression algorithm with the following predictors: gender, race/ethnicity, age group, state residency, remedial status at matriculation, average income in a student's zip code, full-time enrollment in the first semester in college, 35 dummy variables of the attended institution, and 22 dummy variables of broad major categories that are declared in the first semester (See Appendix Table A1). The second model is the EMM model, which is the model of interest in this study. It also uses a conventional logistic regression but includes only the 13 EMMs. The third model—referred to as the adjusted EMM model—also employs a conventional logistic regression and uses the variables that are used in the demographic and EMM models.

The fourth through seventh models are referred to as the kitchen sink models, which use the predictors of the adjusted EMM model and all the other predictors that are available from the transcripts. There are four kitchen sink models because I use four different algorithms to model the same set of predictors: conventional logistic regression, regularized logistic regression, decision tree, and random forest. The variables extracted from each transcript file include average grades and number of courses taken and completed, which are grouped according to the two-digit CIP code and academic class level (remedial, first year, or second year) during the first fall semester and the first year. As discussed in the data section, the kitchen sink models exclude EMMs but instead include the total number of college-level, remedial, and overall attempted and earned credit hours in the first term and the first academic year, which are used to derive EMMs. This data arrangement is important so that ML models are not constrained by EMMs, which are human-made metrics: ML algorithms

may not necessarily find them optimizing the predictive performance of the model. The kitchen sink models are the most saturated, containing 497 predictors.

I built these seven models to determine how well the logit-based EMM model predicts student outcomes in comparison to the other six models. A comparison with the demographic model reveals the extent to which a student's first-year performance—as measured by the EMMs—informs the likelihood of credential completion, as well as what is predicted by the information available at matriculation. A comparison with the adjusted EMM model facilitates the determination of whether the role of EMMs differs among demographic groups. Lastly, a comparison with the four kitchen sink models—which use substantially more predictors, including a set of continuous credit-hour-related variables but excluding EMMs—indicates how EMMs perform relative to models that rely on machine intelligence to choose and derive variables for optimal predictive performance. If the kitchen sink models substantially outperform the logit-based EMM model, then there may be important hitherto unknown predictors that EMMs fail to capture.

4.3 Model Validation Approaches

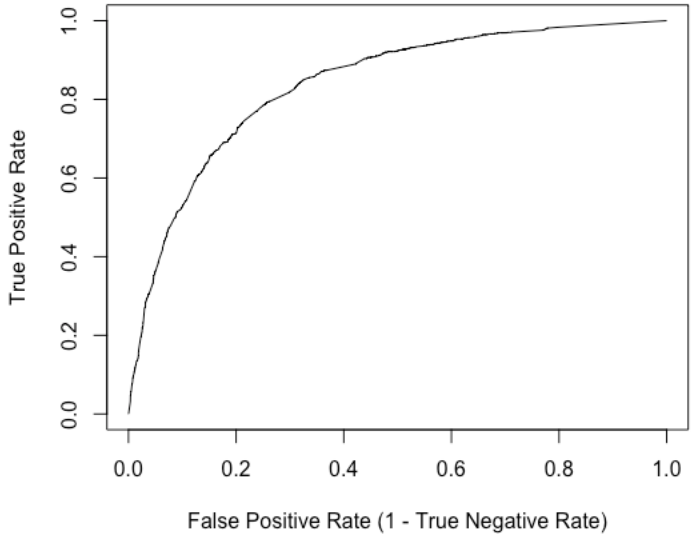
For model validation, I employed 10-fold cross-validation by dividing the entire dataset into 10 segments and using one segment as the testing dataset and the remaining segments as the training dataset. I repeated the analysis 10 times so that each segment served as the testing dataset once. When tuning was necessary, I further divided the training dataset into training and validation sets, with the former consisting of 90% of the training sample and the latter of the remaining 10% of the training sample.

I evaluated the predictive performance of each model with respect to two predictive dimensions: classification and probability estimation. The former refers to the ability to correctly predict whether a student graduates or not. The latter refers to the ability to approximate a student's probability of graduating from the true probability of graduating, which is defined as the actual graduation rate for students within a specified probability range. Because a model that performs well in terms of classification may perform poorly in terms of probability estimation and vice versa, it is important to evaluate each model's predictive performance along these two dimensions so that the conclusion is based on a more balanced perspective.

For classification, my model assessment relies on the true-positive rate, the true-negative rate, and the AUC. The true-positive rate (sensitivity) is defined as the proportion of actual completers who are correctly predicted as completers, while the true-negative rate (specificity) corresponds to the proportion of actual noncompleters who are correctly predicted as noncompleters. The area under the curve (AUC) is a measure that is commonly used to assess the overall predictive performance of a model so that a researcher can avoid overreliance on the true-positive and true-negative rates for model validation (Hastie et al., 2009).

