Using machine learning to predict physics course outcomes

Cabot Zabriskie, Jie Yang, Seth DeVore, and John Stewart[*]

Department of Physics and Astronomy, West Virginia University, Morgantown, West Virginia 26506, USA

(Received 23 April 2019; published 28 August 2019)

The use of machine learning and data mining techniques across many disciplines has exploded in recent years with the field of educational data mining growing significantly in the past 15 years. In this study, random forest and logistic regression models were used to construct early warning models of student success in introductory calculus-based mechanics (Physics 1) and electricity and magnetism (Physics 2) courses at a large eastern land-grant university. By combining in-class variables such as homework grades with institutional variables such as cumulative GPA, we can predict if a student will receive less than a "B" in the course with 73% accuracy in Physics 1 and 81% accuracy in Physics 2 with only data available in the first week of class using logistic regression models. The institutional variables were critical for high accuracy in the first four weeks of the semester. In-class variables became more important only after the first in-semester examination was administered. The student's cumulative college GPA was consistently the most important institutional variable. Homework grade became the most important in-class variable after the first week and consistently increased in importance as the semester progressed; homework grade became more important than cumulative GPA after the first in-semester examination. Demographic variables including gender, race or ethnicity, and first generation status were not important variables for predicting course grade.

DOI: 10.1103/PhysRevPhysEducRes.15.020120

I. INTRODUCTION

Workforce demand for science, technology, engineering, and mathematics (STEM) graduates has grown significantly over the past decade, with the number of jobs requiring at least a STEM bachelor's degree growing to comprise around 20% of the workforce [1]. This growth in the STEM job sector has put significant pressure on universities to increase the number of students who graduate with STEM degrees. In their 2012 report, the President's Council of Advisors on Science and Technology [2] emphasized the need to improve retention of STEM students to avoid a projected 1 million STEM job candidate shortfall over the next decade. Despite the recognized importance of improving STEM graduation rates, only 40% of STEM majors successfully complete their degrees [2].

In a 2014 report on STEM attrition, the U.S. Department of Education found a wide range of attrition (defined as leaving a degree or university) rates across STEM disciplines ranging from a high of 59% for computer or information science majors to a low of 38% for mathematics majors, with an average rate of 48% [3]. The attrition rate was much lower than that of students in humanities, health science, and education, whose attrition rates range from 56% to 62% and is comparable to those of students in business and social or behavioral science [3]. While STEM majors are retained at a higher rate than students in other disciplines, the projected STEM degree shortfall suggests additional steps should be taken to retain more STEM majors.

Improving STEM degree retention is not a new problem for educational research with many studies exploring this issue [3–10]. These studies have often found similar results showing measures of prior preparation such as high school GPA (HSGPA) and ACT or SAT scores coupled with student performance once arriving at college measured by successful credit completion and college GPA (CGPA) produce statistically significant models of persistence.

Introductory physics courses along with introductory mathematics and chemistry courses form key early college hurdles for many STEM majors. While many factors affect the retention of students to STEM degrees, academic success in college classes must be viewed as of central importance to college completion. As such, promoting student success in core STEM courses may be one path to improving STEM retention. In the last 20 years, machine learning has been used to provide new insights into retention [11–15]. Machine learning techniques have only recently begun to be implemented in physics education research (PER) to understand the retention of physics students [16]. At the same time, the use of reformed instruction as an effective means of

[*] jcstewart1@mail.wvu.edu

decreasing failure rates in STEM courses has grown [17]. Docktor and Mestre provide a thorough overview of the use of reformed instruction in PER in their 2014 synthesis of the field [18]. Studies of the effect of reformed instruction have usually focused on either course grade [19–21] and/or student gains on conceptual instruments [20,22–24] as measures of student success. This work develops models for the early semester identification of students at risk of receiving a grade of “C” or lower in introductory physics; these models could be used to direct reformed instructional interventions toward at-risk students, thus further improving student success and retention.

A. Research questions

This study seeks to answer the following research questions:

RQ1: How well can introductory physics course grades be predicted early in the semester?

RQ2: What variables are most important for the accurate prediction of physics course grades early in the semester?

B. Educational data mining

Educational data mining (EDM) is a field that uses statistical, machine learning, and data mining techniques to understand large systems of educational data. Unlike data mining in other fields, such as business or genetics, EDM encompasses predictive modeling and integration with education research techniques such as psychometric modeling [15]. In 2014, Peña-Ayala reviewed 240 EDM studies published between 2010 and 2013; 88% used probability, machine learning, and statistics as their analysis method [25]. Studies evaluating student performance, whether in-class or overall, comprised 21% of 240 studies and were the second most common type of study. The growth in EDM has led to many universities adopting systems utilizing these methods to improve their course outcomes and in-term retention of STEM students [26].

A number of attempts have been made to classify the methods used in EDM. In 2008, Romero *et al.* [27] identified logistic regression, decision trees, random forests, neural networks, naive Bayes, support vector machines, and K-nearest neighbor algorithms as the most commonly applied EDM methods. In 2014, Peña-Ayala [25] examined the methods used in EDM. Classification was used in 42% of the studies while either clustering or regression were used in an additional 42% of studies. Decision trees and logistic regression were used in 18% of works with only Bayes theorem analysis employed more frequently in 20% of the studies. A review by Shahiri *et al.* of the prediction of students’ academic performance using data mining techniques [28] compared 5 major algorithms applied in 30 studies published between 2002 and 2015: decision trees, neural networks, naive Bayes, K-nearest neighbors, and support vector machines. Neural networks and decision trees were the most commonly used techniques. In both reviews, most studies focused on overall academic performance and not on course-level performance.

The methods used in the current study, decision trees, random forests (a method using many decision trees), and logistic regression, will be discussed in detail in Sec. II. Additional information on other machine learning techniques may be found in a number of machine learning texts [29,30].

C. EDM and grade prediction

There have been several studies in EDM that produced models that predicted student grades in undergraduate courses. Huang and Fang [31] used linear regression, multilayer perceptron network modeling, radial basis function network modeling, and support vector machines to predict performance on the final exam in a high-enrollment core engineering course. This study used CGPA, performance in 4 prerequisite courses (Statics, Calculus 1, Calculus 2, and Physics 1), and scores on the three in-semester exams as independent variables and found minimal differences in the accuracy of models constructed using different algorithms. They recommended using ordinary least squares regression with only CGPA as an independent variable to predict average class performance. However, for individual student grade prediction, they found support vector machines with CGPA, all four prerequisite course grades, and the results of the first in-class examination as independent variables produced the best model. Marbouti, Diefes-Dux, and Madhavan [32] built predictive models with variables measured in the class using six different algorithms: logistic regression, support vector machines, decision trees, multilayer perceptron networks, naive Bayes, K-nearest neighbors, and a final ensemble model consisting of the three most successful individual models. Their study predicted course performance at week 5 of the semester when homework, quiz, and test 1 scores were available; defined success as earning a grade of C or better; and studied a first-year engineering course. They found logistic regression and an ensemble model combining support vector machines, K-nearest neighbors, and naive Bayes to be superior with a prediction accuracy of 94% for logistic regression and 92% for the ensemble model. In 2010, Macfadyen and Dawson [33] mined data from the course learning management system (LMS) of an undergraduate biology course to identify 15 variables with significant correlation to final grade. Logistic regression models predicted students at risk of failure (final grade of less than 60%) with 70% accuracy and correctly identified students that failed the course (final grade less than 50%) with 81% accuracy.

II. METHODS

A. Context for research

This study was performed in the introductory physics classes at a large eastern land-grant university serving approximately 30 000 students. The undergraduate population had ACT scores ranging from 21 to 26 (25th to 75th percentile) [34]. The overall undergraduate demographics were 79% White, 4% Hispanic, 7% international, 4% African American, 4% Hispanic, 4% students reporting with two or more races, and other groups each with 1% or less. The sample was primarily male (81%) [34].

Data were collected from both Physics 1 (introductory calculus-based mechanics) and Physics 2 (introductory calculus-based electricity and magnetism). These courses are required for most physical science and engineering majors at the institution. The structure of the courses was similar, but not identical, for the period studied. Each course was led by a single experienced instructor who implemented research-based instructional practices for the period studied. One to two additional lecture instructors worked closely with the course lead each semester and replicated his or her instructional practices. The courses met for three 50-min lectures and one 3-h laboratory each week. Each lecture used Peer Instruction with clickers [35] to engage students in conceptual learning; the grades for the student's clicker responses were based on participation and are called lecture quiz grades. Lab sessions featured a mixture of inquiry-based hands-on activities, conceptual white boarding activities, group problem solving, and traditional experiments. Students received a grade for completing the laboratory (LabGrade) and also completed a graded quiz (LabQuiz). Lab and lab quiz grades disaggregated by class and week were only available in Physics 2. Both classes assigned homework each week which was a mix of conceptual and quantitative problems and was graded to provide the variable HwkGrade. Physics 1 used an online homework system that assigned problems from a popular textbook and allowed multiple attempts for each problem. Physics 2 assigned problems to be worked on paper which were graded by teaching assistants; the problems were written specifically for the class. All grade variables were cumulative; for example, the week 4 homework grade was the student's average homework grade in the first 4 weeks of class. To measure changes in conceptual understanding, Physics 1 administered the Force and Motion Conceptual Evaluation (FMCE) [36] as a pretest and post-test; Physics 2 administered the Conceptual Survey of Electricity and Magnetism (CSEM) [37] as a pretest and post-test. Only the pretest scores were used in this study.

B. Sample

The Physics 1 sample was collected over four semesters from fall 2015 to spring 2017 in which time 1588 students enrolled in the course. For both classes, for students taking the course more than once, only their final attempt was retained; any records with missing data were also removed. For Physics 1, this left 915 complete records that form the Physics 1 sample for this study. The Physics 2 sample was collected from fall 2015 to spring 2017 in which time 1282 students enrolled. The data were filtered in the same manner leaving 805 complete records. Most students were removed for either missing pretest scores, missing HSGPA, or missing ACT or SAT scores. Most Physics 2 students have also taken Physics 1; the restriction to complete records removed students who did not have a grade for Physics 1.

