Modeling student pathways in a physics bachelor's degree program

John M. Aiken,[1,2] Rachel Henderson,[2] and Marcos D. Caballero[1,2,3]

[1]*Center for Computing in Science Education & Department of Physics, University of Oslo, N-0316 Oslo, Norway*

[2]*Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA*

[3]*CREATE for STEM Institute, Michigan State University, East Lansing, Michigan 48824, USA*

(Received 26 October 2018; published 15 May 2019)

Physics education research (PER) has used quantitative modeling techniques to explore learning, affect, and other aspects of physics education. However, these studies have rarely examined the predictive output of the models, instead focusing on the inferences or causal relationships observed in various data sets. This research introduces a modern predictive modeling approach to the PER community using transcript data for students declaring physics majors at Michigan State University. Using a machine learning model, this analysis demonstrates that students who switch from a physics degree program to an engineering degree program do not take the third semester course in thermodynamics and modern physics, and may take engineering courses while registered as a physics major. Performance in introductory physics and calculus courses, measured by grade as well as a students' declared gender and ethnicity play a much smaller role relative to the other features included in the model. These results are used to compare traditional statistical analysis to a more modern modeling approach.

DOI: 10.1103/PhysRevPhysEducRes.15.010128

I. INTRODUCTION

Physics has long built data driven models to explain and to understand systems of study and physics education research (PER) without exception. In PER, these models have been used to explore learning outcomes [1], understand career choice motivations [2], and explore the use of different instructional strategies [3], amongst many other topics. Generally within the PER literature, models are typically within the family of linear models (e.g., ordinary least squares, logistic regression) and are often evaluated by their ability to explain results through odds ratios and p values as well as by using goodness-of-fit tests [4]. These types of methods are adequate when the goal is to explain and/or describe the data collected; however, it becomes difficult to reproduce such results when extending beyond the local setting [5]. Assessing the predictive output of these models can be one way that PER can begin to produce data driven models that can be compared and tested across different settings [5].

As the field of PER continues to develop and refine its approaches to quantitative modeling, it is important to discuss the nature of modeling as well as the limitations and affordances of different approaches. Much of the work in PER assesses model fit without assessing the model's predictive power (see, e.g., Refs. [1–3]). These more traditional approaches tend to evaluate the fit of those models in the context of the collected data. In the current work, we employ a machine learning approach that emphasizes the generalizable nature of a quantitative model by first fitting the model to a data set and then separately evaluating the quality of the model using sequestered data, or "hold out" data, for which the model was not fit. This approach of assessing the predictive output of a model provides a direct quantifiable "performance measure" that can be used to compare multiple models or models from different data sets [6,7]. This approach can allow model predictions to be compared across different settings where the models are attempting to predict the same outcome.

Rather than discussing the employed machine learning technique and abstractly comparing it to more common modeling approaches, here we will demonstrate the affordances and limitations of our approach through a specific case of investigating students who stay in the physics program compared to students who leave physics for engineering. With the aim of introducing a new analysis method into the PER literature, we will analyze the predictions from models that are fit to university registrar data. Namely, we will examine the factors that impact a student's choice to switch from a physics to engineering degree program (intra-STEM switchers). In this study, the registrar data from Michigan State University (MSU) was previously investigated in Aiken and Caballero [8]; these types of data are typically collected by many institutions. Through this

paper, we will introduce an algorithm for classification (random forest [9]) and compare this modeling approach to a summary statistic approach of using contingency tables and effect sizes. While the overarching focus of this paper is to compare these two approaches in a given research context, we do find that our work challenges the previous results stated in Aiken and Caballero [8]. We will demonstrate that the single most important component for remaining in the physics major is taking the third semester modern physics course.

For faculty who have spent any time advising physics majors, it might seem like this result is intuitive and it is true that this result served as an assumption in the previous work that analyzed student pathways in physics [10]. It is precisely this intuition that allows us to focus on a new approach of modeling university registrar data and discuss the affordances and limitations of such an approach. Furthermore, although this result may be intuitive, it has gone unreported and thus, this work can provide the basis for future pathway studies.

As we use registrar data to illustrate what we might learn from this modeling approach, it is worth noting that this approach opens additional research questions that are worth reporting. Much of the literature that examines the pathways of science, technology, engineering, and math (STEM) students focuses on students who leave STEM completely [11,12] and not on those who leave one STEM discipline for another. Because at least one-third of the students registering as physics majors at MSU earn degrees in engineering [8], through this work, we can investigate the following three questions related to whether students stay in a physics bachelor's program or leave for an engineering degree:

(1) Factors that describe students who leave STEM are well documented. Some of those factors are present in university registrar data either directly or through various proxies. Which of the factors identified in the literature impact students to remain a physics major or leave the physics bachelor's program for an engineering degree?
(2) In prior work, Aiken and Caballero [8] found that performance metrics between physics and engineering graduates differed. What were the effect sizes of these performance factors and how do they impact models that compare the effect of various factors against each other?
(3) As our goal is to understand intra-STEM switchers, what did we learn about students who register for physics but leave for engineering?

This Letter is organized as follows. In Sec. II, we present prior work on STEM pathways to document specific features that we might be able to find in the MSU registrar data (Sec. III). We discuss the traditional and machine learning approaches in Sec. IV and refer the reader to Young *et al.* [13] for additional details on the random forest technique. We then present the models tested (Sec. V) as well as the resulting output and validation (Sec. VI). We then discuss the findings, in the context of our data to demonstrate what might be gained from using this modern approach (Sec. VII) as well as the specific limitations of our work (Sec. VIII). Finally, we critique the different modeling approaches, discussing what affordances and limitations we find more generally (Sec. IX) and offer some concluding remarks (Sec. X).

II. BACKGROUND

The previous two decades have seen a large increase in student enrollment in STEM including physics bachelor's degree programs [14]. These large changes have been driven by a variety of efforts both nationally and locally including more aggressive recruiting of students in STEM majors and better retention of current majors in STEM programs. Much of this work was informed by research on why students leave STEM majors [11,12,15]. While there is substantial and continued research into how and why students leave STEM, there is little understanding surrounding why students might leave a particular STEM field (such as physics) for another STEM field. Understanding these intra-STEM switchers can help physics departments explain attrition rates as well as assist departments in developing a better of understanding of how they are (or are not) meeting the needs of their current and potential majors. Furthermore, physics departments can directly benefit from understanding why students leave or stay in their programs. These data can be used to advocate for curricular changes, new learning environments, and up-to-date teaching practices if their undergraduate program is not meeting the department's desires.

A. Leaving STEM literature

Leaving or switching from STEM to other majors has been explored in many contexts: educational, sociological, and through a discipline based education research (DBER) lens. Results from these studies have continually identified three reasons why students leave STEM: lack of interest [11,12,16]; poor performance [8,12,17–19]; and differential experiences among groups [11,20]. Additionally, increased retention has been linked to reformations in teaching and learning [11,17,21]. Physics education research has also explored this topic finding similar results [8,10,22–24]. This section summarizes the literature that has explored these themes.

A lack of interest, avoiding STEM courses, and other nonperformance based measures can be indicators that students will switch from STEM [11,12,16]. Seymour [11] cited loss of interest in STEM and an increased interest in non-STEM topics as a predictor of switching out of STEM. More recently, Chen [12] examined STEM attrition rates and focused on students who leave STEM for non-STEM

programs or those who dropped out entirely. Chen [12] found that students who avoided STEM courses in their first year were likely to switch out of STEM. Marra *et al.* [16] found that students who left engineering programs reported that a lack of belonging in engineering is a more important factor than performance related factors.

Teaching and learning reformations also have impacted students leaving and staying in STEM [11,17,21]. Seymour [11] found that poor teaching methods by STEM faculty influenced students to leave. The reverse has also been demonstrated: students who were exposed to interactive engagement techniques in introductory courses were more likely to persist in their STEM programs [17,21]. Student persistence in STEM has also been linked to attending colleges and universities that focus on teaching over research [17].

Performance in coursework has also been highlighted as a contributing factor, such that better performance is tied to persistence in STEM [8,12,17–19]. Seymour [11] found a small but significant fraction of students who were enrolled as STEM majors but left STEM reporting conceptual difficulties with STEM coursework. Students required to enroll in lower math courses, and having poor grades in STEM courses were all indicators that students will leave STEM programs [12].

In addition to performance at the university, the prior preparation of students has been demonstrated to effect outcomes at the university [11,15,25–27]. Seymour [11] found that many students believed their high school STEM education provided little to no preparation for the university. Reasons for this lack of preparation included that high school was too easy for these students, that "gifted" students often times were not taught study skills, that students experienced gender discrimination, and that the student's high school lacked resources. Teaching and learning in high school physics has also been linked to performance in university physics and STEM leaving [11,15,25,27]. Students who took high school physics courses that focused on "deep and narrow coverage" outperformed students who took "broad and shallow" courses in their university physics courses [15,25,27]. Ultimately, Seymour [11] noted that students "complain, with good reason, that they had no way to know how poorly they were prepared." Thus, high performance in certain high school contexts can be a predictor of leaving STEM in university.