The calculation of the AUC relies on the true-positive rate and the false-positive rate, which is equal to one minus the true-negative rate. First, I plotted both true-positive and false-positive rates on a chart called a receiver operating characteristic (ROC) space, which is defined by true-positive rates on the y-axis and false-positive rates on the x-axis (see Figure 1 for an example). In this study, I calculated all of the possible combinations of true-positive and false-positive rates using every decision-making threshold from 0% to 100% in 0.02 percentage point increments. For example, suppose a model predicts that a student's probability of graduating is 24%. This student can be predicted to be either a completer or noncompleter, depending on the decision-making threshold that is used by the researcher. If I use 20% as the cutoff, for example, this student is predicted to be a completer. Meanwhile, if the threshold is set to 45%, then the student is predicted to be a noncompleter.

Figure 1
Example of Receiver Operating Characteristic (ROC) Space



After calculating the true-positive and true-negative rates using all possible decision-making thresholds from 0% to 100%, I plotted all the combinations of true-positive and true-negative rates in the ROC space. Typically, these combinations form a concave line. In principle, AUC ranges from 0.5 to 1.0, with the former corresponding to the worst possible model performance and the latter to perfect prediction of the outcome. I consider the optimal cutoff point the value that minimizes the distance between the ROC curve and the left-top corner of the ROC space (Pepe, 2003). I repeated this process 10 times for each test sample for cross-validation. The reported true-positive and true-negative rates are the average values of the 10 true-positive and true-negative rates at the optimal cutoff point, and the reported AUC is the average of 10 out-of-sample AUCs.

To examine the model's performance in probability estimation, I used a reliability diagram (Murphy & Winkler, 1977), which is a graphical approach used to compare the average out-of-sample predicted probability of graduating to the observed graduation rate among students with similar predicted probabilities using the fall 2011 cohort, where the observed graduation rate is regarded as the true probability of credential completion. I clustered the fall 2011 students into 10 bins according to their out-of-sample predicted probabilities so that the lowest probability bin included students with predicted probabilities from 0% to 10%, while the highest probability bin consisted of students whose predicted probabilities ranged from 91% to 100%. Then, I plotted the average actual graduation rate for each bin. In an ideal scenario, these two rates should be equal so that the curve closely follows the 45-degree line (Niculescu-Mizil & Caruana, 2005a). A model is considered satisfactory if it follows the 45-degree line closely.

5. Results

5.1 Classification

Table 4 lists the true-positive and true-negative rates, as well as the AUC values for the seven models (see the previous section for a description of each model). All rates are the mean values of 10 cross-validation results, and the value in parentheses is the standard error. For the EMM model, the true-positive rate was 77.5%, while the true-negative rate was

75.4%, and the AUC was 0.835. The optimal decision-making cutoff point was 23.2%. The adjusted EMM model outperformed the EMM model in terms of predictive power on all metrics, but only slightly, thereby suggesting that EMMs are similarly satisfactory predictors for students from various demographic groups. Meanwhile, the EMM-logit model substantially outperformed the demographic model in prediction. The true-positive rate for the demographic model was 0.646; hence, the model correctly predicted 64.6% of actual completers. The true-negative rate was 64.7%, while AUC was 0.697.

Table 4
Prediction Results (Classification)

Model	Demog.	EMM	Adj. EMM	Kitchen Sink			
Method	Logistic			Logistic	Reg[a]	Tree[b]	RF[c]
True Positive	0.646	0.775	0.782	0.798	0.794	0.787	0.803
	(0.018)	(0.019)	(0.013)	(0.009)	(0.013)	(0.029)	(0.016)
True Negative	0.647	0.754	0.763	0.773	0.774	0.723	0.779
	(0.019)	(0.017)	(0.016)	(0.014)	(0.015)	(0.027)	(0.015)
AUC	0.697	0.835	0.845	0.858	0.857	0.821	0.863
	(0.006)	(0.008)	(0.007)	(0.007)	(0.005)	(0.007)	(0.006)
Cutoff	0.232	0.234	0.234	0.231	0.240	0.224	0.2746
	(0.007)	(0.011)	(0.013)	(0.016)	(0.013)	(0.026)	(0.017)
Variables Included							
Demographics	x		x	x	x	x	x
Major	x		x	x	x	x	x
Institution	x		x	x	x	x	x
EMMs		x	x				
Academics[d]				x	x	x	x
Courses[e]				x	x	x	x
# of Predictors	70	13	83	497	497	497	497