C. Variables

The variables used in this study were drawn from institutional records and from variables collected within the physics classes and are summarized in Table I. The in-class variables were described in Sec. II A. The institutional variables are defined in Table I. A few variables require additional explanation. The variable MathEntry measures the first mathematics class the student enrolled in at the institution. It has three levels: "Calculus" for students who first enrolled in Calculus 1 or a more advanced mathematics class, "Algebra" for students who first enrolled in College Algebra, and "Pre-Calculus" for students who first enrolled in a class between College Algebra and Calculus 1. The variable STEMHrs captures the number of credit hours of STEM classes completed before the start of the course modeled. STEM classes include mathematics, biology, chemistry, engineering, and physics classes.

D. Classification models

Classification models attempt to predict categorical outcomes. This study predicts the dichotomous outcomes "P1Grade" and "P2Grade" where students who received an "A" or "B" in Physics 1 were coded as P1Grade $= 1$ while students who received a lower grade were coded as P1Grade $= 0$. Similar coding was used for Physics 2 to produce P2Grade. The models constructed "classify" students into one of these two categories for each course. For example, the logistic regression classifier predicts the probability that a student will be measured with P1Grade $= 1$. If this probability is greater than 0.5, then the classification model assigns that student to the class P1Grade $= 1$ otherwise the student is assigned to the class P1Grade $= 0$.

To construct a classification model, the full dataset is split into two subsets: the training and test datasets. This is done by randomly sampling the full dataset without replacement. The training dataset is used to construct or "train" the models, while the test dataset is reserved for the purpose of evaluating the model performance when classifying "new" data. As much data as possible should be allocated to the training dataset to ensure the creation of the most accurate model while retaining sufficient data in the test dataset for accurate characterization of model performance. For this work, 62% of the data were allocated to the training dataset and 38% to the test dataset; splits as low as 50% test, 50% training have been shown to provide accurate results [38]. This choice was made to retain

TABLE I. Full list of both institutional and in-class variables. An X in either the Physics 1 or Physics 2 columns denotes that the variable was available for that dataset. True is abbreviated T, false, F.

Institutional variables			
Variable	Physics 1	Physics 2	Description
Gender	×	×	Gender (Men = 1 Women = 0).
InState	×	×	Student is resident of the state where the institution is located (T = 1, F = 0).
URM	×	×	Student does not identify as White non-Hispanic (T = 1, F = 0).
MathEntry	×	×	First math class taken (Calculus, Pre-Calculus, and Algebra).
FirstFall	×	×	Started in a fall semester (T = 1, F = 0).
FirstGen	×	×	Student is a first generation college student (T = 1, F = 0).
CmpPct	×	×	Percentage of hours attempted that were completed at the start of course.
CGPA	×	×	College GPA at start of course.
STEMHrs	×	×	Number of STEM (Math, Bio, Chem, Eng, Phys) credit hours completed at start of course.
HrsCmp	×	×	Total credits hours earned at start of course.
HrsEnroll	×	×	Total credits hours enrolled at start of course.
P1Grade		×	(Dependent Variable) Grade for last Physics 1 attempt (A or B = 1, CDFW = 0).
P1Atmp		×	Physics 1 attempted more than once (T = 1, F = 0).
P2Grade		×	(Dependent Variable) Grade for last Physics 2 attempt (A or B = 1, CDFW = 0).
HSGPA	×	×	High school GPA.
ACTM	×	×	ACT or SAT mathematics percentile.
ACTV	×	×	ACT or SAT verbal percentile.
Cal1Grade	×	×	Grade for last Calculus 1 attempt (A or B = 1, CDFW = 0).
Cal1Atmp	×	×	Calculus 1 was attempted more than once (T = 1, F = 0).
APCount	×	×	Number of courses where AP credit was received.
APCredit	×	×	Number of credits hours received for AP tests.
TransCnt	×	×	Number of courses where transfer credit was received.
TransHrs	×	×	Number of credits hours received for transfer courses.

In-class variables			
Variable	Physics 1	Physics 2	Description
FMCEPre	×		Percentage score on the FMCE pretest.
Test 1	×	×	Percentage score on the first exam of the semester.
CSEMPre		×	Percentage score on the CSEM pretest.
LecQuiz	×	×	Average grade on the lecture quiz by each week.
LabQuiz		×	Average grade on the lab quizzes by each week.
HwkGrade	×	×	Average grade on the homework by each week.
LabGrade		×	Average grade for the laboratory by each week.

approximately 300 students in the test dataset for each class. This choice was fairly arbitrary and many other criteria for selecting the test dataset size could have been made and should yield similar results.

After a model is constructed, it is then used to make predictions of the outcomes of the test dataset producing a matrix containing the frequency of prediction outcomes called the confusion matrix [39] as shown in Table II. The confusion matrix compares the predicted outcome to the actual outcome for the test dataset. The on-diagonal terms represent correct predictions and off-diagonal terms incorrect predictions. The sum of the entries in the confusion matrix is the size of the test dataset, N_{test}.

TABLE II. Confusion matrix.

	Actual negative	Actual positive
Predicted negative	True negative	False negative
Predicted positive	False positive	True positive

Many different statistics characterizing prediction accuracy can be computed from the confusion matrix; in this work model accuracy, the fraction of correct predictions given by

$$\text{accuracy} = \frac{\text{True Neg.} + \text{True Pos.}}{N_{\mathrm{test}}}. \tag{1}$$

Accuracy can be misleading because models with substantial accuracy can be constructed by pure guessing. For example, if the sample has an outcome equally balanced between two classes, a classification model which assigns all individuals to the same class will have an accuracy of 50%. To compensate for this effect, Cohen's κ was developed to provide a measure of accuracy normalized to the baseline accuracy of random chance [40]. Cohen's κ is given by

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}, \tag{2}$$

where p_0 is the model accuracy and p_e is the probability of randomly guessing the correct classification. Model fit is classified as follows: less than 0.2 as poor agreement, 0.2 to 0.4 as fair agreement, 0.4 to 0.6 as moderate agreement, 0.6 to 0.8 as good agreement, and 0.8 to 1.0 as excellent agreement [41].

The final method of characterizing the quality of the models constructed in this paper is the receive operating characteristic (ROC) curve, a technique originally developed to determine if radar receivers were accurately detecting aircraft. The ROC curve is constructed by plotting the true positive rate against the false positive rate for all values of the decision threshold from 0 to 1. Each classification model produces a value for the probability of each outcome. The decision threshold is the probability where a particular outcome would be selected; it is not always optimal to select at a threshold of 0.5. The ROC curve provides a measure of the model's discrimination between the outcomes as measured by the area under the curve (AUC). In a model that is no better than guessing, the AUC will be 0.5 and the ROC curve will be a straight line. A model with perfect discrimination characteristics would have an AUC of 1.0 [39,42]. Hosmer *et al.* [42] suggest an AUC threshold of 0.80 for excellent discrimination. Examples of ROC curves are presented in the Supplemental Material [43].

A baseline model was created for each sample by predicting all students in the test dataset would have the most common outcome in the training data set. For example, in Physics 1 63% of the students received an A or B, the baseline models classifies all students as students who will receive an A or B producing a classification that is 63% accurate, but represents pure guessing.

E. Classification methods

Two statistical techniques were used in the prediction of student grades in this paper: logistic regression and random forests. Several different classification methods were examined: logistic regression, K-nearest neighbors, classification and regression trees, naive Bayes, support vector machines, and random forests. Logistic regression and random forests were ultimately selected. Logistic regression was often selected as the best model in the literature and random forests represent one of the most commonly used techniques in EDM; each has its own unique advantages and disadvantages which complement each other. These will become evident in the following sections.

1. Logistic regression

Logistic regression represents one of the most widely used classification methods. In logistic regression, the probability $P(Y = 1)$ of a binary dependent variable $Y = 0$, 1 is predicated by a set of independent variables X_i. The probability, which is restricted to the range [0,1], is first projected on the range [0,∞] by calculating the odds; odds $= P/(1 - P)$. The odds are projected into an unbounded range by taking the logarithm. Logistic regression, then, employs methods related to ordinary linear regression to minimize the error by selecting an optimal set of regression coefficients, β_i:

$$\ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i, \tag{3}$$

where k is the number of independent predictor variables. The odds of $Y = 1$ is found by exponentiating Eq. (3). If all other variables are constant, a unit increase in X_1 multiplies the odds by e^{β_1}, the odds ratio. We report the odds ratio instead of the regression coefficient, β_i. The underlying statistical assumptions of logistic regression are different than ordinary least squares and models are estimated using maximum likelihood techniques.

The odds ratio for each coefficient multiplies the base odds, the odds when all coefficients except the intercept are zero. If the odds ratio is above one, then the odds ratio minus one multiplied by 100 is equal to the percentage increase in the odds. For example, if the odds ratio $e^{\beta_1} = 1.4$ then an increase in X_1 by one unit increases the odds of receiving an A or B by 40%. If the odds ratio is less than 1, the odds ratio is inverted before subtracting 1 and multiplying by 100 to yield the percentage decrease in odds. For example, if $e^{\beta_1} = 0.25$, then an increase in X_1 of one unit decreases the odds of receiving an A or B by $(1/0.25 - 1) \times 100\% = 300\%$ [29].

Logistic regression requires certain assumptions to be met to produce valid results. The dependent variable or outcome must follow a binomial distribution [42]. Outcomes must also be statistically independent and the continuous independent variables must be related linearly to the log odds. Logistic regression is not robust against collinearity in the independent variables and as with linear regression, models with high multicollinearity may be problematic.

For each logistic regression model or logistic model, each variable was fit independently in a univariate logistic model. The fits for each were evaluated with a very liberal screening criterion, retaining variables with p value of 0.20 to 0.25 to filter out the least important variables. Once the variables meeting the screening criterion were selected, a logistic model was constructed using all the selected variables and then pruned stepwise to the most parsimonious model by retaining variables with $p < 0.05$ [42]. This model is called the "optimal" model in the present work. Results were, then, confirmed using a separate backwards step-wise process that minimized the Akaike information criterion (AIC) of the model using the "stepAIC" function from the R "MASS" package [44], which stepwise removes parameters from the model until the removal of parameters no longer significantly decreases the AIC.