Different demographic groups have a wide variety of experiences in their education that can impact their choice to stay or leave a major [11,15,20,28–30]. Women and minority groups have reported a "chilly environment" in the classroom and on-campus that is often less experienced by their white male counterparts [28]; this remains true when students from different groups were similar in academic performance [11]. Seymour [11] found that curriculum pace, receiving poor grades while expecting high grades, and competition within STEM majors disproportionately affected men on their decision to leave a STEM major compared to women. Seymour [11] attributed the "weed-out process" having "a greater impact on young men because it carries messages which are intended to have meaning for them…[the weed-out processes] are obscure to young women, and they are thus less directly affected by them." Female students were also less likely to be interested in STEM in their senior year of high school in comparison to their freshman years while male counterparts' interest remained stable over that time [29]. This is true even when they were enrolled in high level STEM courses in high school [30]. Female students have also been shown to earn lower grades in calculus-based physics courses when compared to male students with similar backgrounds [15].

Students from racial and ethnic backgrounds that are underrepresented in STEM also have different experiences in the STEM programs, which can impact their choices surrounding traditional STEM degree programs [11,20,30]. These experiences begin precollege where black and Hispanic students often have had less STEM opportunity than white students [11,30]. For example, black and Hispanic male students were less likely to have substantial STEM preparation in high school compared to white students; however, when these students had similar preparation, they pursued STEM careers in equal measure [30]. Additionally, Seymour [11] found that underprepared students could appear "over confident" in the STEM preparation because they excelled in less than average high school programs. Confronting that reality in college has lead to these students switching from STEM. Finally, participating in high school science and math courses generally has had a positive effect across all groups [20]. However, students who identify as minorities are affected less positively by these programs than white and Asian students.

While a substantial amount of work has investigated students leaving STEM, there has been less work that has focused on leaving physics, specifically. The work that has been done supports the broader work in STEM and has identified better performance [8], use of teaching reformations [10,22,23,31], and increased interest [24] as all having had positive effects on the retention of physics majors. Aiken and Caballero [8] found that students' performance in introductory physics and calculus courses and when these students take these courses was important to completing a physics degree [8]. In other work, physics majors were at least as likely to graduate with a physics degree who had been in reformed introductory courses as those taking traditionally taught courses [10]. Interviews with physics majors who were also learning assistants [31] have suggested that participating in a learning assistant program increased physics major retention [22,23]. Students who become physics majors were also more likely to have expertlike beliefs as measured by the Colorado Learning Attitudes about Science Survey (CLASS) [32] when they entered the university [24].

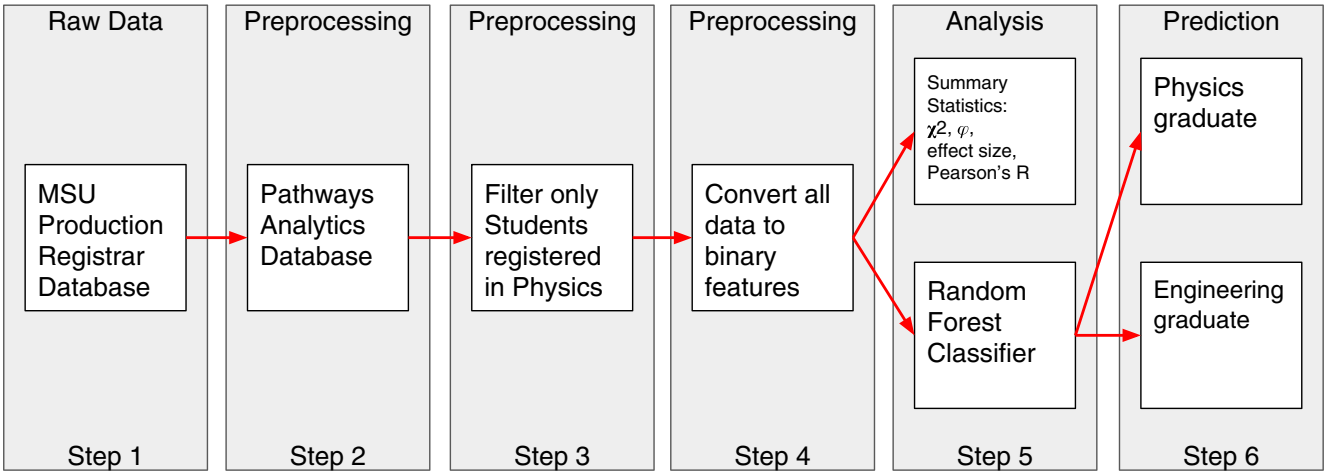

FIG. 1. Raw data are collected by the MSU Office of the Registrar and are used to build the pathways analytics database. The data is filtered to only students who registered as physics majors and then further reduced to binary features. The analysis includes summary statistics (Table I and Fig. 2) and the random forest classifier predicting "physics graduates" and "engineering graduates" (Figs. 3, 4, 5).

III. DATA AND CONTEXT

This work focused on students enrolled in the physics bachelor's program who left for an engineering degree using data collected by the Office of the Registrar at Michigan State University. MSU is a large, land-grant, American university in the Midwest. It typically has had an enrollment of approximately 50 000 students and is predominately white (76.5% across the entire university) [33]. University STEM degree programs are split between the College of Natural Science and the College of Engineering.

The data collected by the MSU Registrar included time-stamped course information that allowed for the examination of features related to students showing a lack of interest (e.g., students take courses much later than their first year). It also contained grade data for courses so performance metrics were examined. Throughout the study, teaching reformations were controlled for by examining the dataset during a time (1993–2013) when there were no research-based teaching reformations enacted at MSU.

The data set used in the following analysis was built from a database collected by the MSU Registrar. Prior to performing any analyses, the data were preprocessed to form what will be called the "pathways analytics database" or "pathways database," for short. The data were preprocessed following the pipeline shown in Fig. 1, which is described below.

A. Pathways analytics database

The pathways database contains three tables describing student demographics, time-stamped course data, and time-stamped major registration data (step 2 in the data pipeline diagram in Fig. 1). The pathways database catalogs every course a student takes and every major they declare including their final degree. It includes grades for every course as well as transfer institution data for courses with transfer credit. It includes demographic information such as gender and ethnicity. For some students, prior preparation data such as high school GPA and performance on the MSU math placement test is also available.

The data within the pathways database was valid for students who began studying between 1993 and 2013. The 2013 cutoff was used because students might not have had adequate time to reach graduation by Spring 2017 (the date when the data was pulled) if they enrolled past 2013. Prior to 1993, MSU was on a quarter system and there were major changes to degree programs and courses after the switch, thus, data prior to 1993 were not used in the analysis.

B. Filtering

To investigate the research questions outlined in Sec. I, a set of appropriate filters was developed and applied to the pathways database. With the focus toward investigating why students leave the physics major for an engineering degree, the data were filtered for only undergraduate students who at some point during their undergraduate program declared a major offered by the Department of Physics and Astronomy (step 3 in the data pipeline diagram in Fig. 1). Additionally, the data set was filtered to only students who ultimately received a degree from either the Department of Physics and Astronomy or from the College of Engineering. The final dataset did not include students who never completed a degree program nor did it include students who switched to non-STEM programs or other STEM programs. Students registered as physics majors who received degrees in physics or engineering made up 66.5% of the students who ever registered as a major in the Department of Physics and Astronomy between 1993 and 2013. A total of 1422 students were analyzed in this study. In this work, the students who registered as physics students and then were awarded degrees from either the

TABLE I. Summary statistics and contingency table analysis. The summary statistics are presented as percentages of the students who graduated with a degree (physics or engineering). Superscript "a" denotes $p < 0.05$, "b" denotes $p < 0.01$, and "c" denotes $p < 0.001$.

Feature	Physics graduate (%)	Engineering graduate (%)	χ^2	ϕ	V
Took Modern Physics	87.09	8.86	476.19[c]	0.34	0.15
Took engineering course	51.35	79.63	42.65[c]	0.03	0.05
High Physics 1 grade	37.84	19.84	40.57[c]	0.03	0.04
High Physics 2 grade	41.59	23.55	36.04[c]	0.03	0.04
High Calculus 2 grade	25.98	12.30	35.39[c]	0.03	0.04
Physics 1 on time	69.22	46.96	30.58[c]	0.02	0.04
Calculus 2 on time	72.37	53.31	20.68[c]	0.02	0.03
Female	16.67	10.71	9.29[b]	0.01	0.02
Calculus 1 transfer credit	55.56	44.58	8.58[b]	0.01	0.02
Physics 2 transfer credit	20.72	14.82	7.03[b]	0.01	0.02
Calculus 1 on time	84.08	73.28	5.28[a]	0.00	0.02
Physics 2 on time	31.83	26.46	3.53	0.00	0.01
Physics 1 transfer credit	29.73	25.27	2.58	0.00	0.01
White or Asian	86.64	82.14	0.85	0.00	0.01
High Calculus 1 grade	16.82	15.08	0.67	0.00	0.01
Calculus 2 transfer credit	32.88	33.33	0.02	0.00	0.00
Total (N)	666	756			

Department of Physics and Astronomy or the College of Engineering are referred to as either "physics graduates" or "engineering graduates." respectively.