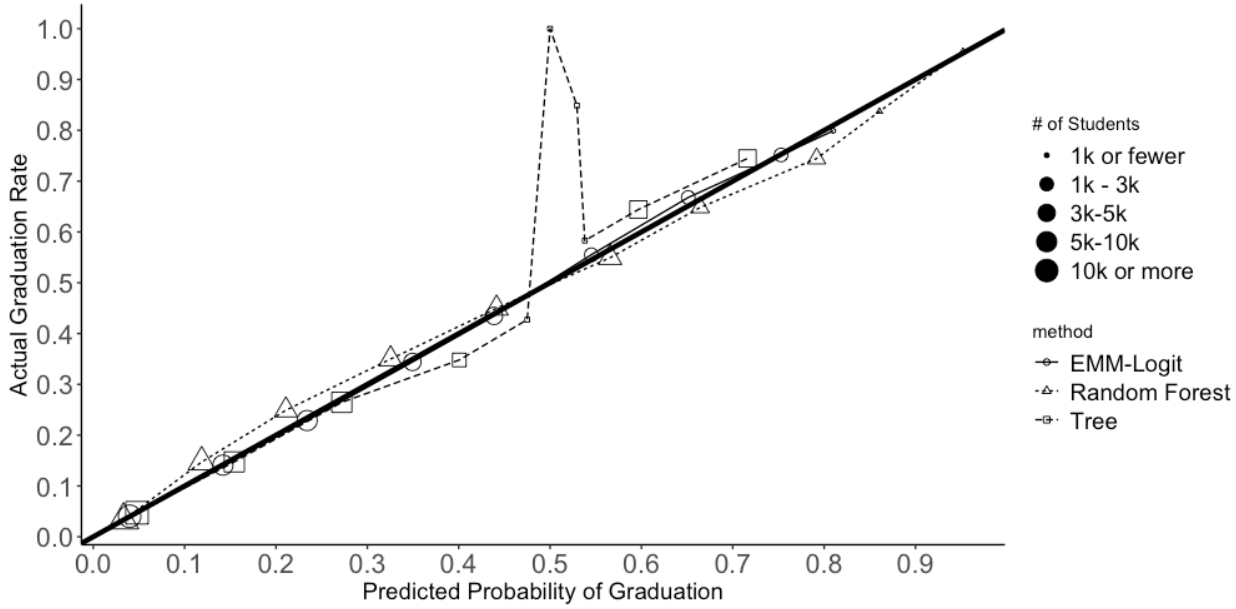
[a] Regularized logistic regression using

The kitchen sink models include 497 variables, and the three ML algorithms (i.e., regularized logistic regression, decision tree, and random forest) used the hyperparameters tuned to maximize out-of-sample predictive power as measured by the highest AUC. Tuning was not necessary for the conventional logistic regression. All four of the kitchen sink models outperformed the EMM model and the adjusted EMM model, albeit only slightly. Random forest demonstrated the best performance with an AUC of 0.863, a true-positive rate of 80.3%, and a true-negative rate of 77.9%. The other kitchen sink models, including the logit-based kitchen sink model, demonstrated similar but slightly lower levels of predictive performance. Overall, the kitchen sink models outperformed the demographic, EMM, and adjusted EMM models in prediction. However, compared to the logit-based EMM model, the best kitchen sink model (i.e., random forest) improves AUC by only approximately 0.02 (on a scale of 0.5 to 1.0, with the closer to the latter the better), and the true-positive and true-negative rates by only 2 percentage points. And so, the improvement in predictive performance that is realized by using ML with many predictors is only marginal.

Figure 2 presents a reliability diagram, which compares the observed graduation rates to the predicted probability of graduation for the EMM-logit, random forest, and tree methods. For the sake of clarity, this figure omits the results of the demographic model, which greatly underperforms compared to the other six models as expected from the classification result. For the same reason, I also omit the adjusted EMM-logit and the regularized logistic regression models whose results are almost identical to that of the EMM-logit model.

Logistic regression is designed to minimize the log-odds, which is a transformed expression of the class probability (Niculesco-Mizil & Caruana, 2005b). As a result, logistic regression algorithms are known for out-performing tree algorithms in estimating the true probability. Such properties are observed in Figure 2. The logit-based EMM model (and the kitchen sink regularized logistic regression model, which is not shown for the sake of clarity) closely follows the 45-degree line; hence, these models estimated the likelihood of graduation accurately. Decision tree widely deviates from the 45-degree line for students in the middle range of predicted probability of graduation. Random forest predicted more accurately than decision tree but still underperformed compared to logistic regression, although only slightly.

Figure 2
Reliability Diagram



In addition, Figure 2 shows that students cluster toward the lower end of the probability distribution, as signified by the size of each mark's shape, which corresponds to the number of students. Hence, most students do not accumulate any early momentum or accumulate very little early momentum. And so, helping students to build momentum by the end of the first year is a key challenge that many community colleges face.

5.2 Robustness Checks

. Two operational challenges of using the kitchen sink models are the sparsity (i.e., too many zero values in the dataset) and high dimensionality of the dataset (i.e., too many predictors) due to the inclusion of subject-level credit hour and grade data. These variables dramatically increase the dimensionality of the dataset, while the high sparsity could degrade the model's performance (Hastie et al., 2009). Since this limitation may explain why the kitchen sink models only minimally improve the predictive power over the EMM-logit model, I reduced the high dimensionality of the dataset via principal component analysis (PCA). PCA transforms the variables into a set of orthogonal variables, which are referred to as principal components. PCA also estimates the eigenvalue of each principal component, which represents the degree of variation that each

principal component accounts for in the dataset. I selected principal components that explain 90% of the total variation and entered them into the model in lieu of the course enrollment and grade variables. This technique reduced the number of predictors from 497 to 154, and, consequently, reduced dataset sparsity. Then I tested how this new dataset affects the predictive performance of the kitchen-sink random forest model, which demonstrated the best predictive performance in terms of classification in the main analysis. I found that the use of PCA slightly lowered the predictive performance for the kitchen-sink random forest model (see Table A2 and Figure A1 in the Appendix). This result suggests that the sparsity and high dimensionality of the data file are unlikely to be the reasons for the minor improvement in predictive performance by the ML-based models over the EMM model.