For the logistic modeling of the weekly data, the in-class-only model was fit first. Once the optimal in-class model was found, the institutional variables selected in the optimal institutional logistic model were then added to these models and the pruning process was performed again to select the best fitting combined model.

2. Decision trees

In order to understand random forests, it is first necessary to understand the decision trees upon which they are built. Decision trees are a machine learning algorithm which splits the dataset into two or more "most" homogeneous groups based on the measured variables. The algorithm works by taking the dataset, splitting it by each of the independent variables, and then measuring the degree to which these splits have created subsets of data which are maximally homogenous by outcome (each split should contain as large a percentage of one outcome as possible). Each split subset is then split using the same criteria producing a tree where each node is characterized by the criteria used to make the split (the decisions). If allowed, this algorithm will continue to split the data until each terminal node or leaf is perfectly homogenous (contains only one of the possible outcomes). To ensure that the model is not overfit, decision trees are "pruned" back to a model that balances the complexity of the model with the predictive power of the model. Because the models are based solely on homogenizing outcomes, decision trees are not susceptible to multicollinearity [45].

3. Random forests

Random forests are an extension of the decision tree algorithm where, instead of a growing single tree for the model, thousands of trees are grown. This "forest" of decision trees is used to fit the data and then the ensemble "votes" for the most likely outcome. Each decision tree is used to classify each participant and the most commonly occurring classification of the forest is selected. Random forests are a bootstrapping technique where the individual trees are grown on Z samples of size N generated by sampling the training dataset with replacement. Each of these Z unique samples is fit with a decision tree which uses a subset $m = \sqrt{k}$ of the available variables k [46]. Using a subset of the available parameters accomplishes two goals: first, the trees are decorrelated from each other, and second, the strongest predictors are prevented from always overwhelming some of the weaker predictors. With the exception of these differences, the trees are grown using the same technique as in the previous section without pruning [29]. Random forests also provide a characterization of the relative importance of the independent variables through a number of variable importance indices which indicate the degree to which each variable is influential in the model. The "randomForest" package in R produces two commonly used versions of variable importance: mean decrease in Gini index and mean decrease in accuracy [47]. This paper reports the more intuitive mean decrease in accuracy. The mean decrease in accuracy is determined as the average decrease in accuracy across all trees if the variable was removed [46].

In order to construct the optimal random forest models, 10 000 decision trees were constructed using all available variables. The 1-SE rule [45] was then applied to select the "optimal" random forest model. This was accomplished by fitting the ROC curve for the full model, then removing variables and selecting the most parsimonious model whose ROC curve was within 1-SE, 1 standard error, of the full model.

For the random forest analysis in this work, the optimal in-class model was found and the variables from the institutional random forest model were added; however, no pruning was done due to the limited number of parameters in the model.

F. Opening the "black box" of machine learning: Local interpretable model-agnostic explanations (LIME)

While machine learning provides powerful tools for understanding large datasets, the algorithms work as "black boxes" where the model builds itself. Even if the model fits well and predicts the test dataset adequately, it is difficult to extract additional meaning from the predictions. Ribeiro, Singh, and Guestrin [48] developed a method for explaining how any machine learning model makes its predictions by assuming that locally all models behave linearly. Their local interpretable model-agnostic explanations (LIME) algorithm works by selecting the record of a single student and fitting a sparse linear model to predict that one outcome. The process then perturbs the model around that fit and uses data from other students close by to determine the degree to which each variable is necessary to the predicted outcome. Using this method, it is possible to see which variables influence a model's decision and the degree of importance these variables have to that decision. This can be used to determine if a model is making decisions based upon variables that make sense (for example, a student with a high CGPA is more likely to graduate) rather than spurious correlations (for example, not attending class increases the odds of course success).

The LIME algorithm can also be used diagnostically to determine the features of an individual student's performance which are most predictive of their success or failure.

III. RESULTS

The goal of this work was to understand the variables important to the prediction of physics course grades at six time points early in two introductory physics classes: before the class begins using only institutional data and at the end of weeks 1 through 5. The results for Physics 1 and 2 are discussed separately.

A. Physics 1 classification accuracy

As the Physics 1 class progresses and the students complete assignments, the classification models become more accurate. Table III reports the accuracy, kappa, and AUC for the in-class models and the combined institutional and in-class models for week 1 through week 5 as well as the institutional model using data available before students begin the course. For the institutional model, both results using all variables and results pruning the model to an optimal model are presented. Figure 1 plots the evolution of these quantities for the weekly models. The random guessing baseline model is also presented; the horizontal axis represents the performance of the baseline model in each plot in Fig. 1.

Using logistic regression, the third week of the semester was the first where the models using only in-class variables outperformed the model using only the institutional variables with superior kappa and AUC. We will focus on kappa and AUC because, unlike accuracy, these two measures correct for the effect of guessing. In weeks 1 through 4, the combined in-class and institutional models outperformed each of the in-class only models for that week; however, this was not the case in week 5. The week 5 logistic models were statistically indistinguishable using DeLong's test comparing the AUC of the separate ROC curves, indicting that by the first in-semester examination the institutional variables were no longer necessary for the prediction of student grades. The weekly logistic models demonstrate the importance of including the institutional data if student risk is to be accessed very early in the semester.

The random forest results in Table III show that the random forest models performed almost identically to the logistic models. Unlike the logistic models, the kappa of the in-class-only random forest model only exceeded the optimal institutional model in week 5. The optimal in-class-only model AUC exceeded the institutional model AUC in week 3, but the two AUC values were equal in week 4. In week 5, the in-class-only model once again was no longer distinguishable from the combined institutional and in-class model.

B. Physics 1 variable importance

1. Institutional model

For Physics 1, the optimal logistic model of the institutional data required only a small subset of the available variables, as shown in Table IV. This optimal model had acceptable, but average, fit statistics with an accuracy of 70%, an improvement of 7% over the baseline model, and Cohen's κ of 0.32 representing fair agreement of the model with the test dataset [41]. The ROC curve produced an AUC of 0.79, with a 95% confidence interval that included the 0.80 threshold for excellent discrimination. Each logistic model reports the odds-ratio and its 95% confidence interval using unstandardized variables. For example, in Table IV a one-point increase in CGPA on a four-point scale multiplies the odds by 11.4. The normalized odds ratio uses standardized continuous variables. To standardize or normalize a variable, the mean is subtracted from the variable and the result is divided by the standard deviation. A one-standard deviation increase in CGPA multiplies the odds by 3.47, an increase in odds of 247%. Each model also reports an intercept; this is the base odds if all the independent variables are zero. Most unnormalized intercepts are zero; a

TABLE III. Physics 1: Logistic and random forest model performance. The 95% confidence interval for AUC is shown in brackets.

		Logistic regression				Random forests		
Model	Variables	Accuracy	κ	AUC	R^2	Accuracy	κ	AUC
Baseline	None	0.63	0.00	0.50		0.63	0.00	0.50
Institutional	All	0.71	0.35	0.78 [0.72, 0.83]	0.35	0.73	0.40	0.78 [0.73, 0.83]
	Optimal	0.70	0.32	0.79 [0.74, 0.84]	0.33	0.72	0.40	0.77 [0.71, 0.82]
Week 1	In-Class	0.70	0.26	0.71 [0.65, 0.76]	0.16	0.67	0.20	0.67 [0.61, 0.73]
	In-Class and Institutional	0.73	0.39	0.81 [0.77, 0.86]	0.37	0.72	0.40	0.79 [0.75, 0.84]
Week 2	In-Class	0.71	0.31	0.78 [0.72, 0.83]	0.24	0.71	0.34	0.75 [0.69, 0.80]
	In-Class and Institutional	0.78	0.50	0.84 [0.80, 0.89]	0.41	0.75	0.46	0.81 [0.76, 0.86]
Week 3	In-Class	0.76	0.45	0.81 [0.76, 0.86]	0.30	0.73	0.38	0.80 [0.75, 0.85]
	In-Class and Institutional	0.79	0.53	0.86 [0.81, 0.90]	0.43	0.75	0.44	0.82 [0.78, 0.87]
Week 4	In-Class	0.76	0.46	0.81 [0.77, 0.86]	0.32	0.74	0.39	0.77 [0.72, 0.83]
	In-Class and Institutional	0.78	0.50	0.86 [0.81, 0.90]	0.43	0.75	0.46	0.82 [0.78, 0.87]
Week 5	In-Class	0.82	0.59	0.88 [0.85, 0.92]	0.46	0.78	0.52	0.85 [0.81, 0.89]
	In-Class and Institutional	0.80	0.55	0.88 [0.84, 0.92]	0.51	0.80	0.56	0.86 [0.82, 0.90]