The data for each of 1422 students was organized into single vectors of features (or variables). These vectors contained all of the model features that were used to predict their final graduated degree. The features with summary statistics can be found in Table I.

Several features were included in the analysis based on previous literature described in Sec. II. For example, the demographic features were included because the experiences of female students and students who identify as ethnic or racial minorities have been shown to affect graduation outcomes in STEM [11,12,20]. In addition, grades in introductory physics and calculus courses were also included since, in prior work, course performance has been shown to be linked to students switching from physics programs to engineering programs [8].

Furthermore, the time when students take STEM courses has been shown to impact a student's success in a STEM program (earlier is better) [12]; thus, the time at which students take introductory physics and calculus courses have all been included in the analysis.

In addition to the features that have been discussed in prior literature, the impact of the individual courses on earning a degree in physics or engineering was also explored. Including such course-level registration features (specifically taking engineering courses or the first modern physics course) was based upon two hypotheses: (i) students who switched to engineering degree programs might have chosen to do so prior to actually switching in the MSU registrar database and thus may have signed up for required engineering courses, and (ii) students who took the first modern physics course had invested in the physics degree, as this was the first course that was offered with only physics bachelor degree students being required to take it [10].

C. Converting to binary features

For the analysis, all model features were converted to binary features (step 4 in the data pipeline diagram in Fig. 1). While some features were collected as binary (e.g., gender), some were converted to binary. An explanation for each converted feature appears below.

Grades.—Grade features were reduced to a "high or low" binary feature and with a chosen cutoff at ≥ 3.5. Grades can be considered ordinal data [34], and indeed were not on a strict interval scale (e.g., while the minimum grade point at MSU is 0.0 and the maximum is 4.0 there was no ability to earn a 0.5 grade point score). Grades in introductory physics and calculus courses rose beginning in the early 1990s. The average grade in physics and calculus rose by approximately 0.5 grade points (from ~2.5 to ~3.0 for physics, and ~2.0 to ~2.5 in calculus). Thus, the cutoff of ≥ 3.5 GPA was chosen because it was always above the grade increase.

Transfer credit.—Students could come to MSU with transfer credit from Advanced Placement or from other institutions of higher education. When a student had a transfer credit for a particular course instead of a grade, the data was coded with a 0 in the high or low grade feature and a 1 in the transfer grade feature for the course.

Time when courses were taken.—Time features were converted to on-time or late for each course. The split between on-time and late is different for first semester courses and second semester courses. For Physics 1 and Calculus 1, the cutoff was set after the first semester.

This was because the MSU physics department recommends that students take these courses in their first semester of enrollment. For Physics 2 and Calculus 2, the cutoff was set after the first year of enrollment. This was because the MSU physics department recommends that students take these courses in their second semester of enrollment. In all cases, a positive response is when the student has taken the course prior to the established cutoffs.

Ethnicity.—MSU changed the way students reported their ethnicity over the course of the data set due to changes in the Integrated Postsecondary Education Data System (IPEDS) ethnicity definitions in 2007 [35]. Thus, ethnicity definitions were first collapsed into the pre-2007 IPEDS definitions; they cannot be conversely expanded because the new definitions are more nuanced. The data were further collapsed into a binary feature indicating whether the student identified as white or Asian, or as a different reported ethnicity.

Gender.—Gender was collected by the MSU Office of the Registrar as binary data (either male or female) and we used it as such in our model. While we recognize that gender is a complex issue [36], we lack data beyond a binary description.

Engineering courses.—A feature was created that assessed whether or not a student had taken engineering courses. The data were restricted, per student, to the semesters when the student was registered as a physics major in the Department of Physics and Astronomy. For example, a student who took one or more engineering courses during the semesters in which they were registered as a physics major was coded with a 1 for this feature.

First modern physics course.—A feature was created as to whether a student took the first physic course, modern physics, offered at MSU. This feature was coded with a 1 if a student was enrolled in this course at any point.

In addition, two model features which focused on prior preparation were also available for some students. These features included students' reported high school GPA and the score, if taken, that a student received on the MSU math placement exam. These two features require additional explanation that will appear in Sec. VI. The model features described above form all of the features used in the various models throughout the study.

IV. METHODS

In this work, contingency tables [37] and binary classification models [38], specifically random forests, were analyzed to predict whether a student would receive a degree from the Department of Physics and Astronomy or switch to an engineering degree (Step 5 in Fig. 1). Contingency tables have been widely used in PER (see, e.g., Finkelstein *et al.* [39]). Often, researchers employ the χ^2 statistic and effect sizes to assess the correlation of frequency data. For all features described above χ^2 statistics and Cramer's V effect sizes [40] were calculated (Table I).

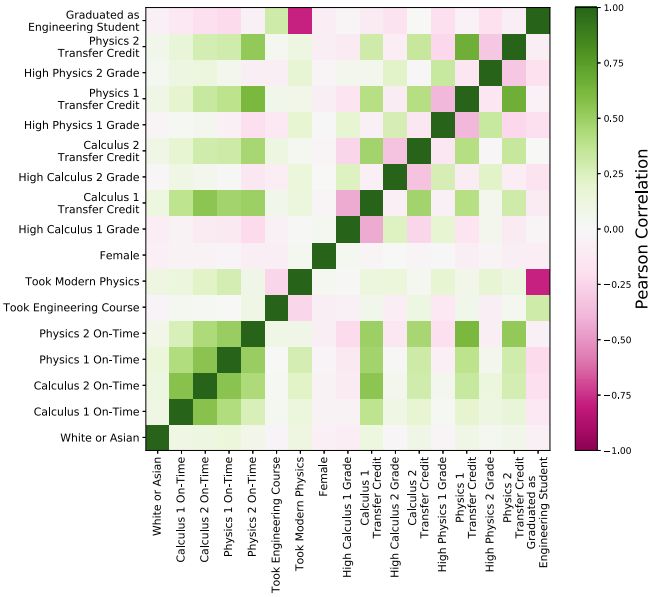

FIG. 2. Pearson correlation matrix for features in the main model.

Additionally the Pearson correlation coefficients [41] were also calculated (Fig. 2).

Binary classification models, specifically logistic regression, have also been used in PER (e.g., Dabney and Tai [2]). In general, a binary classification model predicts an outcome that is discrete and (usually) binary. For example, logistic regression can be used to predict if a student stays or leaves STEM, earns a grade above or below a certain level, or whether or not a student completes a specific course. This technique can be extended to a multiclass outcome variable; however, for the purpose of our analysis, the outcome variable of interest (physics graduate vs engineering graduate) was binary.

Binary classification models are a supervised learning technique that can consist of various algorithms (e.g., logistic regression). In this paper, we have used the random forest algorithm for classification [9] because the underlying features in our analysis were binary (Sec. III C). As it is a fairly new method to PER, this section will provide a summary of the random forest algorithm [9]. Additionally, it will introduce how to assess the model predictions and associated inferential statistics. For a thorough review of the random forest algorithm, see Young *et al.* [13].

A. Random forest

Random forest classifiers use the mode output of a collection of decision trees to predict if the input data belongs in one class or another [9]. The model produces "feature importances" which represent the average change in the decision criterion for each feature in the data set. The larger the feature importance, the more important the feature is to the model. It is important to note that feature

importance is measured *relative* to the other features in the model. An exhaustive review of random forest algorithms and their uses can be found in Refs. [9,38,42].

The power that the random forest algorithm has over a more traditional model (e.g., logistic regression) is that it is an ensemble model [9]. Ensemble models are a collection of submodels whose outputs, taken together, form the prediction. Because random forest is a collection of decision trees, the random forest can fit multiple subpopulations within a dataset. Thus the ensemble of trees in the random forest can fit subsets of the data without overly biasing the model output [43]. By contrast, a model like logistic regression attempts to fit a hyperplane to the entire input dataset [44]. In a random forest, no one tree in the forest is required to be predictive of all the data. Rather, it is the collection of decision trees that make up the random forest that form the predictions.

B. Evaluating a classification model

In PER, researchers are often concerned with not only providing predictive outcomes, but explaining the system that is being studied. However, analyzing the predictive output of an explanatory model is important even if it is not the intent [5]. In doing so, studies can provide a basis for comparison with other models that may also fit data from different settings. Below, we introduce how to evaluate the predictive output of a classification model. These methods are applicable to all classification models (e.g., logistic regression), not only the random forest model used in this paper.

Model predictions are evaluated in a number of ways such as accuracy, receiver operating characteristic (ROC) curves, and the area under the ROC curve (AUC) [38]; the best model can be found using a grid search [42]. For this work, the best model was defined as the model with the highest combination of metrics that produced good predictions. Below, we discuss these evaluations metrics and the relationships between them.

Accuracy is the ratio of true predictions to all predictions made. This measurement has a caveat; when the data are class imbalanced (i.e., classes are not split evenly [45]), the accuracy can be less meaningful [42]. For example, if 90% of the data belong to class A, having a prediction algorithm that assigns class A to all data produces a 90% accuracy. In this analysis, the data do not have a large class imbalance (see Section IV C).