The results of this study that have been presented so far are based on a relatively large student-level data sample that was collected from two states. However, at the institutional level, the cohort size is substantially smaller and varies from several hundred to several thousand students. Whether a single institution with a smaller sample size can expect a similar level of predictive performance is an important question for institution-level practitioners. To address this question, Figure 3 plots the true-positive rates, true-negative rates, and AUC values for the EMM model with logistic regression and the kitchen-sink random forest model for each of the 36 institutions. I chose the random forest model for comparison because it demonstrated the best performance in the state-level analysis. A mark (in the shape of a circle or triangle) in each boxplot represents an institution-level prediction result. Each boxplot contains results from the bottom 25% to the top 75% of the corresponding measure. The horizontal line inside the box corresponds to the median, while the vertical lines that point outward from each box indicate the 5^{th} and 95^{th} percentiles of the measure. For the logit-based EMM model, the true-positive rates mostly cluster in the high-70s to the low-80s, the true-negative rates gather around the high-70s, and the AUC values are concentrated from 0.82 to 0.85. The kitchen-sink random forest model performed better than the logit-based EMM model across all metrics, but only slightly. These results are comparable to the state-level results, thereby suggesting that EMMs that were modeled via logistic regression maintain a similar level of predictive performance at the institutional level.

Figure 3
True-Positive Rate, True-Negative Rate, and AUC of the Logit-based EMM Model and the Kitchen-Sink Random Forest Model by Institution

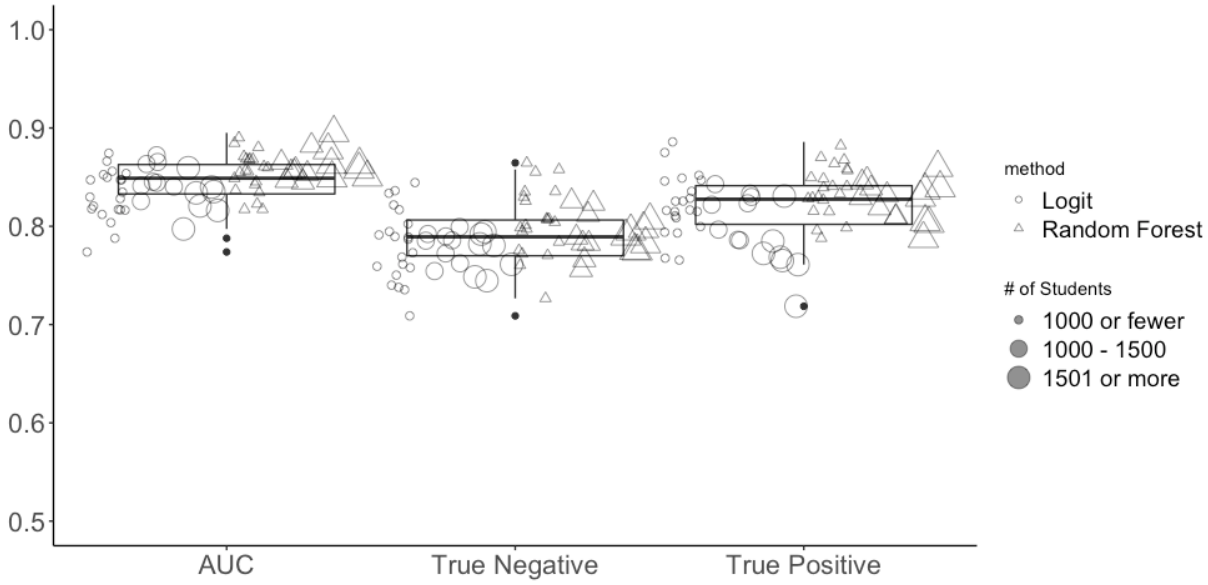
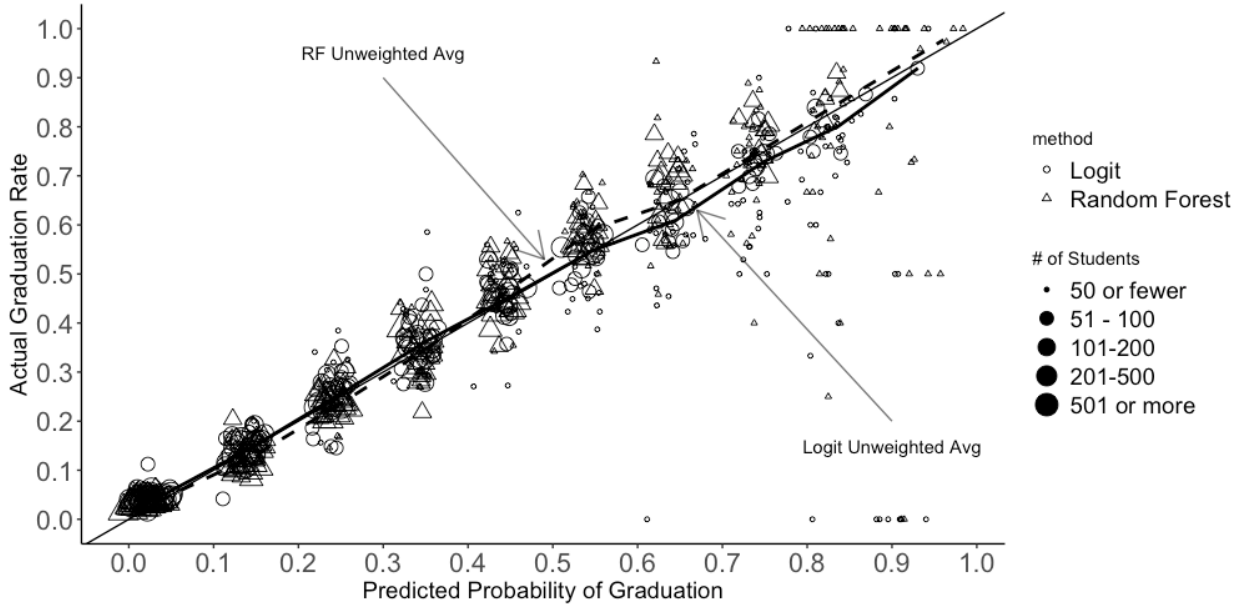