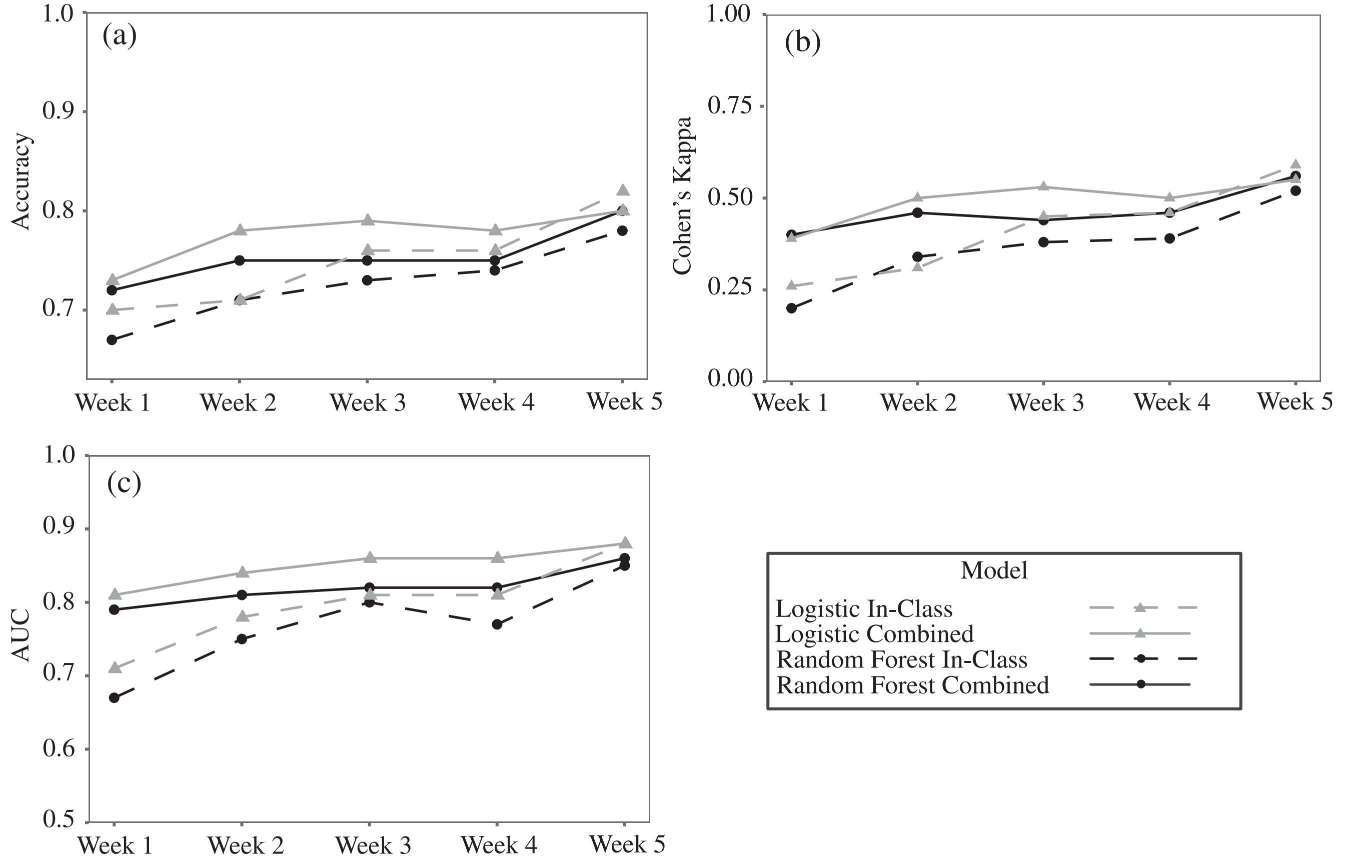

FIG. 1. Physics 1: Model accuracy (a), Cohen's κ (b), AUC (c). The horizontal axis represents the performance of the baseline model in each plot.

student with a zero CGPA or HwkGrade has a very small chance of passing the class. The normalized intercept represents the odds of a student with all dichotomous variables equal to zero and average values on all continuous variables of receiving an A or B in the course.

The optimal institutional logistic model required four variables: CGPA, ACTM, CmpPct, and MathEntry. Higher CGPA, ACTM, and CmpPct increased the odds of a student receiving an A or B in the course. CmpPct is the percentage of credit hours attempted that were completed. Math Entry Point is a 3-level categorical variable with levels: Calculus, Pre-Calculus, and Algebra where Calculus was used as the reference level for the variable. Students who did not arrive at the university ready to take calculus had lower odds of receiving an A or B in Physics 1. CGPA was much more important than either CmpPct or ACTM with an increase in CGPA of 1 standard deviation corresponding to an increased odds of receiving an A or B of 247%. Students starting mathematics in a course below Calculus 1 had significantly lower odds of earning an A or B; the odds for students entering in College Algebra decreased by 270%, 245% for those entering in Pre-Calculus. This has important implications for physics instruction, because all students in Physics 1 had passed Calculus 1, a prerequisite for the course; therefore, weak high school mathematics preparation is not completely remediated by matriculating through college mathematics classes.

TABLE IV. Physics 1: Optimal institutional logistic model.

Variable	Odds ratio	95% CI	Norm odds ratio	Z score	p value
Intercept	0.00	[0.00, 0.00]	4.93	−8.69	<0.001
CmpPct	1.03	[1.02, 1.05]	1.68	3.90	<0.001
CGPA	11.40	[6.62, 19.64]	3.47	8.77	<0.001
ACTM	1.02	[1.01, 1.04]	1.45	2.64	0.008
MathEntry: Algebra	0.27	[0.13, 0.53]	0.27	−3.73	<0.001
MathEntry: PreCal	0.29	[0.16, 0.52]	0.29	−4.15	<0.001

The performance of the optimal random forest institutional model was better than that of the optimal logistic model producing an accuracy of 72% and a kappa of 0.40 representing moderate agreement of the model with the test data. The model discrimination was somewhat worse than that of the logistic model with an AUC of 0.77 representing adequate model performance. The optimal model selected by the 1-SE rule contained the 4 variables in the logistic

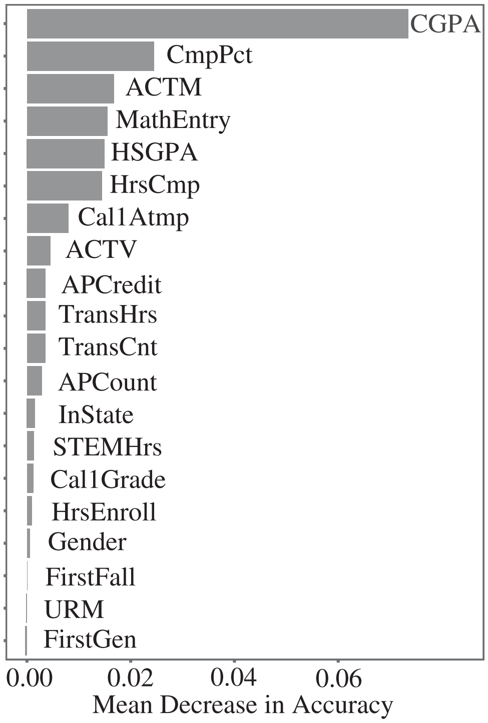

FIG. 2. Physics 1: Variable importance for institutional variables. High values represent more important variables for the goodness of fit of the random forest models.

model: CGPA, CmpPct, ACTM, and MathEntry, as well as a fifth variable, HrsCmp (the total credit hours completed).

Figure 2 plots the mean decrease in accuracy of the institutional random forest model. The mean decrease in accuracy is the average decrease in classification accuracy across all decision trees using the variable if the variable is removed from the tree. Analysis of the importance indices in Fig. 2 showed that despite the optimal model requiring only five variables, several other parameters were important when predicting student success in Physics 1. All five of the optimal variables show high or reasonably high importance in Fig. 2. The largest drop in importance occurs after CGPA where the next most important parameter was only one-third as important to the model fit. Although not included in the optimal model, HSGPA performs very similarly to HrsCmp, MathEntry, and ACTM in variable importance and would likely have been included had CGPA not been included. Demographic factors including first generation status, in-state residence, ethnic or racial minority status, and gender were all minimally important to the model. These results further support the variables selected for the previous logistic model with all 4 of the logistic variables with medium to high importance.

Substantial additional analysis is suggested by Fig. 2. Measures of college success, CGPA and CmpPct are the most important in the model; however, measures of pre-college preparation are also important, HSGPA, ACTM, and MathEntry. The three high school variables are approximately equally important suggesting that it is not only math readiness, but general academic success in high school that is important. Taking college-level preparatory courses in high school either through Advanced Placement (AP) or through transfer credit seems of less importance than general success in high school measured by HSGPA or overall preparation measured by ACTM.

TABLE V. Physics 1: Logistic models in-class variables only.

Variable	Odds ratio	95% CI	Norm odds ratio	Z score	p value
		Week 1			
Intercept	0.01	[0.00, 0.02]	0.87	−6.68	<0.001
FMCEPre	1.04	[1.02, 1.05]	1.88	4.60	<0.001
HwkGrade	1.05	[1.03, 1.07]	2.77	6.24	<0.001
LecQuiz	2.34	[1.37, 3.98]	2.34	3.12	0.002
		Week 2			
Intercept	0.00	[0.00, 0.00]	1.69	−8.05	<0.001
FMCEPre	1.05	[1.03, 1.07]	2.33	5.19	<0.001
HwkGrade	1.08	[1.06, 1.10]	4.95	7.56	<0.001
LecQuiz	3.58	[1.54, 8.34]	1.41	2.95	0.003
		Week 3			
Intercept	0.00	[0.00, 0.00]	0.40	−9.12	<0.001
FMCEPre	1.06	[1.04, 1.08]	2.91	5.64	<0.001
HwkGrade	1.10	[1.07, 1.12]	6.89	8.68	<0.001
LecQuiz	5.47	[1.84, 16.24]	1.46	3.06	0.002
		Week 4			
Intercept	0.00	[0.00, 0.00]	0.28	−9.30	<0.001
FMCEPre	1.06	[1.04, 1.08]	2.84	5.55	<0.001
HwkGrade	1.10	[1.08, 1.13]	7.64	8.67	<0.001
LecQuiz	7.90	[2.26, 27.58]	1.53	3.24	0.001
		Week 5			
Intercept	0.00	[0.00, 0.00]	0.32	−10.72	<0.001
FMCEPre	1.03	[1.01, 1.05]	1.74	2.75	0.006
HwkGrade	1.10	[1.08, 1.13]	7.35	8.13	<0.001
LecQuiz	7.66	[1.76, 33.29]	1.51	2.72	0.007
Test 1	1.10	[1.08, 1.13]	4.64	8.26	<0.001

2. Physics 1 in-class-only models

The weekly models use in-class and institutional variables on a by-week basis to predict the course grade students will receive with data available in each week. In this section, models using only in-class variables are presented. These models use data that is easily accessible for most introductory physics instructors. The logistic models from week 1 through week 5 are presented in Table V. Because of the small number of variables available, the random forest analysis will only be discussed in detail for the combined models.