Receiver operator characteristic curves compare the true positive rate (TPR) and the false positive rate (FPR) for a variety of cutoffs [46]. The TPR is the ratio of true positives to the sum of true positives and false negatives. The FPR is 1 minus the ratio of true negatives to the sum of true negatives and false positives.

ROC curve plots have a certain geography to them. The diagonal serves as a boundary; the model is better than guessing if the curve is above the diagonal and worse than guessing if the curve is below it. A curve trending towards the upper left is one that is approaching perfect classification [13].

The area under the ROC curve [7] provides an additional summary statistic for interpreting the quality of a classification model. The closer the AUC is to 1, the better the performance of the model. Performance in this case is defined as the ratio of TPR to FPR.

Random forest models have a number of parameters that govern the size and shape of the forest (e.g., number of decision trees, number of features allowed per tree, the decision criterion, etc.). Because different parameter choices can produce models with different qualities of predictions, we employed a grid search to determine the best choice of parameters for our models [47]. A grid search is a method used to test the total combination of a range of possible model parameters to return the highest fitting scores. In this work, we assessed the model's performance using accuracy and AUC for classifiers. The range of parameters and the Python scripts used to test and evaluate the models in this paper can be found in the online Jupyter notebook [48].

C. Training and testing the model

Testing a model's predictive power requires splitting the data into a training and a testing data set; this worked used a ratio of 70:30. That is, 70% of the data was used to train the model and 30% of the data was used to test the model's predictions. This 70:30 ratio is a common choice for many machine learning techniques [42]. The data is randomly sampled into a testing and a training set without replacement. The testing data is sometimes called "hold out data" because it is held from the model and is not used for fitting the data [42]. In this work, the ratios between the two graduating outcomes for the training and the testing data were the following:

(i) **Physics:** Training: 45.7%, Testing: 47.3%
(ii) **Engineering:** Training: 54.3%, Testing: 52.7%

Thus within the data, the graduating degrees were roughly equivalent in shape and there was little class imbalance.

To produce the highest AUC and accuracy, the model was run repeatedly through parameter variations via a grid search. By performing a grid search, 1920 parameter combinations were tested. The parameter ranges were chosen to vary around the default values of the parameters as specified in the sci-kit learn documentation for random forest classifiers [49]. For example, the maximum number of features available to a tree uses the minimum of the default: $n_{\text{features}} = \sqrt{n_{\text{features,max}}}$ up to half the maximum. The maximum number of feature combinations would exceed the computational capacity available for this project thus some limits were placed on the grid search (such as not varying n_{features} between the true minimum and maximum number of available features) to minimize computational time. The final output model parameters were then used to fit the random forest classifier to the training data set.

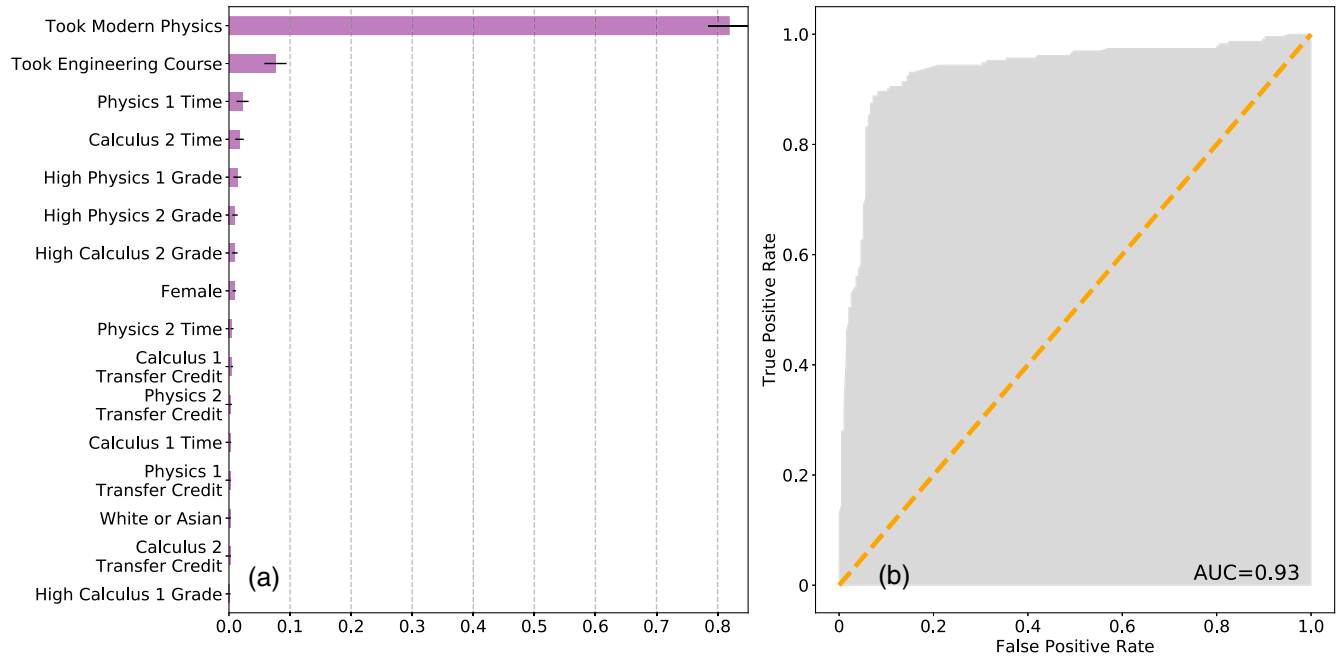

FIG. 3. (a) Feature importances for the main model. The error bars represent the standard error for the distribution of feature importances across all trees within the forest. (b) The ROC curve for the main model. The dashed line represents random chance.

The various parameters found in the Jupyter notebook mentioned above [48] include commonly varied parameters such as the number of trees in the forest and the depth of the forest.

Because of the observed grade increase over time (described in Sec. III C, grades in introductory courses have some time dependence. Thus, it was expected that other features might be time dependent as well. To investigate this and how it might effect the analysis, a sliding window approach was used to analyze this time dependence [50]. To explore if our model and the resulting features were temporally invariant, the data were split into 17 windows centered around each year from 1996 to 2013. For example, a 4 year window centered on 1996 included data from 1994–1998. The model was then refit on each individual window of data. The feature importances were compared for each time window to investigate which features remained invariant over time. To assess the quality of the window size, we used 4, 6, and 10-year window sizes from the beginning of the dataset (1996) to the end (2013) in increments of one year. Since each window size produced similar results, only the 6-year window is reported in this paper.

D. Reducing the model to the minimum viable features

The final way that a model can be evaluated is combining all of the previously mentioned evaluation methods (accuracy and AUC) to determine what is the minimum subset of features necessary to construct a viable model. This method is known as "recursive feature elimination" [51,52]. Using the best parameters provided by the grid search, the recursive feature elimination (RFE) algorithm begins with all of the features available, and removes the least important features (based on feature importance magnitude) one by one until only the most important feature is left. This process produces AUC and accuracy curves that can be compared to one another to determine the most predictive model with the least number of features. In a sense, this process is similar to finding the fewest number of features in a linear regression model that produces a fit that is not statistically different from the model with all the features. In RFE, features that are less important to model prediction have minimal effects on AUC or accuracy when included in the model. It is then assumed that such features do not represent an aspect of the data that is important for prediction.

V. ANALYZED MODELS

Throughout the analysis, multiple models were investigated. Below we describe the details of each model that was used:

(1) **Main model**: The main model included all available features and was used to perform a grid search to find the best parameters that defined the forest (Fig. 3). This model used the entire data set ($N = 1422$ students).

(2) **Recursive model**: The recursive feature elimination (RFE) produced as many models as there were features and used the best parameters to define the forest from the grid search in the main model. This was used to describe which features impacted

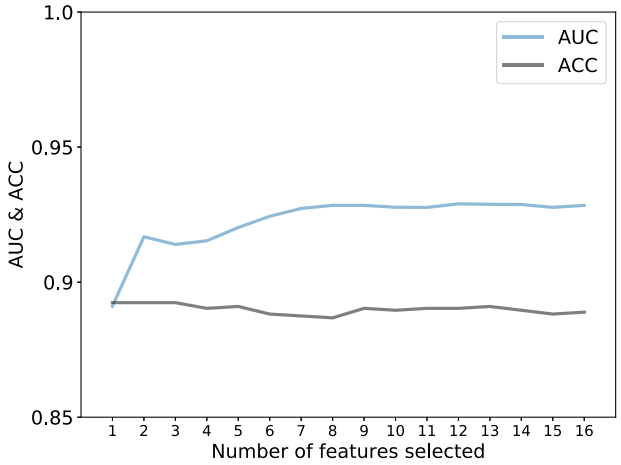

FIG. 4. The AUC and ACC for each number of features selected. The features are ordered the same as Fig. 3.

the model the most (Fig. 4). This model used the entire data set ($N = 1422$ students).