Figure 4 shows a reliability diagram by institution for the same models as in Figure 3. Each shape represents an institution, and its size corresponds to the number of students within the same probability bin, with the circle and triangle shapes corresponding to the logit-based EMM and kitchen-sink random forest models, respectively. The 45-degree line represents perfect probability estimation, and a model's performance in estimating the likelihood of graduation is considered better the closer its line is to the 45-degree line. The solid line represents the unweighted average of institutional performance based on the logit-based EMM model, and the dashed line corresponds to the same indicator for the kitchen-sink random forest model. Shapes that correspond to fewer than 30 students are excluded from the calculation of the unweighted average. The figure shows that the EMM and random forest models estimate the likelihood of credential completion with approximately equal accuracy for students with lower probabilities, whereas the latter outperforms the former in the high probability ranges, but only slightly.

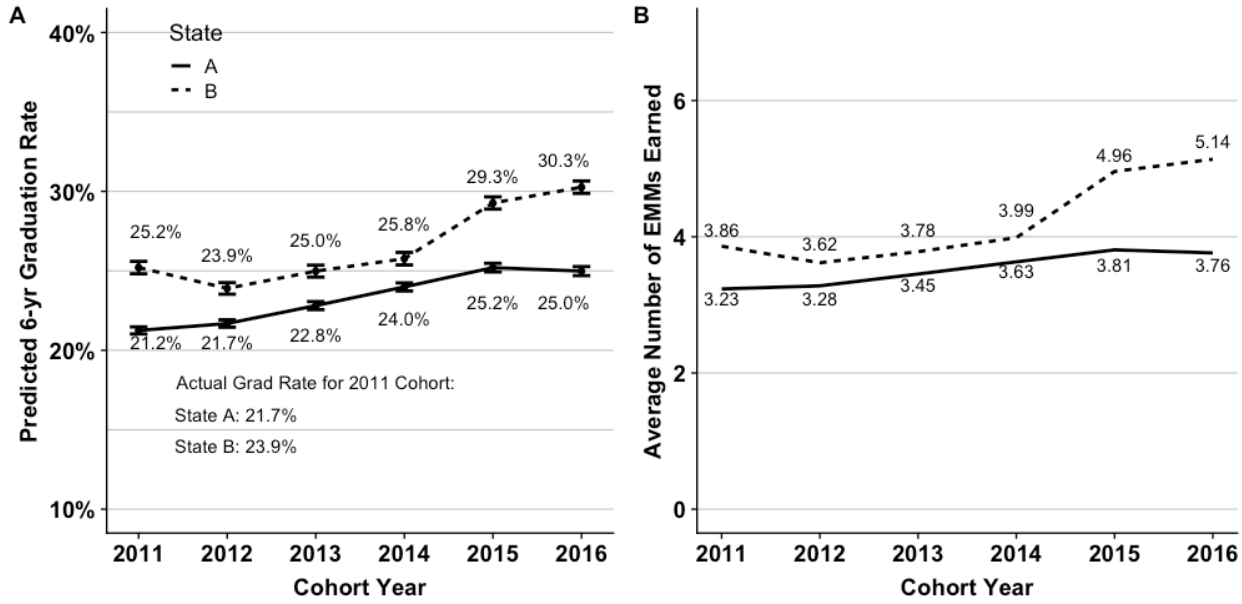
Figure 4
Reliability Diagram for the Logit-based EMM and Kitchen-Sink Random Forest Models by Institution



5.3 Applications of the Model

Decision-makers can use EMMs to predict state- or institutional-level graduation rates by averaging individual students' predicted probabilities, as presented in Figure 5, Panel A. The predicted six-year graduation rates for six recent cohorts in two states are obtained by using EMMs as predictors that are modeled by logistic regression with no interaction term. The 2011 predicted graduation rate is the average of the in-sample prediction results, whereas the predicted graduation rates for the 2012 through 2016 cohorts are the averages of out-of-sample predictions. The error bar indicates the 95% confidence interval of the predicted graduation rate. For comparison, I also present Figure 5, Panel B, which plots the actual trend in the average number of EMMs that are satisfied by each cohort by state.