The variables selected by the in-class logistic models were consistent across all five weeks with FMCEPre, HwkGrade, and LecQuiz significant in all models. The normalized odds ratio of the homework grades increased from 2.77 in week 1 to a high of 7.64 in week 4 demonstrating the increasing predictive importance of homework grades as the semester progressed. The odds ratio for the normalized lecture quiz grade was largest in week 1 when the lecture quiz captured whether the student enrolled on time and whether they promptly obtained a clicker as well as first week attendance. In the subsequent weeks, the normalized coefficient was smaller and fairly constant as the variable transitioned to primarily a measure of attendance.

The significance of HwkGrade and LecQuiz were not surprising. Student success should depend on successful homework completion and consistent class attendance. These results provide support that the analysis method is working and that the class studied is functioning as intended.

3. Physics 1 in-class and institutional models

Models were also constructed using a combination of in-class and institutional variables. The logistic models are presented in Table VI. In weeks 1 to 4, CmpPct, CGPA, and MathEntry were significant institutional variables and FMCEPre, HwkGrade, and LecQuiz were significant in-class variables (LecQuiz was not significance in week 1). With the inclusion of test 1 in week 5, MathEntry was no longer significant. ACTM was only significant in week 2 with a p value of 0.044. The odds ratio of CmpPct remained relatively stable across all models and MathEntry became fairly constant after week 1. The normalized CGPA odds ratio decreased from a high of 3.55 in week 1 to a low of 1.96 in week 5 when test 1 grades became available.

The combined models show that both in-class-only and institutional-only models provide an incomplete picture of student risk and that both sets of variables are important to fully evaluate the chances of success of students in a physics class.

Random forests provided additional insight by estimating variable importance, which characterizes the value of each parameter without sensitivity to multicollinearity. These models combine the optimal institutional model with the optimal in-class model without further pruning. Figure 3 shows that CGPA was the most important variable in weeks 1 to 4 by a factor of at least 2 when compared to every other variable except homework grade. Test 1 became the most important variable in week 5. The random forest models routinely performed more weakly than the logistic models by a small margin.

In all weeks, the in-class variables FMCEPre and LecQuiz were of little importance to the accuracy of the random forest models. This does not agree with the parameter estimates of the logistic models in Table VI where FMCEPre often had a normalized odds ratio commensurate with HwkGrade. This suggests the collinearity expected between the grades of different assignments in the same class may be influencing the parameter estimates of grades in the logistic models. This shows the efficacy of using both logistic and random forest models in parallel.

TABLE VI. Physics 1: Logistic models combining in-class and institutional variables.

Variable	Odds ratio	95% CI	Norm odds ratio	Z score	p value
		Week 1			
Intercept	0.00	[.00, .00]	5.32	−9.34	<0.001
FMCEPre	1.04	[1.02, 1.06]	1.94	3.89	<0.001
HwkGrade	1.03	[1.02, 1.05]	1.96	3.95	<0.001
CmpPct	1.03	[1.02, 1.05]	1.66	3.68	<0.001
CGPA	11.91	[6.62, 21.44]	3.55	8.26	<0.001
MathEntry: Algebra	0.23	[0.13, 0.43]	0.23	−4.60	<0.001
MathEntry: PreCal	0.29	[0.16, 0.53]	0.29	−4.07	<0.001
		Week 2			
Intercept	0.00	[0.00, 0.00]	1.47	−9.32	<0.001
FMCEPre	1.05	[1.02, 1.07]	2.24	4.28	<0.001
HwkGrade	1.05	[1.03, 1.08]	2.87	4.74	<0.001
LecQuiz	3.03	[1.13, 8.15]	1.35	2.20	0.028
CmpPct	1.03	[1.01, 1.05]	1.65	3.63	<0.001
CGPA	8.70	[4.73, 16.01]	3.02	6.95	<0.001
ACTM	1.02	[1.00, 1.04]	1.36	2.01	0.044
MathEntry: Algebra	0.42	[0.19, 0.89]	0.42	−2.25	0.024
MathEntry: PreCal	0.43	[0.23, 0.81]	0.43	−2.62	<.009
		Week 3			
Intercept	0.00	[0.00, 0.00]	1.16	−9.81	<0.001
FMCEPre	1.06	[1.03, 1.08]	2.72	4.89	<0.001
HwkGrade	1.07	[1.04, 1.09]	3.79	5.74	<0.001
LecQuiz	4.35	[1.26, 15.00]	1.39	2.32	0.020
CmpPct	1.03	[1.01, 1.05]	1.63	3.57	<0.001
CGPA	7.65	[4.11, 14.21]	2.83	6.43	<0.001
MathEntry: Algebra	0.33	[0.17, 0.63]	0.33	−3.35	<0.001
MathEntry: PreCal	0.43	[0.23, 0.81]	0.43	−2.63	0.008
		Week 4			
Intercept	0.00	[0.00, 0.00]	0.77	−9.82	<0.001
FMCEPre	1.05	[1.03, 1.08]	2.59	4.72	<0.001
HwkGrade	1.07	[1.05, 1.10]	3.98	5.67	<0.001
LecQuiz	6.49	[1.59, 26.50]	1.47	2.61	0.009
CmpPct	1.03	[1.01, 1.05]	1.63	3.56	<0.001
CGPA	7.09	[3.79, 13.25]	2.72	6.14	<0.001
MathEntry: Algebra	0.33	[0.17, 0.64]	0.33	−3.29	0.001
MathEntry: PreCal	0.45	[0.24, 0.84]	0.45	−2.52	0.012
		Week 5			
Intercept	0.00	[0.00, 0.00]	0.40	−10.69	<0.001
FMCEPre	1.04	[1.02, 1.06]	2.00	3.23	0.001
HwkGrade	1.09	[1.06, 1.11]	5.54	6.69	<0.001
LecQuiz	7.33	[1.57, 34.26]	1.49	2.53	0.011
CmpPct	1.03	[1.01, 1.05]	1.57	3.29	0.001
CGPA	3.72	[1.91, 7.24]	1.96	3.87	<0.001
Test 1	1.09	[1.07, 1.12]	4.10	7.12	<0.001

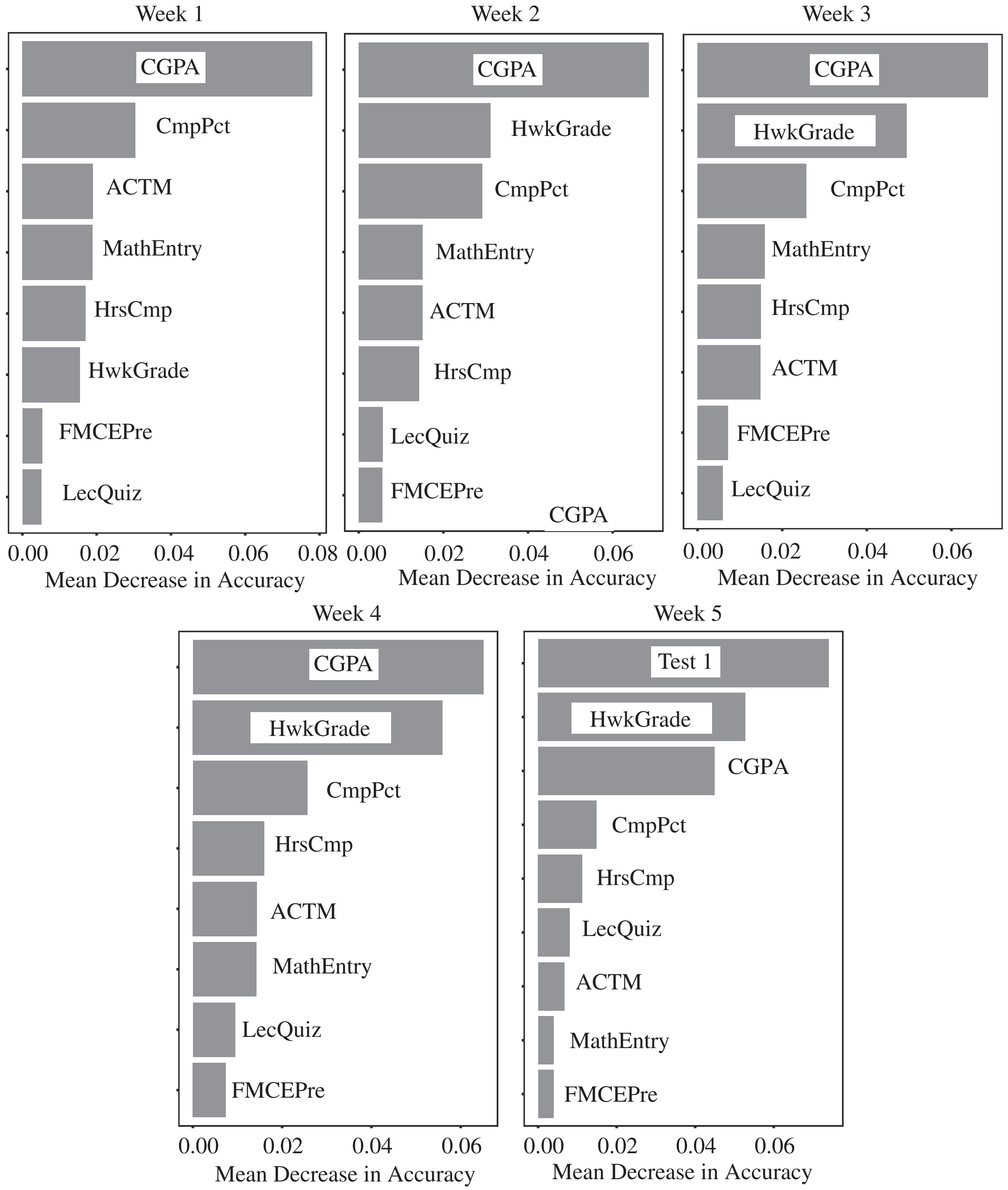

FIG. 3. Physics 1: Weekly variable importance of the combined in-class and institutional models.

C. Physics 2

Physics 2 is usually taken the semester after Physics 1 and has Physics 1 as its prerequisite. As such, the Physics 2 models contain additional important variables not available to the Physics 1 models: the grade in Physics 1 (P1Grade) and whether Physics 1 was taken more than once (P1Atmp). Physics 2 also maintained more detailed weekly records allowing the use of lab grades and lab quiz grades.