(3) **Sliding time window model**: The sliding time window model produced a model for each time window using the best parameters defined by the grid search in the main model. It used subsets of data built from each time window for the training and the testing data (Fig. 5). This model used the entire data set ($N = 1422$ students).

(4) **Prior preparation model**: A final model was built from training data that had been reduced using complete case analysis. Complete case analysis excludes all data that have missing data. In this case, high school GPA and the MSU math placement test score were added to the model. The model was then trained using the complete case for high school GPA or the MSU math placement test score to determine if these features were important predictors to students staying in physics. This model also used the parameters from the grid search in the main model. This model used two subsets of the data set [N students with HS GPA = 1037 (73%), N students with math placement score = 700 (49%)].

VI. RESULTS

A contingency table analysis demonstrated that taking modern physics had the highest effect ($\phi_{\text{modern}} = 0.34$, $p < 0.001$) on students who earned a degree from the Department of Physics and Astronomy (Table I). While this feature held the highest effect size, it is still relatively small [53]. All other features demonstrated negligible effect sizes. Previously, it was demonstrated that high grades in physics and calculus might indicate students will remain in a physics degree program [8]; however, through this analysis it was found that high grades in both courses had negligible effect sizes on earning a degree from the Department of

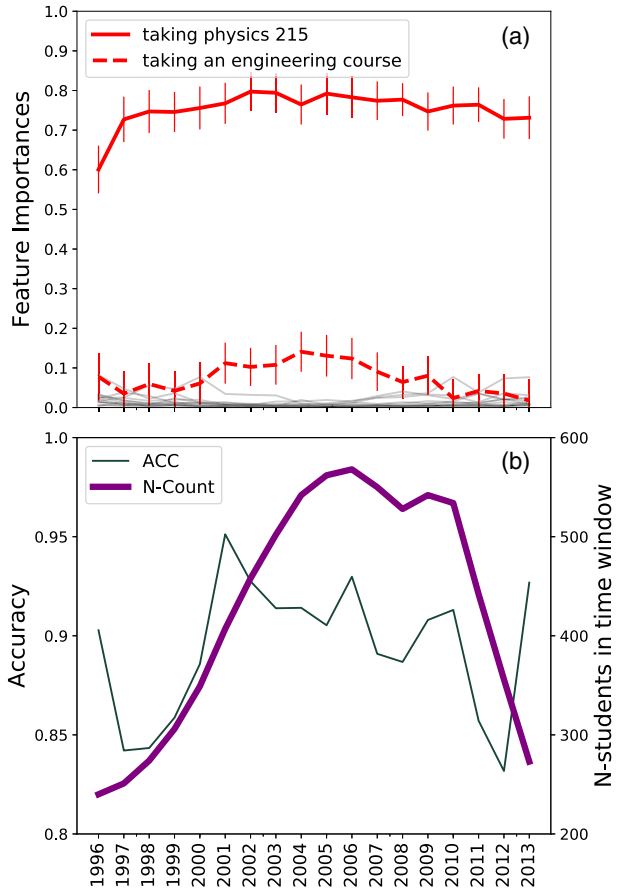

FIG. 5. Sliding time window model. (a) The original grid searched cross validated model parameters were fit to data from sliding time windows (±3 years) with the window center starting in 1996 and running through 2013. The error bars represent the standard error of the feature importance. (b) The accuracy (ACC) for each of the time windowed models. The dip in the number of students in later years was due to the sampling method; since no students who begin after 2013 are included in the dataset, the total number of students from 2010 begins to decrease.

Physics and Astronomy. Outside of taking modern physics and the prediction feature (graduating with a physics or engineering degree), features in this data set showed low linear correlations with each other, but there were still interesting structures (Fig. 2). For example, taking the modern physics course showed a low anticorrelation with taking an engineering course while registered as a physics major (Pearson's $R = -0.24$). Results also showed that there was a positive correlation between the times at which students take courses. This is likely due to these courses having prerequisites from one to the other. In addition, there were small anticorrelations between high grades and transfer grades. This does not indicate that a student with transfer credits for a course might receive low grades for the same course. Given that transfer grades and high grades were mutually exclusive, a student could not have a high grade from a MSU course and have transferred in credit for

that course; this can be explained by how we constructed our data set. Contingency tables and Pearson correlations can give some insight into the data we gathered; however, they cannot give predictions about more individualized results or provide general inference about physics students who switch to engineering. To further explore this, a random forest classifier was employed.

A random forest classifier (the main model from Sec. V) was built to predict whether a student would graduate with a physics or engineering degree based on the features in Table I. The model demonstrated that taking modern physics was of greatest importance to the model's prediction (Fig. 3). This model's predictive ability is high, demonstrated by (AUC = 0.93) the ROC curve being well above the random guessing discrimination line. No other feature had an importance above 0.1. The calculated feature importances are a measure of the mean decrease in Gini impurity each time the feature was used in a tree [9,49]. The second most important feature was a student taking an engineering course while registered as a physics major. Using recursive feature elimination (Recursive model in Sec. V) and comparing the AUC and accuracy, the optimum model had two features: taking modern physics and taking an engineering course as a physics major (Fig. 4). Including any additional features reduced the accuracy of the model. Figure 4 also shows that the AUC increased at a small expense of accuracy by including additional features: having a high grade in Physics 1, Physics 2, and Calculus 2, taking Calculus 2 in the first year, and whether or not the student is female. Including features beyond this did not increase AUC or cause noticeable changes in accuracy.

Given the observed increase in grades in introductory physics and calculus courses over time, the time-dependent nature of our findings were investigated. A random forest classifier trained on sliding windows of data subsets (the sliding time window model in Sec. V) also demonstrated that taking a modern physics course was the most important predictor for a student to receive a physics degree (Fig. 5) at any point in time (i.e., for every window scale centered on any given year). There was little variation in the model for other features outside of the slight increase and decline in the feature importance for taking an engineering course while registered as a physics major.

MSU uses a custom math placement test to place students in math courses who do not have AP credit or high SAT or ACT math scores. These data were not present for all students as they might have had a high school transcript that resulted in them not having to take such assessments. Thus, student performance on this test and a reported high school GPA existed for a subset of the data. 1037 students (73%) had a reported high school GPA; 700 students (49%) had a reported math placement score; and 604 students (43%) had both reported. These features were not preprocessed like the other features in the dataset and thus represent the actual high school GPA reported or the score the student received on the math placement exam. These features were used in the prior preparation model (see Sec. V). These features demonstrated no increase in the overall predictive power of the models for this subset of data. Because these features had no impact on model prediction, and they were missing for a large fraction of the data, these prior preparation features were excluded from all other models. Furthermore, because these features did not impact model prediction, they were not imputed.

VII. DISCUSSION OF RESULTS

This work demonstrated that taking the first course in modern physics was the single most important feature for predicting if a student earned a bachelor's degree from the Department of Physics and Astronomy at Michigan State University. It additionally demonstrated that taking an engineering course while registered as a physics major was an indicator that students will switch and eventually earn a degree from the College of Engineering. Further, it was found that performance as measured by grades in introductory physics and calculus could have small effects on whether a student earned a physics degree or not.

We set out to answer three questions:

(1) Which of the factors identified in the literature impact students to remain a physics major or leave the physics bachelor's program for an engineering degree?
(2) What were the effect sizes of these performance factors and how do they impact models that compare the effect of various factors against each other?
(3) Through this analysis, what did we learn about students who register for physics but leave for engineering (i.e., intra-STEM switchers)?

Section II describes multiple reasons why students might leave STEM. These include a lack of interest, poor teaching in STEM courses, performance in university STEM courses and prior preparation, and differential experiences for different demographic groups. For most of the time period of this study, there were no systematic research-based changes to teaching practices at MSU. Thus, from this data, it is not possible to provide commentary on how poor teaching and how course transformations or changes to pedagogy might have affected student retention in the physics major. Future work will examine how current course transformations in the Department of Physics and Astronomy at MSU (e.g., Ref. [54]) have changed physics major recruitment and retention.

Below, we discuss the unique results from our analysis, namely, the role of the third-semester, modern physics course and how taking engineering courses as a physics major play in earning a physics degrees. Then, we discuss the roles that previous observed features (ie., lack of interest and performance in STEM courses) played in our work. In future sections, we will discuss the limitations and

implications to this research study (Sec. VIII) along with the affordances that the model prediction assessment has above more traditional modeling approaches (Sec. IX).

A. The importance of the first modern physics course

We found that the most important feature in our model that predicted which degree a student earned was whether or not a student took the modern physics course [Fig. 3(a)]. This result remained important as other features were eliminated (Fig. 4) and was consistent over time (Fig. 5). Finding that taking modern physics is the most predictive feature in our model bolsters the assumptions made in prior work. Rodriguez *et al.* [10] decided to filter their data based on whether or not students had taken their modern physics course. If a student had not taken this course, Rodriguez *et al.* [10] did not consider the student to be a "physics major" in their data set. Our work further supports this assumption of taking the first required physics course for all physics majors (i.e., modern physics) was the most predictive feature in the data for finishing with a physics degree through all analyses.