Figure 5
Predicted Six-Year Graduation Rate via Logistic Regression and the Average Number of EMMs Attained for the 2011 through 2016 Cohorts, by State



According to the figure, the predicted graduation rate trends upward in both states and closely follows the progression of the average number of attained EMMs. In state A, the graduation rate is projected to increase by approximately 3.8 percentage points from the 2011 to the 2016 cohort, while the average number of attained EMMs is expected to increase by 0.53 points. In state B, the graduation rate is projected to increase by 5.1 percentage points from the 2011 to the 2016 cohort, while the average number of satisfied EMMs is expected to increase by 1.28 points. In addition, the fall 2015 cohort appears to have performed particularly well in state B, according to the sudden upticks in the trend lines for both EMMs and the graduation rate. Such perspectives would have not been available to decision-makers until years later if they had to wait until actual graduation rate data became available.

Another tempting area of application for EMMs is an early alert practice, which identifies students who demonstrate "risky" behaviors, as determined by the algorithms. However, I do not recommend using EMMs for early alerts for the following reasons: First, EMMs still fail to correctly predict graduation outcomes for 20% of students. This error rate is high for individual-level prediction. Second, EMMs become available one year after initial enrollment, which is probably too late for institutions to intervene. A reasonable early alert

system requires more accurate prediction performance and must enable an institution to intervene as early as possible, not one year after enrollment.

6. Conclusion

This study examines the predictive power of early momentum metrics (EMMs), which are 13 near-term leading indicators for community college credential completion suggested by prior literature. The study examines how this simple set of metrics performs in comparison to more complex machine learning (ML)-based prediction models that utilize substantially more predictors. Using transcript-level data from approximately 50,000 students from 36 institutions in two states, I found that EMMs that were modeled via logistic regression accurately predict a student's six-year credential completion status with an accuracy of approximately 80%. The predictive power increased only slightly when I used ML algorithms with 497 predictors. These results suggest that EMMs are reasonable, if not perfect, predictors for credential completion, which is consistent with the claim of Belfield et al. (2019). A comparison of the performances of EMMs and ML-based models also suggest—at least in the context of predicting student completion rates—that the domain knowledge that has been accumulated in the academic momentum literature leaves little room for additional insights by ML. At this point, any additional gain gleaned from using ML techniques over EMMs for the prediction of community college credential completion is small if the additional predictors have weak rationales to be included in the model, no matter how large the number of those predictors may be.

Important questions remain unanswered. For example, this study does not examine how institutions can help students gain academic momentum, which few students are able to do. The equity gap in EMM attainment among demographic groups is also large. Because the equity gap in long-term success is unlikely to be closed without addressing the EMM attainment gap in the near term (Belfield et al., 2019), it is important for future research to investigate methods by which students can improve their attainment of EMMs. Which variables could potentially increase the predictive power of the ML-based model also remains an open question. The predictive performance of the ML-based models may improve

if researchers have access to new types of data beyond transcripts. Examples include data that are accumulated in a learning management system (LMS), measures of student "mindset," students' writings, and image data that capture how faculty and students interact in the classroom. Access to data that capture students' behaviors in novel ways may provide insights that increase our ability to predict community college student completion rates.

Finally, the discussion of how to use predicted data in practice must continue. Human behaviors are difficult to predict, and no prediction is likely to be perfect, regardless of how powerful prediction algorithms may become or how much access to novel data may grow in the future. Overreliance on prediction results could also be harmful in terms of equity, since ML prediction is known for being prone to bias against individuals from less privileged backgrounds (Corbett-Davies & Goel, 2018). Prediction performance may improve over time, but the interpretation capacity of leaders as data users also must progress by the same degree.