1. Institutional model

The accuracy, kappa, and AUC of each Physics 2 model is presented in the Table VII. The optimal logistic model using only institutional variables produced an accuracy of 80%, an improvement of 12% over the baseline model and 10% over the performance of the same model in Physics 1. This model also had kappa of 0.55 and AUC of 0.89, values that were not obtained until the week 5 models in Physics 1.

TABLE VII. Physics 2: Model accuracy, κ, and AUC.

		Logistic regression				Random forests		
Models	Variables	Accuracy	κ	AUC	R^2	Accuracy	κ	AUC
Baseline	None	0.68	0.00	0.50		0.68	0.00	0.50
Institutional	All Variables	0.81	0.57	0.89 [0.86, 0.93]	0.34	0.80	0.55	0.88 [0.84, 0.92]
	Optimal	0.80	0.55	0.89 [0.85, 0.93]	0.31	0.81	0.58	0.86 [0.82, 0.90]
Week 1	In-Class	0.73	0.33	0.75 [0.69, 0.81]	0.12	0.74	0.36	0.74 [0.68, 0.80]
	In-Class and Institutional	0.81	0.57	0.90 [0.87, 0.94]	0.35	0.81	0.57	0.88 [0.84, 0.92]
Week 2	In-Class	0.75	0.40	0.78 [0.73, 0.84]	0.21	0.75	0.42	0.75 [0.70, 0.81]
	In-Class and Institutional	0.81	0.57	0.91 [0.87, 0.94]	0.38	0.82	0.61	0.88 [0.84, 0.92]
Week 3	In-Class	0.78	0.49	0.85 [0.80, 0.89]	0.30	0.79	0.52	0.87 [0.82, 0.91]
	In-Class and Institutional	0.82	0.59	0.91 [0.88, 0.94]	0.42	0.84	0.62	0.91 [0.87, 0.94]
Week 4	In-Class	0.78	0.48	0.85 [0.81, 0.90]	0.31	0.75	0.43	0.83 [0.78, 0.88]
	In-Class and Institutional	0.84	0.62	0.91 [0.88, 0.95]	0.44	0.81	0.57	0.89 [0.85, 0.92]
Week 5	In-Class	0.85	0.65	0.94 [0.91, 0.97]	0.56	0.85	0.66	0.93 [0.90, 0.96]
	In-Class and Institutional	0.85	0.67	0.95 [0.93, 0.97]	0.60	0.86	0.68	0.95 [0.92, 0.97]

The optimal logistic model of the institutional data for Physics 2 was strongly related to performance in Physics 1 as shown in Table VIII. The normalized odds ratios for receiving an A/B in Physics 2 improved by 169% if the student received an A or B in Physics 1. CGPA remained an influential variable with a one standard deviation improvement in CGPA improving the odds of receiving an A or B in Physics 2 by 243%. Taking Physics 1 more than a single time decreased a student's odds of receiving an A or B by 426%. Physics 1 is the direct prerequisite for Physics 2 and, therefore, this result was not surprising; however, Calculus 1 is a direct prerequisite for Physics 1, but was not significant in any of the Physics 1 models. Cal1Atmp (whether Calculus 1 was taken more than once) was significant in the Physics 2 random forest models.

The institutional random forest models for Physics 2 also performed very well with the optimal institutional model achieving an accuracy of 81%, kappa of 0.58, and AUC of 0.86. This high degree of predictive power was achieved using CGPA, P1Grade, Cal1Atmp, HrsCmp, and CmpPct. The variables CGPA, HrsCmp, and CmpPct were also included in the corresponding random forest model in Physics 1. The variable importance indices in Fig. 4 of the full model show similar results to Physics 1 with CGPA the highest importance variable with a mean decrease in accuracy of more than twice that of the next most important variable. The demographic variables were once again unimportant for predicting grade. Only a small subset of institutional variables were necessary to build highly performing models of student course success in the random forest model.

TABLE VIII. Physics 2: Optimal institutional logistic model.

Variable	Odds ratio	95% CI	Norm odds ratio	Z score	p
Intercept	0.00	[0.00, 0.01]	1.76	−7.53	<0.001
P1Grade	2.69	[1.56, 4.63]	2.69	3.57	<0.001
P1Atmp	0.19	[0.07, 0.54]	0.19	−3.12	0.002
CGPA	11.66	[6.01, 22.59]	3.43	7.27	<0.001

2. Weekly models

The detailed weekly logistic models and the weekly random forest variable importance estimates are presented

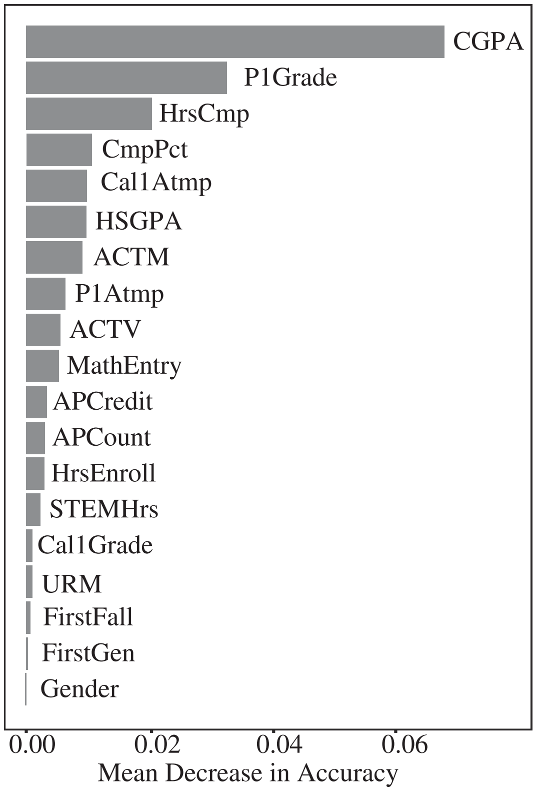

FIG. 4. Physics 2 variable importance for the institutional variables.

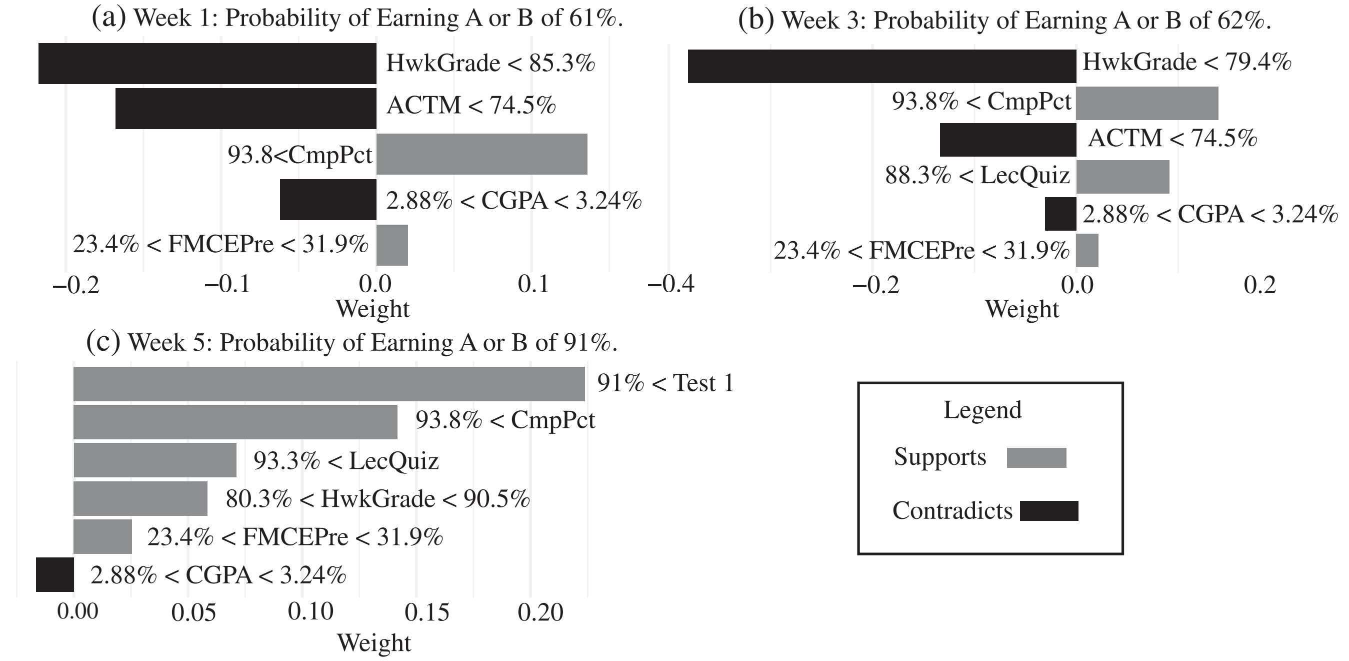

FIG. 5. Physics 1 LIME results for one student for week 1, week 3, and week 5. The student received a B in the class.

in the Supplemental Material [43]. The performance of the optimal logistic model of the institutional data was not matched by the in-class-only model until week 5 and test 1, much later than the Physics 1 logistic models. All of the logistic models were highly accurate with all combined institutional and in-class models having accuracies in excess of 80%, kappas in the range of good agreement, and excellent discrimination characteristics as measured by AUC. In each case, the Physics 2 models performed better than similar models in Physics 1. This may have resulted from the additional in-class variables available, the availability of P1Grade as a variable, the increased difficulty and unfamiliarity of the material, or from hand-graded homework providing a more accurate measure of students' understanding.

The in-class logistic models were remarkably consistent from week 1 to week 5 with only LabQuiz and HwkGrade significant in the first 4 weeks. The week 5 model retained both parameters along with test 1 grade. Unlike the behavior of LecQuiz in Physics 1, there was no large difference in the LabQuiz normalized odds ratio between week 1 and week 2. The normalized odds ratio for HwkGrade increased from 2.13 in week 1 to 4.40 in week 5 showing the increasing importance of homework grades for predicting student success in this course.