The nature and role of this course in MSU's Department of Physics and Astronomy suggests that such a course should be a strong predictor of completing the physics degree program. The Department expects that students will take modern physics in their third semester of enrollment. It is the student's first exposure to thermodynamics and quantum mechanics at MSU and is also the first physics course that is not required for any major outside of Department of Physics and Astronomy. The course requires students to have completed the introductory physics sequence and be, at a minimum, concurrently enrolled in a multivariate calculus course.

Because this course predicted whether a student stayed in physics or left for an engineering degree, it could serve as a good outcome variable for future analyses. That is, future work will investigate which features predict if a student is likely to take the modern physics course. Understanding which features impact whether or not a student will take this course can help departments understand the potential pathways for future physics majors.

B. Leaving for engineering

Taking an engineering course while registered as a physics major was a smaller, but still predictive, feature in the models. This could indicate that students who leave physics for engineering might not have ever intended to stay in physics, or, perhaps, that they might have found the applied nature of engineering more attractive. Furthermore, the experiences that students encounter in early introductory physics courses could of driven them away from physics and ultimately graduating with an engineering degree.

In contrast to students that leave the physics program for other majors [8], students at MSU who leave the physics for an engineering degree rarely take physics courses beyond the two-semester introductory sequence. Thus, taking an engineering course while registered as a physics major might be a relatively strong signal that a student intends on leaving physics for engineering. It could also be a signal that the student plans to dual major (something our models do not account for). These explanations cannot be confirmed by the data and approaches used in this study and will be further explored in future work. The precise reasons underlying intra-STEM switching are better unpacked using qualitative approaches.

Through the sliding window analysis, results suggested that students taking engineering courses while registered as physics majors had its largest feature importance in the mid 2000s (Fig. 5). This was due, in part, to a sharp visual, but not statistically significant ($\chi^2 = 2.66$, $p = 0.26$), increase in students switching from physics to engineering between 2005 and 2011 (see the Jupyter notebook [48] for supplemental figures). Through discussion with relevant MSU faculty, we could find no credibly documented reason why these students might have left the Department of Physics and Astronomy for the College of Engineering in higher numbers during this period.

Though the more traditional, population-level analysis showed that the effect size of taking an engineering course was negligible ($\phi_{\text{engineering}} = 0.05$, $p < 0.001$, see Table I), our random forest model did see an increase in the predictive power via the AUC when this feature was included (see Fig. 4). However, ultimately, taking an engineering course had a negligible effect size. This implies that subtle features in our dataset might not be best explained by population statistics (e.g., χ^2, effect size). This observation should encourage researchers to use both descriptive and inferential analysis to conduct studies.

C. Lack of interest

While not measured directly, three of the features served as proxies for interest in earning a physics degree:

(1) The time when students took introductory physics or calculus courses relative to their enrollment at MSU (later may indicate less interest),
(2) Whether or not a student took an engineering course while registered as a physics major (doing so may indicate less interest), and
(3) If a student had transfer credit for physics courses (having credit may indicate greater interest).

Aiken and Caballero [8] demonstrated that students who switched from physics to engineering were likely to enroll in introductory physics courses later than those who stay in physics; also confirmed in Table I). In the main model, the 3rd most important feature measured whether students take physics 1 in their first semester (Fig. 3). However, ultimately the feature has such little importance that it does not show a large change in AUC scores when added to the model (Fig. 4). Thus, while it is true that engineering

graduates take introductory physics courses later in their academic career, it does not seem to have a large impact on whether a student earns a physics or engineering degree relative to the other features in the model. When compared to the Chen [12] result, namely, if students avoid STEM courses in their first year that is indicative of leaving STEM, the results presented above demonstrated that perhaps when students take their physics courses was inconsequential to switching from physics to engineering. However, this could be a signal that only certain STEM courses are indicative of moving within STEM. Chen [12] highlights mathematics, saying "proportionally more STEM leavers than STEM persisters did not earn any math credits in their first year." In this study, this effect was not observed as students are required to complete the introductory calculus sequence to earn a degree in either physics or engineering.

Ultimately, 613 (81%) students who left for engineering took at least one engineering course while registered as a physics major. These courses are likely to be prerequisites to switch majors. For example, the most likely engineering course to be taken is CSE 231, Introduction to Programming 1; 326 (43.1%) students switching from a physics degree to an engineering degree took this course. It is a prerequisite for admission into the College of Engineering to pursue a degree in Computer Engineering, Computer Science, and Mechanical Engineering [55]. Additionally, Aiken and Caballero [8] found that students who switched to engineering were more likely to take introductory physics in their second and third semesters whereas students who stayed in the physics program were more likely to take physics in their first semester. Taking introductory physics in the second and third semester is explicitly recommended by the Mechanical Engineering department at MSU [56]. Thus, the courses that students take and when they choose to take them are indicators of a lack of interest in physics and therefore, will ultimately leave the physics major for an engineering degree.

Finally, the engineering graduates also had fewer transfer credits for physics courses than physics degree earners. Students who switched to engineering from physics were also less likely to take upper division physics courses in comparison to the students who switched to other majors from physics [8]. These findings suggest that students who switched to engineering demonstrate less interest in physics based on what courses they choose to take early in their college experience, including their pre-MSU academic careers.

It is important to note that the above claims about students' interest in physics are made from features that may be considered "measures of interest"; these features are only proxies for a student's expressed interest in physics. Further work should be done to explore these claims explicitly since this is outside of the scope of this paper.

D. Performance in coursework

A result of the above analysis that is in tension with prior work performed by Aiken and Caballero [8] is the small impact that grades have in predicting whether a student will earn a physics or engineering degree [8]. Previously, researchers found that students who switched from physics to engineering performed below average in introductory physics and calculus courses in comparison to students who stayed in physics [8]. It was assumed in the prior study that performance in a course could have a profound impact on a student's persistence in the physics major. Moreover, it has been well documented that such performance measures are important to STEM persistence [11].

In the current work, course performance, measured by grades, was significant but with a negligible effect size in the population level analysis (Table I and Fig. 2). However, none of the random forest models found these performance features to be particularly important (measured by feature importance) to make accurate predictions. We posit three reasons for this finding. First, the methods used in the previous study [8] were different than those used in the analysis described above. Here, using random forests, models were constructed to compare the relative effect of the explored features. In Aiken and Caballero [8], only the effects of individual features were explored. Second, in previous work, grades were treated as a continuous feature and in the above analysis, they are treated as ordinal features. Finally, it might be that grades were not an important indicator for students leaving one STEM program for another STEM program (i.e., leaving physics for engineering).

In Aiken and Caballero [8], grades were compared using Z scores [57]. The Z score normalizes all data to unit variance; thus, the larger an outlier, the more weight (visually) can be given to the score. As the goal of prior work was to provide methods for visualizing these types of data, the decision to use Z scores was driven by considering compelling ways to represent our data, but the resulting analyses might not have been appropriate. Additionally, the previous work did not make any comment on an "effect size" that could represent the population comparisons effectively. While Aiken and Caballero [8] state that physics graduates performed better than engineering graduates in their introductory physics and calculus courses, the degree to which they performed better was not well quantified. This could be because the analyzed Z score did not provide a proper normalization given that the underlying grade data is ordinal [58].

In the present case, conflicting results were found. At the population level, having a high score in Physics 1, Physics 2, and Calculus 2 was statistically significant. But, each feature had a negligible effect on graduating with a physics degree (Table I, Fig. 2). However at the individual level, (i.e., the random forest classification modeling) grades were not important features (Figs. 3 and 5). In fact,

through recursive feature elimination, performance in these courses did not substantially impact the reduced models (Fig. 3).

By comparing the descriptive analysis (using population statistics) to the inferential analysis (using the random forest model), conflicting results were found, which indicate how the two different approaches can compare the relative impact of different features. This conflict might simply be due to the differences one encounters when conducting analysis using descriptive statistics compared to modeling; the descriptive population statistics aggregate all of the available data into a single result. In the descriptive analysis, a comparison was made between the two groups (physics and engineering graduates) and because the number of student records was reasonably high, we might expect to find a statistically significant difference in these distributions. The random forest model, however, places each student record into a class (physics or engineering) depending on the available data and computes the *relative* importance of each feature against the others.

Grades might also not have been important for students leaving physics programs for engineering programs. Seymour [11] identified the assumption of grade differences as a "barrier" to understanding why "able students" leave STEM. In the above study, students who left physics for engineering are "able" in Seymour's words. Course grades were included in the above study because Aiken and Caballero [8] demonstrated that they might be important and that initial findings demanded further exploration. However, in this analysis, grades were shown to have no effect on the random forest models and negligible effect sizes. Seymour [11] emphasized alternate reasons for leaving such as faculty obsession with "weeding out" as opposed to supporting students and a lack of peer support. It could be that these alternative reasons (or other reasons like interest in engineering) are what motivates these students to leave the physics bachelor's degree program for engineering and further research is needed to fully understand why.