Appendix

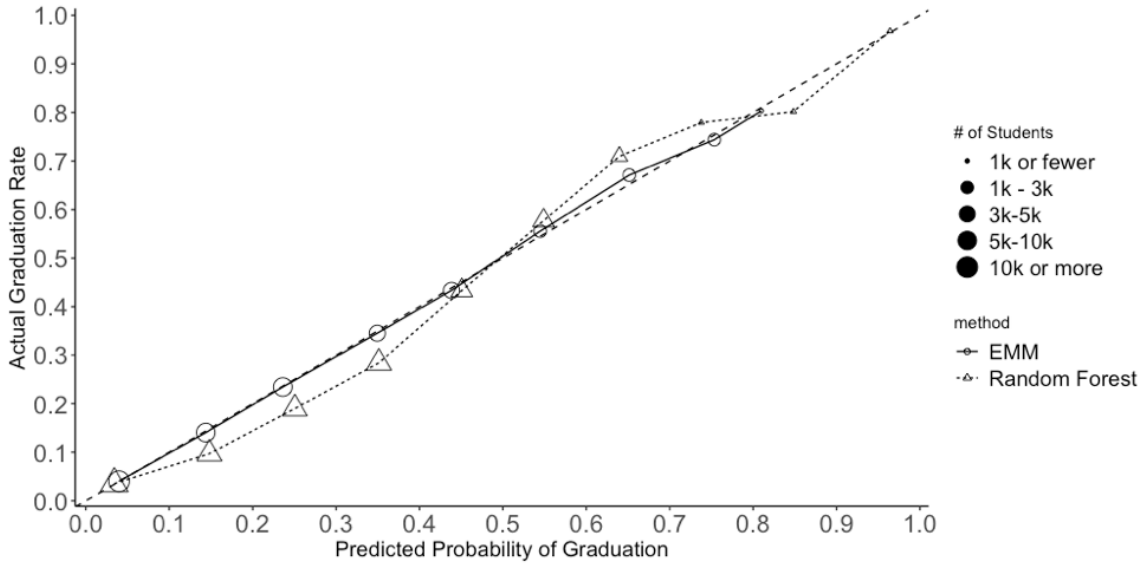
Table A1
Broad Major Categories for Six-Digit Classification of Instructional Programs (CIP) Codes

Broad Major Category	CIP Codes
1) Art, humanities, and English	9.0000 - 9.9999, 16.0000 - 16.9999, 23.000 - 24.9999, 30.0000 - 30.0099, 30.1301, 30.2101, 30.2201, 30.2202, 30.2310, 30.26 - 30.2699, 30.9999, 38.0000 - 39.9999, 50.0000 - 50.0399, 50.0500 - 50.9999
2) Mathematics and science (STEM)	26.0000 - 27.9999, 40.0000 - 40.9999, 30.0101, 30.0601, 30.1001,30.1801, 30.1901, 30.2501, 30.3200 - 30.3399
3) Social and behavioral sciences	5 - 5.9999, 22-22.0102, 22.0103 - 22.0299, 22.04 - 22.9999, 30.0501, 30.1101, 30.1501, 30.1701,30.2001, 30.1200 - 30.1299, 30.1401, 42-42.9999, 45 - 45.9999, 54 - 54.9999, 30.27-30.28
4) Agriculture and natural resources	1-1.9999, 3 - 3.9999
5) Automotive and aeronautical technology	15.0800 - 15.0899
6) Business and marketing	52 - 52.0399, 52.05 - 52.9999, 19.0505, 19.0604, 8 - 8.9999
7) Secretarial and administrative services	22.0103, 22.0301, 22.0302, 52.04 - 52.0499, 23.0303, 23.0399
8) Communications and design	10 - 10.9999, 19.0202, 19.0906, 50.04 - 50.0499
9) Computer and information sciences	11 - 11.9999, 25 - 25.9999, 30.0801, 30.1601
10) Cosmetology	12.04 - 12.0499
11) Culinary services	12.05 - 12.0599
12) Engineering and architecture	4 - 4.9999, 14-14.9999, 19.06 - 19.0603, 19.0605 - 19.0699
13) Engineering/science technologies	15 - 15.0799, 15.09 - 15.9999, 41 - 41.9999
14) Education and child care	13 - 13.9999, 19.0706, 19.0709, 20.0102, 20.0107, 20.02 - 20.299
15) Allied health	51 - 51.3799, 51.40 - 51.9999, 19.05 - 19.0599, 34 - 34.9999
16) Nursing	51.38 - 51.3999
17) Construction	46 - 46.9999
18) Manufacturing	19.09 - 19.0905, 19.0907 - 19.0999, 48 - 48.9999
19) Mechanics and repair	47 - 47.9999
20) Transportation	49 - 49.9999
21) Protective services	28 - 29.9999, 43 - 43.9999
22) Other career-technical	12 - 12.0399, 12.06 - 12.9999, 19 - 19.0499, 19.07 - 19.0705, 19.0706 19.0708, 19.0710 - 19.0899, 19.10 - 20.0101, 20.0103 - 20.0106, 20.0108 - 20.0199, 20.03 - 20.9999, 31 - 31.9999, 44 - 44.9999
23) Missing or uncategorized	

Table A2
Classification Results With Reduced Dimensions Obtained via Principal Component Analysis (Kitchen-Sink Random Forest Model Only)

Method	Random Forest (PCA)
True Positive	0.782
	(0.020)
True Negative	0.749
	(0.017)
AUC	0.837
	(0.008)
Cutoff	0.31
	(0.017)
# of Predictors	153

Figure A1
Reliability Diagram Using the Sample With Reduced Dimensions Obtained via Principal Component Analysis



References

Adelman, C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment* (Report No. PLLI-1999-8021). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, DC: U.S. Department of Education.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 272–278.

Attewell, P., Heil, S., & Reisel, L. (2012). What is academic momentum? And does it matter? *Educational Evaluation and Policy Analysis*, *34*(1), 27–44.

Attewell, P., & Monaghan, D. (2016). How many credits should an undergraduate take? *Research in Higher Education*, *57*(6), 682–713.