The combined institutional and in-class models also performed similarly to those of Physics 1 with all variables except CGPA and LabQuiz consistently significant over the five weeks. As the course accumulated grade data, the normalized odds ratio of CGPA dropped from 2.89 in week 1 to 2.17 in week 4; in week 5, when the test 1 grade was available, CGPA was no longer significant. LabQuiz was only significant in weeks 3 and 4.

The random forest results in Table VII show that the random forest models performed similarly to the logistic models. The in-class model was not an improvement over the institutional model until week 5.

Variable importance in Physics 2 was similar the variable importance in Physics 1 with CGPA the most important variable and homework grade increasing in importance throughout the five week period to match CGPA in week 5. Unlike Physics 1, when test 1 was included in week 5, it was nearly twice as important as either CGPA or HwkGrade. Physics 2 logistic models and importance plots can be found in the Supplemental Material [43].

D. LIME results

Once a classification model is constructed, the variable importance and the odds ratios can help instructors and researchers understand the factors that are important for predicting students' success. The classification model can then be used to identify students at risk of receiving a low grade. The LIME algorithm allows further understanding of the classification model, showing the factors that were important to making the prediction for individual students.

Figure 5 shows the LIME analysis of the progression of the optimal logistic model for a single Physics 1 student in weeks 1, 3, and 5. The student ultimately earned a B in the class. In the first week of the class, the model predicted that the student would receive an A or B with 61% probability. Poor performance on the first in-class homework assignment, a relatively weak ACTM score, and a weak CGPA were the largest contributors to lowering the probability, shown in Fig. 5(a). By week 3, the probability of earning an A or B was 62%. The student's low homework grade was the strongest factor lowering this probability. A strong record of previous course completion, strong lecture quiz scores, and a fair (for the class) score on the FMCE pretest supported the prediction of earning an A or B. By week 5, the probability increased dramatically due to a very strong performance on the first examination. Despite a number of variables indicating that this student would not do well in the course prior to week 5, the model accurately predicted that the student would receive a B or better in the first week of class.

The LIME analysis opens a number of exciting instructional possibilities. While tailored to individual students, examination of the LIME results for the set of at-risk students can allow an instructor to develop a greater understanding of the risk factors as well as the critical thresholds for those factors. For example, for students with moderate CGPA, what homework average represents the threshold for increased risk of failure. The LIME analysis could also allow instructors to provide more detailed advice to at risk students; it is possible this process could be automated to provide feedback to all students.

IV. DISCUSSION

This study investigated two research questions; they will be discussed in the order proposed.

RQ1: How well can introductory physics course grades be predicted early in the semester? Highly accurate models of student success were constructed in both Physics 1 and Physics 2 using institutional variables, in-class variables, and a combination of in-class and institutional variables. In both classes, the models with only institutional variables performed well with strong fit statistics. These models outperformed similar models reported by Huang and Fang [31] that had a maximum classification prediction accuracy of 67% compared to the 70% to 80% for our logistic models. The optimal logistic institutional-only models outperformed, as measured by kappa, models using only in-class variables early in the class. In Physics 1, three weeks of in-class data were required before in-class-only models matched the performance of the institutional-only models. In Physics 2, in-class-only models did not match the performance of institutional-only models until week 5 when the score on the first in-semester examination became available. In Physics 1, models combining institutional and in-class variables had superior performance, measured by kappa, in week 2 to in-class-only models in week 4; thus allowing the accurate identification of at-risk students much earlier in the semester. In Physics 2, the week 1 combined institutional and in-class model had superior properties to the week 4 in-class-only model, allowing very early identification of at-risk students. This early identification may allow interventions to be directed toward students in time to positively affect their first test grade.

The in-class models presented in this study performed similarly to those developed by Macfadyen and Dawson [33], but were around 10% to 15% less accurate than the in-class models developed by Marbouti *et al.* [32]. There are many potential reasons for the lower accuracy of our models: the current study used grades less than a B as a classification criteria while Marbouti *et al.* classified grades less than a C. The course studied by Marbouti *et al.* was an introductory engineering course which had students with a fairly homogeneous preparation and may have weighted participation more strongly than mastery of content. Physics 1 and 2, as well as the introductory biology course studied by Macfadyen and Dawson [33], are courses taken by a variety of majors at different points in their academic careers.

RQ2: What variables are most important for the accurate prediction of physics course grades early in the semester? In both courses, the institutional models were constructed using only a small subset of available parameters: CGPA, CmpPct, ACTM, and MathEntry in Physics 1 and CGPA, P1Grade, and P1Atmp in Physics 2. As such, only a limited amount of additional information would need to be provided to an instructor to greatly increase the accuracy of the identification of at-risk students. Additional institutional variables are important predictors of student success; however, they were unnecessary for the identification of at-risk students in the introductory physics course in this study.

The differences in variable selection between the two courses was most likely the result of the availability of Physics 1 course information for the Physics 2 models. Physics 1 performance serves as a strong indicator of student performance in a Physics 2. The high importance of CGPA is consistent with other research into grade prediction [31,49]. As the direct prerequisite for Physics 2, Physics 1 grade was expected to be the strongest predictor; however, its variable importance was less than CGPA showing that overall academic performance at the college level was more important than prior physics performance. All Physics 2 students in the sample had a grade for Physics 1. The grade in the primary prerequisite course for the Physics 1, Calculus 1, was not as important as was expected from prior research [31,49]. Whether Calculus 1 was attempted more than once was not important to the Physics 1 models and was only significant in the Physics 2 random forest models. The grade in Calculus 1 was not significant in any model. In Physics 1, the student's math entry point was an important variable in both the institutional-only and the combined models. All students in Physics 1 had completed Calculus 1; therefore, completing a college mathematics class did not completely remove the effects of weak high school mathematics preparation. This was consistent with previous work showing prior mathematics performance was predictive of current course performance [31,50]. The observation that Calculus 1 was important in Physics 2 models, but not in Physics 1 models may suggest that Calculus 1 might be more appropriate as a co-requisite to Physics 1 than as a prerequisite.

For the in-class models for both courses, HwkGrade was the most important variable until test 1 grade became available. In the courses studied, the student's homework grade was the variable most directly related to mastery of the topics tested by the in-semester examinations. The importance of this variable increased as the class progressed. This may have been a result of the increasing weekly difficulty of assignments early in a physics class providing a better measure of student capability. It could also result from HwkGrade measuring whether the student self-regulates to improve homework scores early in the class.

The conceptual pretest was of lower importance than homework grade in the Physics 1 models and was not significant in Physics 2 models. This may have resulted from the low pretest scores in both classes making the pretest scores less predictive of success. Henderson *et al.* demonstrated that CSEM pretest scores were less strongly related to conceptual preparation for women than for men; they attributed this difference to the lower scores of women shifting their score distribution nearer to the pure guessing distribution [51]. A similar effect may lessen the predictive power of the pretest in the classes studied.

Gender, race or ethnicity, and first generation college-goer status were not important variables to predicting students' physics course success. Prior work has suggested that race or ethnicity and first generation status are strongly mediated by academic preparation variables such as ACT or SAT scores [52,53]. While there is substantial literature suggesting gender is an important variable for the predication of physics post-test scores [54], there is an equally significant literature suggesting that, in general, women earn higher course grades than men [55]. Neither effect was important to the prediction of physics grades when general college-level control variables, such as CGPA, were available.

A. Additional observations

Beyond the possibility of constructing accurate classification models, the logistic regression results, the variable importance results, and the LIME results provide additional insight into student outcomes. Variable importance measures show where an instructor should look to identify at-risk students and show that different measures are important at different times in a physics class. This also allows the identification of grades that are collected which do not predict student course outcomes; it may be efficacious to revisit these assignments to determine why grades on the assignment are not related to overall success in the course. Logistic regression odds ratios quantify the size of the effects of different variables. The LIME algorithm, Fig. 5, provides a detailed, by-student, characterization of how the models made their decision when predicting student outcomes and provides insights into where interventions could have the greatest impact. It is important to understand that the LIME results are particular to each case and should not be generalized to all students; however, examining LIME results for the cohort of students identified as at risk can provide substantial additional insight.

The conceptual pretest scores had very low variable importance in any model that included institutional variables, particularly CGPA. While this may be the result of the low pretest scores as discussed above, it may also result from the pretest providing primarily a measure of general academic ability rather than specific physics prior preparation. This would explain why pretest scores were unimportant in models containing superior measures of general academic ability such as CGPA.

B. Recommendations

The methods explored in this paper allow any instructor to develop risk models for introductory physics classes. These models could be used to target interventions to at-risk students, possibly strongly improving STEM retention. The models are more accurate earlier in the semester if a few institutional variables are available, particularly CGPA. Departments and institutions should develop practices to make these variables easily available to instructors.

The LIME models provide a detailed student-level picture of probable class outcomes and a time-resolved evolution of the outcome along with the factors influencing the prediction. In the future, further studies may investigate whether it is productive to make these models available to the students.

Once at-risk students have been identified, a broad set of possible interventions become available. These can be as simple as communicating to the student that they are at risk of not successfully completing the class (with appropriate messaging) [33,56,57]. This could be done electronically through an indicator in the LMS or personally through a message from the course instructor or from the student's laboratory teaching assistant (TA). Beyond an expression of concern, the message could suggest connecting with existing campus resources such as office hours or supplemental instruction. By examining the student's record to date in the class (perhaps informed by the LIME algorithm), behaviors that enhance risk of failure can be identified and communicated to the student. If the class uses interactive instruction, learning or teaching assistants can be directed to ensure the student is involved. This might involve reshuffling laboratory or recitation groups. If the source of the student's risk is failure to complete assignments, the student could be encouraged to turn in a late assignments for reduced credit. If the source of risk is failure to attend class, the student's absences could be excused under the condition of improved future attendance. Instructors may consider forming optional components of the course open to all students such as study groups overseen by a TA. These additional components could implement alternate pedagogical methods addressing the needs of students who were not prospering in the instructional environment provided to the majority of students. At-risk students could be directed to these resources.