VIII. LIMITATIONS AND IMPLICATIONS

A limitation of this study is that the data were confined to MSU and thus, cannot comment on how broad the findings might be regarding physics programs in general. Additionally, because the physics program at MSU is populated largely by white or Asian students (84.2%), there is not enough statistical power to make strong predictions about the effects of race or ethnicity in our model. We aim to develop models with other institutional partners in order to provide discussion about the robustness of our claims. However, we find it promising that our work supports the choices made in Rodriguez *et al.* [10] given the broad differences in student populations between the two institutions.

In addition, prior preparation data (e.g., high school GPA, math placement score, etc.) was not present for all students in the study. With the aim of studying intra-STEM switching, a review of the STEM retention literature suggested that prior preparation is a key factor in STEM persistence [11,12]. On the other hand, some work suggests that high school GPA might not be very predictive of student outcomes at the university [59]. To explore the impact of prior preparation, the models were analyzing with complete cases; some students had incomplete records prior to enrolling at MSU. Because of this, we are unable to comment on precisely how accurate our predictions might be regarding prior preparation. Through our analysis, we believe that prior preparation (as we have measured it) has a small feature importance with regard to intra-STEM switching. But, we acknowledge that it could be that both physics and engineering graduates have higher than average high school GPAs and math placements and this is why we observe that these features have a small importance in our prior preparation model.

Finally, our analysis is entirely predicated on data collected by the MSU registrar. As such, we can point to features that we find to be analytically predictive of students earning a degree in physics or engineering from MSU. However, we are unable to comment on the mechanism by which this happens or to provide any clear narratives beyond what has been presented. Qualitative work that includes interviews, focus groups, and case studies would be needed to unpack the underlying mechanisms and narratives that underpin the results presented here.

In addition to limitations within the data, the methods presented have limitations as well. First, ROC curves represent the full decision threshold space for the classification using the output probabilities from the classification model [46]. Thus the tails of the ROC curves may be less valuable. Additionally, we have not presented any curve fitting method in this paper for determining the optimum decision threshold via the ROC curve [60]. Instead we have opted to use visual inspection and area under the curve methods to determine if the classification model is producing believable predictions. With regards to the random forest model, random forest feature importances are relative measurements with regards to the decision trees within the forest [9]. They represent average values of changes in Gini importance and cannot be used to produce odds ratios like a logistic regression model. Additionally, for very large datasets random forests are computationally intensive. Even in this study ($N = 1422$) the grid search was limited due to available computing resources.

The results presented suggest several implications that physics departments could put into practice. First, providing strong early motivations for taking the first modern physics course could help students decide if physics is a better choice for them as the introductory course sequence revisits most of high school physics. Second, if students come to the physics department with the express intent to switch to engineering, supporting these students'needs is

likely different from supporting the needs of students who intend to stay in physics. Lastly, performance does not appear to be a deciding factor for intra-STEM switching (at least with regard to physics and engineering). Thus, departments should place less emphasis on grades in introductory courses as a determining factor for which students to "actively recruit" into the major.

This work suggests several new lines of research that were not explored here but we intend to explore in the future. First, given that taking modern physics is such an important indicator for graduating with a physics degree, we intend to explore which features are predictive of students taking this course. Second, we intend to investigate what other features characterize a student who switches to engineering or those who switch from another degree program. Third, we plan to research what narratives underlying this intra-STEM switching can research using other methods discover as mechanisms for our observations here and in the aforementioned planned work. Finally, broadening the scope of this work, we intend to explore which features are important for students who switch between other degree programs, leave STEM, or leave the university all together. Extending this work will provide a more complete view of the complex system students navigate from freshman year to graduation.

IX. DISCUSSION OF METHODS

Using the predictive output of a model is important to adopt because these methods allow direct comparison to other models from different settings attempting to predict student degree outcomes in physics. For example, the fitted model from MSU data can be directly applied to other institution data and the predictive output can be assessed. Second, a model fit on other data may provide other explanatory factors while still offering the same predictive power. In both cases this aids in verifying the explanatory features in the model as those that actually describe the system being sampled. This research provides an example, grounded in PER data and possibly a intuitive result (e.g., modern physics is the most important feature when predicting if a student will switch to engineering from a physics program), as to why prediction is important in quantitative research [5].

Figure 3 demonstrates the feature importance and ROC curve that was calculated for the main model. Most models used in PER provide some feedback on the importance of each feature to the model's performance (e.g., calculating the odd's ratios for a logistic regression model's coefficients [2]). These feature importances give researchers a measurement that can be used to compare one feature in a model to another. In the case of logistic regression, the significance of these features is commonly evaluated via p values and goodness-of-fit tests [44]. p values have been demonstrated to be an incomplete measurement of the significance of a statistic [61]. Goodness-of-fit tests frequently rely on residual analysis and do not always use sequestered hold out data [5,44,62]. The above analysis used hold out data to analyze the predictive output of the models used in this paper. Hold out data provide an additional component of model analysis which helps researchers determine what the model means by examining the predictive output [5,62]. Analyzing a model's predictive ability does not give researchers any additional insight into explaining the system they are studying [5]. It does, however, provide a standard way of comparing model results. Additionally, analyzing the ROC curve is a method available to all classification models (e.g., logistic regression) and is not reserved only for random forest classifiers. Thus, analyzing the predictive output can be directly applied to other models published in PER in addition to the already tried and true methods of p-value analysis and goodness-of-fit tests.

Figure 4 provides an analysis of the model AUC and accuracy for the recursive model. The recursive feature elimination method is similar in concept to methods of dimensionality reduction [e.g., factor analysis, principal component analysis (PCA)] already used in PER (see, e.g., Ref. [63]). However, unlike factor analysis and PCA, RFE allows researchers to completely exclude certain features as opposed to projecting existing features into a new space [51]. However in some cases, researchers may see completely removing features from a model as inappropriate. Removing features could be inappropriate because there is a strong theoretical framework as for why a system would behave a certain way. Thus, if a model does not fit the data, then the model is rejected, not the data. Using RFE does not suggest that some features are theoretically inappropriate for a model. RFE is a way to measure the contribution of a feature to model prediction or goodness of fit without projecting the feature into a new space such as factor analysis or PCA. RFE is somewhat similar to methods already used to assess regression models in PER (i.e., explained variance in a model due to a feature [64]). However, in the case of a regression model, this typically uses an R^2 statistic that is calculated from the residuals used to fit the entire data set rather than on additional hold out data. There are R^2 statistics for regression modeling that use hold out data [52]; however, to the best of our knowledge, we are unaware of them being used in PER. RFE can show, relatively easily, the contribution that each feature has to the predictive power of the model (see Sec. VII).

In addition to examining the model's predictive output via accuracy and AUC, the above analysis used the large dataset to explore the stability of the result over the entire time domain (Fig. 5). Analyzing the data for different time windows confirmed the result that taking modern physics was the most likely indicator for completing a degree in physics. The use of a direct model comparison assessing each time window's model accuracy and feature importances provides researchers with further evidence that supports the predictive power of the model of interest. Finally, the

time domain analysis allowed for the examination of grade inflation (see Sec. III C) and what effect it may have had on the model. Since the total number of 3.5 or 4.0 grades grew over time, this modern approach of a sliding time window allowed us to explore the time dependence of the independent variables in relation to the model's predictive power. While not every dataset in PER is as extensive as the dataset used in this paper, this research demonstrates a modern way to compare complex models.

X. CONCLUSION

This research study has introduced to the PER community a new approach to evaluate classification models. Evaluating the predictive output of a model can provide a basis for comparing complex models across different institutional settings. In fact, the models used in the analysis can be readily made available for testing on other institutional data upon request [48].

This work attempted to take a first look at intra-STEM switching. The analysis focused on students who register as a physics major and either stay or leave the physics program for an engineering degree at Michigan State University. Using registrar data from MSU, a random forest classifier demonstrated that taking the first course in modern physics is a strong indicator that a student will stay in the physics program. In addition, results demonstrated that students that leave for engineering programs may "prepare" to do so by taking engineering courses while registered as physics majors. Finally, through this current analysis, it seems that Aiken and Caballero [8] overstated the importance of performance in introductory physics and calculus courses; grades in these courses were not of high importance in any of the models evaluated.

With the focus on assessing the quality of the predictive output of models, in this case, the ability to accurately predict who will complete a degree in physics or switch to engineering, the analysis allowed for direct comparison between contingency table results and a random forest model results. Summary statistics were shown to be "significant" via their calculated p values and the size of their effect helped determine the magnitude of this significance. However, p values have been demonstrated to lack the power to conclusively demonstrate significance [61]. In this study, although taking the modern physics course was statistically significant (see Table I), the effect size was "small" and negligible for all other features. Descriptive statistics indicated that taking this course was not very important to graduating with a degree in physics. However, when used in the random forest classification model, it was shown that taking this first required physics course for physics majors was, intuitively, very important to predicting who will remain in the physics program or leave for an engineering degree. The model result does not invalidate the contingency table analysis. In fact, both results support each other. The contingency table demonstrates a statistical significance, while the model analysis provides its *relative* importance to other explanatory features in the data. The model also demonstrates that while the effect size is small for taking modern physics, this feature still should not be dismissed from understanding why students leave for engineering. While ultimately, this problem may not have needed this sophisticated analysis to demonstrate that modern physics is important for graduating with a bachelor's degree in physics, this research has provided the basis for demonstrating the importance of model building and assessing the predictive output of complex models.