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). *Predicting student dropout in higher education*. Ithaca, NY: Cornell University, arXiv. arXiv:1606.06364

Barshay, J., & Aslanian, S. (2019, August). Under a watchful eye: Colleges are using big data to track students in an effort to boost graduation rates, but it comes at a cost. *APMReports*. Retrieved from https://www.apmreports.org/story/2019/08/06/college-data-tracking-students-graduation

Belfield, C., Jenkins, D., & Fink, J. (2019). *Early momentum metrics: Leading indicators for community college improvement.* New York, NY: Columbia University, Teachers College, Community College Research Center.

Belfield, C., Jenkins, D., & Lahr, H. (2016). *Momentum: The academic and economic value of a 15-credit first-semester course load for college students in Tennessee* (CCRC Working Paper No. 88). New York, NY: Columbia University, Teachers College, Community College Research Center.

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

Brint, S., & Karabel, J. (1989). *The diverted dream: Community colleges and the promise of educational opportunity in America, 1900-1985*. New York, NY: Oxford University Press.

Calcagno, J. C., Crosta, P. M., Bailey, T., & Jenkins, D. (2007). Stepping stones to a degree: The impact of enrollment pathways and milestones on community college student outcomes. *Research in Higher Education*, *48*(7), 775–801.

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning.* Ithaca, NY: Cornell University, arXiv. arXiv:1808.00023.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(1-4), 131–156.

Dougherty, K. J. (1994). *The contradictory college: The conflicting origins, impacts, and futures of the community college*. Albany, NY: State University of New York Press.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 17–35.

Doyle, W. R. (2011). Effect of increased academic momentum on transfer rates: An application of the generalized propensity score. *Economics of Education Review*, *30*(1), 191–200.

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

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.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Second Ed.). New York, NY: Springer.

Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in Higher Education*, *46*(8), 883–928.

Jenkins, D., & Bailey, T. (2017). *Early momentum metrics: Why they matter for college improvement* (CCRC Brief No. 65). New York, NY: Columbia University, Teachers College, Community College Research Center.

Jenkins, D., Brown, A. E., Fink, J., Lahr, H., & Yanagiura, T. (2018). *Building guided pathways to community college student success: Promising practices and early evidence from Tennessee*. New York, NY: Columbia University, Teachers College, Community College Research Center.

Jenkins, D., & Cho, S.–W. (2014). Get with the program… and finish it: Building guided pathways to accelerate student completion. *New Directions for Community Colleges*, *2013*(164), 27–35.

Leathart, T., Frank, E., Holmes, G., & Pfahringer, B. (2017). Probability calibration trees. *Proceedings of Machine Learning Research, 77*, 145–160.

Leinbach, D. T., & Jenkins, D. (2008). *Using longitudinal data to increase community college student success: A guide to measuring milestone and momentum point attainment* (CCRC Research Tools No. 2). New York, NY: Columbia University, Teachers College, Community College Research Center. Retrieved from https://ccrc.tc.columbia.edu/publications/research-tools-2.html

Long, B. T., & Kurlaender, M. (2009). Do community colleges provide a viable pathway to a baccalaureate degree? *Educational Evaluation and Policy Analysis, 31*(1), 30–53.

Luan, J. (2002, June). *Data mining and knowledge management in higher education - potential applications.* Paper presented at the Annual Forum of the Association for Institutional Research, Toronto, Canada.

Mathewson, T. G. (2015). Beyond student progress: How Ivy Tech is approaching data analytics. *Education Dive.* Retrieved from https://www.educationdive.com/news/beyond-student-progress-how-ivy-tech-is-approaching-data-analytics/405110

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics, 26*(1), 41–47.

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems With Applications, 38*(12), 14984–14996.

Niculescu-Mizil, A., & Caruana, R. (2005a). Predicting good probabilities with supervised learning. In L. De Raedt & S. Wrobel (Eds.), *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). New York, NY: ACM Digital Library.

Niculescu-Mizil, A., & Caruana, R. (2005b). Obtaining calibrated probabilities from boosting. In F. Bacchus & T. Jaakkola (Eds.), *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence* (pp. 413–420). Arlington, VA: AUAI Press.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction.* Vol. 28. Oxford, UK: Oxford University Press.

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

Reynolds, C. L., & DesJardins, S. L. (2009). The use of matching methods in higher education research: Answering whether attendance at a 2-year institution results in differences in educational attainment. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 24, pp. 47–97). Dordrecht, The Netherlands: Springer.

Taddy, M. (2019). *Business data science.* New York, NY: McGraw-Hill Education.

Wang, X. (2013). Modeling entrance into STEM fields of study among students beginning at community colleges and four-year institutions. *Research in Higher Education, 54*(6), 664–692.

Wang, X. (2015). Pathway to a baccalaureate in STEM fields: Are community colleges a viable route and does early STEM momentum matter? *Educational Evaluation and Policy Analysis*, *37*(3), 376–393.

Wang, X. (2017). Toward a holistic theoretical model of momentum for community college student success. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (Vol. 32, pp. 259–308). Cham, Switzerland: Springer.