C. Limitations

Even with both in-class and institutional data, the models were not completely accurate. Analysis of the confusion matrices in Supplemental Material [43] shows a bias toward false positives in many of the models. The decision threshold should be determined on a case-by-case basis depending on instructor comfort with misclassification. These models may not be entirely generalizable as they depend on the local educational environment at the institution studied. Because of this, the techniques for constructing the models should be adopted instead of the individual models when extending to different environments.

D. Future work

This work provided an introduction to the use of classification algorithms to predict student success. A significant amount of additional research is needed to fully understand the application of these algorithms to physics classes. The classification algorithms require the researcher to make a number of choices such as the relative size of the test and training datasets, the number of decision trees grown to form the random forest, and the minimum required sample size. Optimal criteria for these choices should be determined. This study used the default parameter choices for the machine learning algorithms; different values of these parameters should be explored to determine the optimal values for prediction of physics grades. This work chose to predict whether a student would achieve an A or B; this choice was made partially to have an outcome variable the was somewhat balanced between two outcomes. Many interesting outcome variables are substantially unbalanced, for example whether a student receives a "D" or "F" in the classes studied; the behavior of classification algorithms for very unbalanced outcomes should be investigated.

This study found that demographic variables were not important predictors of success. This was true in both the logistic and random forest models and, therefore, seems a robust result. The samples for each minority demographic group were extremely unbalanced with members of the group in the minority forming less than 20% of the sample. Additional research is needed to ensure this imbalance did not bias the results of the classification algorithms.

V. CONCLUSIONS

Machine learning techniques produced accurate predictions of student course outcomes. Models using institutional data were more accurate before the class began than models using only in-class variables were in the second week of class in Physics 1, the fourth week of class in Physics 2. The accuracy of the models varied between the two courses studied; however, many of the same variables were common to the optimal models for the two courses. By the first examination in both courses, the institutional variables no longer improved the models. The importance of CGPA as a measure of preparedness was reinforced in this study with none of the additional institutional variables near the level importance of this single measure. Using the LIME algorithm, it was possible to show which factors were most important for each student in predicting his or her most likely course outcome. The combination of these models and the LIME algorithm produced a highly accurate and detailed classification of student risk that can be used to target interventions at the student level in introductory physics.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant No. ECR-1561517.

[1] National Science Board, *Revisiting the STEM Workforce: A Companion to Science and Engineering Indicators 2014* (National Science Foundation, Arlington, VA, 2015).

[2] President's Council of Advisors on Science and Technology, *Report to the President. Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics* (Executive Office of the President, Washington, DC, 2012).

[3] X. Chen, *STEM Attrition: College Students' Paths into and out of STEM Fields. NCES 2014-001* (National Center for Education Statistics, Washington, DC, 2013).

[4] K. Rask, Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences, Econ. Educ. Rev. **29**, 892 (2010).

[5] E. J. Shaw and S. Barbuti, Patterns of persistence in intended college major with a focus on STEM majors, NACADA J. **30**, 19 (2010).

[6] A. V. Maltese and R. H. Tai, Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among US students, Sci. Educ. **95**, 877 (2011).

[7] G. Zhang, T. J. Anderson, M. W. Ohland, and B. R. Thorndyke, Identifying factors influencing engineering

student graduation: A longitudinal and cross-institutional study, J. Eng. Educ. **93**, 313 (2004).

[8] B. F. French, J. C. Immekus, and W. C. Oakes, An examination of indicators of engineering students' success and persistence, J. Eng. Educ. **94**, 419 (2005).

[9] R. M. Marra, K. A. Rodgers, D. Shen, and B. Bogue, Leaving engineering: A multi-year single institution study, J. Eng. Educ. **101**, 6 (2012).

[10] C. W. Hall, P. J. Kauffmann, K. L. Wuensch, W. E. Swart, K. A. DeUrquidi, O. H. Griffin, and C. S. Duncan, Aptitude and personality traits in retention of engineering students, J. Eng. Educ. **104**, 167 (2015).

[11] P. Baepler and C. J. Murdoch, Academic analytics and data mining in higher education, Int. J. Scholarsh. Teach. Learn. **4**, 17 (2010).

[12] R. S. J. D. Baker and K. Yacef, The state of educational data mining in 2009: A review and future visions, J. Educ. Data Mining **1**, 3 (2009); https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8.

[13] Z. Papamitsiou and A. A. Economides, Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. J. Educ. Tech. Soc. **17**, 49 (2014); https://www.jstor.org/stable/jeductechsoci.17.4.49.

[14] A. Dutt, M. A. Ismail, and T. Herawan, A systematic review on educational data mining, IEEE Access **5**, 15991 (2017).

[15] C. Romero and S. Ventura, Educational data mining: A review of the state of the art, IEEE T. Syst. Man Cy. C **40**, 601 (2010).

[16] J. M. Aiken, R. Henderson, and M. D. Caballero, Modeling student pathways in a physics bachelor's degree program, Phys. Rev. Phys. Educ. Res. **15**, 010128 (2019).

[17] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. Pat. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).

[18] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).

[19] R. J. Beichner, J. M. Saul, D. S. Abbott, J. J. Morse, D. Deardorff, R. J. Allain, S. W. Bonham, M. H. Dancy, and J. S. Risley, The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project, in Research-based Reform of University Physics, Vol. 1, edited by E. F. Redish and P. J. Cooney (American Association of Physics Teachers, College Park, MD, 2007), pp. 2–39.

[20] N. D. Finkelstein and S. J. Pollock, Replicating and understanding successful innovations: Implementing tutorials in Introductory Physics, Phys. Rev. ST Phys. Educ. Res. **1**, 010101 (2005).

[21] E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez, and P. Pamelá, Toward equity through participation in modeling instruction in introductory university physics, Phys. Rev. ST Phys. Educ. Res. **6**, 010106 (2010).

[22] C. H. Crouch and E. Mazur, Peer instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).

[23] K. Cummings, J. Marx, R. Thornton, and D. Kuhl, Evaluating innovation in studio physics, Am. J. Phys. **67**, S38 (1999).

[24] B. Thacker, H. Dulli, D. Pattillo, and K. West, Lessons from a large-scale assessment: Results from conceptual inventories, Phys. Rev. ST Phys. Educ. Res. **10**, 020104 (2014).

[25] A. Peña-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works, Expert Syst. Appl. **41**, 1432 (2014).

[26] U. bin Mat, N. Buniyamin, P. M. Arsad, and R. Kassim, An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention, in *Engineering Education (ICEED), 2013 IEEE 5th Conference* (IEEE, Bellingham, WA, 2013), pp. 126–130.

[27] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, Data mining algorithms to classify students, in *Proceedings of the 1st International Conference on Educational Data Mining*, edited by R. S. Joazeiro de Baker, T. Barnes, and J. E. Beck (Montreal, Quebec, Canada, 2008).

[28] A. M. Shahiri, W. Husain, and N. A. Rashid, A review on predicting student's performance using data mining techniques, Procedia Comput. Sci. **72**, 414 (2015).

[29] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R* (Springer-Verlag, New York, NY, 2017), Vol. 112.

[30] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists* (O'Reilly Media, Boston, MA, 2016).

[31] S. Huang and N. Fang, Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models, Comput. Educ. **61**, 133 (2013).

[32] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, Models for early prediction of at-risk students in a course using standards-based grading, Comput. Educ. **103**, 1 (2016).

[33] L. P. Macfadyen and S. Dawson, Mining LMS data to develop an early warning system for educators: A proof of concept, Comput. Educ. **54**, 588 (2010).

[34] U.S. News & World Report: Education, U.S. News and World Report, Washington, DC, https://premium.usnews.com/best-colleges. Accessed 2/23/2019.

[35] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).

[36] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[37] D. P. Maloney, T. L. O'Kuma, C. Hieggelke, and A. Van Huevelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

[38] D. Conway and J. White, *Machine Learning for Hackers* (O'Reilly Media, Boston, MA, 2012).

[39] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. **27**, 861 (2006).

[40] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, New York, NY, 1977).

[41] D. G. Altman, *Practical Statistics for Medical Research* (CRC Press, Boca Raton, FL, 1990).

[42] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression* (John Wiley & Sons, New York, NY, 2013), Vol. 398.

[43] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.020120 for complete Physics 2 analysis, examples of the ROC curves, and a complete list of confusion matrices.

[44] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S* 4th ed. (Springer-Verlag, New York, NY, 2002).

[45] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees* (Wadsworth & Brooks/Cole, Monterey, CA, 1984).

[46] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer-Verlag, New York, NY, 2009).

[47] A. Liaw and M. Wiener, Classification and regression by random forest, R News **2**, 18 (2002); https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf.

[48] M. T. Ribeiro, S. Singh, and C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2016), pp. 1135–1144.

[49] A. Wojciechowski and L. B. Palmer, Individual student characteristics: Can any be predictors of success in online classes, Online J. Distance Learn. Adm. **8**, 13 (2005); https://www.westga.edu/~distance/ojdla/summer82/wojciechowski82.htm.

[50] C. L. Ballard and M. F. Johnson, Basic math skills and performance in an introductory economics class, J. Econ. Educ. **35**, 3 (2004).

[51] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).

[52] R. Henderson and J. Stewart, Racial and ethnic bias in the Force Concept Inventory, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH*, edited by L. Ding, A. Traxler, and Y. Cao (AIP, New York, 2017), pp. 172–175.

[53] R. Henderson, C. Zabriskie, and J. Stewart, Rural and first generation performance differences on the Force and Motion Conceptual Evaluation, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (AIP, New York, 2018).

[54] A. Madsen, S. B. McKagan, and E. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).

[55] D. Voyer and S. D. Voyer, Gender differences in scholastic achievement: A meta-analysis., Psychol. Bull. **140**, 1174 (2014).

[56] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, Academic analytics: A new tool for a new era, EDUCAUSE Rev. **42**, 40 (2007); https://er.educause.edu/articles/2007/7/academic-analytics-a-new-tool-for-a-new-era.

[57] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauria, J. R. Regan, and J. D. Baron, Early alert of academically at-risk students: An Open Source Analytics Initiative, J. Learn. Analytics **1**, 6 (2014).