ACKNOWLEDGMENTS

This project was supported by the Michigan State University College of Natural Sciences including the STEM Gateway Fellowship, the Association of American Universities, the Olav Thon Foundation, and the Norwegian Agency for Quality Assurance in Education (NOKUT), which supports the Center for Computing in Science Education. Additionally, we would like to thank Gerald Feldman who suggested the concept of the sliding window analysis.

[1] L. Ding and M. D. Caballero, Uncovering the hidden meaning of cross-curriculum comparison results on the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **10**, 020125 (2014).

[2] K. P. Dabney and R. H. Tai, Comparative analysis of female physicists in the physical sciences: Motivation and background variables, Phys. Rev. ST Phys. Educ. Res. **10**, 010104 (2014).

[3] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj, Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?, Phys. Rev. ST Phys. Educ. Res. **8**, 020104 (2012).

[4] L. Ding, X. Liu, and K. Harper, *Getting Started in PER-Reviews in PER*, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2012).

[5] G. Shmueli, To explain or to predict?, Stat. Sci. **25**, 289 (2010).

[6] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. **30**, 1145 (1997).

[7] J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology **143**, 29 (1982).
[8] J. M. Aiken and M. D. Caballero, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 28–31.
[9] L. Breiman, Random forests, Mach. Learn. **45**, 5 (2001).
[10] I. Rodriguez, G. Potvin, and L. H. Kramer, How gender and reformed introductory physics impacts student success in advanced physics courses and continuation in the physics major, Phys. Rev. Phys. Educ. Res. **12**, 020118 (2016).
[11] E. Seymour, *Talking about Leaving: Why Undergraduates Leave the Sciences* (Westview Press, Boulder, CO, 2000).
[12] X. Chen, STEM Attrition: College Students' Paths Into and Out of STEM Fields, Statistical Analysis Report No. NCES 2014-001, Technical Report, 2013.
[13] N. T. Young, G. Allen, J. M. Aiken, R. Henderson, and M. D. Caballero, arXiv:1810.07859.
[14] N. R. Council, *Adapting to a Changing World: Challenges and Opportunities in Undergraduate Physics Education* (National Academies Press, Washington, DC, 2013).
[15] R. H. Tai and P. M. Sadler, Gender differences in introductory undergraduate physics performance: University physics versus college physics in the USA, Int. J. Sci. Educ. **23**, 1017 (2001).
[16] 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).
[17] A. L. Griffith, Persistence of women and minorities in STEM field majors: Is it the school that matters?, Econ. Educ. Rev. **29**, 911 (2010).
[18] E. J. Shaw and S. Barbuti, Patterns of persistence in intended college major with a focus on STEM majors., NACADA J. **30**, 19 (2010).
[19] G. Hackett and N. E. Betz, An exploration of the mathematics self-efficacy/mathematics performance correspondence, J. Res. Math. Educ. **20**, 261 (1989).
[20] X. Wang, Am. Educ. Res. J. **50**, 1081 (2013).
[21] J. Watkins and E. Mazur, Retaining students in science, technology, engineering, and mathematics (STEM) majors, J. Coll. Sci. Teach. **42**, 36 (2013).
[22] E. Williams, J. Zwolak, and E. Brewe, 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. 436–439.
[23] L. Harris, I. Beatty, and W. Gerace, in *Proceedings of the 2013 Physics Education Research Conference, Portland, OR*, edited by P. V. Engelhardt, A. D. Churukian, and D. L. Jones (AIP, New York, 2013), pp. 173–176.
[24] K. Perkins and M. Gratny, in *Proceedings of the 2010 Physics Education Research Conference, Portland, OR*, edited by C. Singh, N. Rebello, and M. Sabella (AIP, New York, 2010), pp. 253–256.
[25] P. M. Sadler and R. H. Tai, Success in introductory college physics: The role of high school preparation, Sci. Educ. **85**, 111 (2001).
[26] Z. Hazari, G. Sonnert, P. M. Sadler, and M.-C. Shanahan, Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study, J. Res. Sci. Teach. **47**, 978 (2010).
[27] M. S. Schwartz, P. M. Sadler, G. Sonnert, and R. H. Tai, Depth versus breadth: How content coverage in high school science courses relates to later success in college science coursework, Sci. Educ. **93**, 798 (2009).
[28] M. Crawford and M. MacLeod, Gender in the college classroom: An assessment of the "chilly climate" for women, Sex Roles **23**, 101 (1990).
[29] P. M. Sadler, G. Sonnert, Z. Hazari, and R. Tai, Stability and volatility of STEM career interest in high school: A gender study, Sci. Educ. **96**, 411 (2012).
[30] W. Tyson, R. Lee, K. M. Borman, and M. A. Hanson, Science, technology, engineering, and mathematics (STEM) pathways: High school science and math coursework and postsecondary degree attainment, J. Ed. Stud. Pl. Risk **12**, 243 (2007).
[31] V. Otero, S. Pollock, R. McCray, and N. Finkelstein, Who is responsible for preparing science teachers?, Science **313**, 445 (2006).
[32] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).
[33] Fall Enrollment Numbers at MSU, https://opb.msu.edu/functions/institution/documents/FallEnrollment.pdf.
[34] G. R. Pike, G. D. Kuh, and R. C. Massa-McKinley, First-year students' employment, engagement, and academic achievement: Untangling the relationship between work and grades, J. Stud. Aff. Res. Pract. **45**, 560 (2008).
[35] IPEDs definitions, https://nces.ed.gov/ipeds/report-your-data/race-ethnicity-reporting-changes (accessed: 10-23-2018).
[36] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).
[37] B. S. Everitt, *The Analysis of Contingency Tables* (Chapman and Hall/CRC, London, 1992).
[38] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, Springer Series in Statistics (Springer, New York, USA, 2001), Vol. 1.
[39] N. D. Finkelstein, W. K. Adams, C. Keller, P. B. Kohl, K. K. Perkins, N. S. Podolefsky, S. Reid, and R. LeMaster, When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment, Phys. Rev. ST Phys. Educ. Res. **1**, 010103 (2005).
[40] A. C. Acock and G. R. Stavig, A measure of association for nonparametric statistics, Social Forces **57**, 1381 (1979).
[41] J. Lee Rodgers and W. A. Nicewander, Thirteen ways to look at the correlation coefficient, Am. Statistician **42**, 59 (1988).
[42] A. Géron, *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media, Inc., Boston, MA, 2017).

[43] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, J. Chem. Inf. Comput. Sci. **43**, 1947 (2003).
[44] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression* (John Wiley & Sons, New York, 2013), Vol. 398.
[45] N. V. Chawla, N. Japkowicz, and A. Kotcz, Special issue on learning from imbalanced data sets, SIGKDD Explor. Newsl. **6**, 1 (2004).
[46] T. Fawcett, ROC graphs: Notes and practical considerations for researchers, Technical Report, 2004.
[47] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, A practical guide to support vector classification, Technical Report, 2003.
[48] Figures and supplementals for this paper can be found in this github repository, https://github.com/learningmachineslab/publication_notebooks/blob/master/physics_engineering_pathways.ipynb.
[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).
[50] T. G. Dietterich, in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (Springer, New York, 2002), pp. 15–30.
[51] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Mach. Learn. **46**, 389 (2002).
[52] I. Guyon and A. Elisseeff, *Feature Extraction* (Springer, New York, 2006), pp. 1–25.
[53] C. J. Ferguson, Prof. Psychol. Res. Pract. **40**, 532 (2009).
[54] P. W. Irving, M. J. Obsniuk, and M. D. Caballero, P^3: A practice focused learning environment, Eur. J. Phys. **38**, 055701 (2017).
[55] Admission to College of Engineering Requirements, https://www.egr.msu.edu/undergraduate/academic/admission-engineering (accessed: 2018-10-23).
[56] Suggested course schedule for Mechanical Engineering, https://www.egr.msu.edu/sites/default/files/content/UGS/prereq_flowchart_me_fs16_0.pdf (accessed: 2018-10-23).
[57] E. Kreyszig, *Advanced Engineering Mathematics* (John Wiley & Sons, New York, 2010).
[58] A. Fielding, M. Yang, and H. Goldstein, Multilevel ordinal models for examination grades, Stat. Model **3**, 127 (2003).
[59] J. Noble and R. Sawyer, Technical Report No. ACT-RR-2002-4 (American College Testing Program, Iowa City, IA, 2002).
[60] R. M. Centor, Signal detectability: The use of ROC curves and their analyses, Med. Dec. Making **11**, 102 (1991).
[61] G. M. Sullivan and R. Feinn, Using effect size—or why the P value is not enough, J. Grad. Med. Educ. **4**, 279 (2012).
[62] D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow, A comparison of goodness-of-fit tests for the logistic regression model, Stat. Med. **16**, 965 (1997).
[63] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).
[64] G. Potvin and Z. Hazari, Student evaluations of physics teachers: On the stability and persistence of gender bias, Phys. Rev. Phys. Educ. Res. **12**, 020107 (2016).