Journal of Data Science 21 (4), 745–760
October 2023

DOI: 10.6339/22-JDS1068
Data Science in Action

Identifying Prerequisite Courses in Undergraduate Biology Using Machine Learning

Youngjin Lee[1,*]

[1]3940 N. Elm St., Denton, TX 76207, University of North Texas, USA

Abstract

Many undergraduate students who matriculated in Science, Technology, Engineering and Mathematics (STEM) degree programs drop out or switch their major. Previous studies indicate that performance of students in prerequisite courses is important for attrition of students in STEM. This study analyzed demographic information, ACT/SAT score, and performance of students in freshman year courses to develop machine learning models predicting their success in earning a bachelor's degree in biology. The predictive model based on Random Forest (RF) and Extreme Gradient Boosting (XGBoost) showed a better performance in terms of AUC (Area Under the Curve) with more balanced sensitivity and specificity than Logistic Regression (LR), K-Nearest Neighbor (KNN), and Neural Network (NN) models. An explainable machine learning approach called break-down was employed to identify important freshman year courses that could have a larger impact on student success at the biology degree program and student levels. More important courses identified at the program level can help program coordinators to prioritize their effort in addressing student attrition while more important courses identified at the student level can help academic advisors to provide more personalized, data-driven guidance to students.

Keywords *attrition; Educational Data Mining; Learning Analytics; STEM education; student success*

1 Introduction

As our society becomes more complex, it is important to be able to solve problems by gathering and evaluating information, and making data-driven decisions. Since these are the kinds of skills students can develop in STEM courses, education researchers and policy makers have been emphasizing importance of attracting and maintaining students in STEM programs in college (Ehrenberg, 2010; Olson and Riordan, 2012; Sullivan, 2006; Xie and Killewald, 2012). Unfortunately, however, the number of students who choose a STEM degree is decreasing, and many students who matriculated as a STEM major either drop out or change their major. A report from the National Center for Educational Statistics indicated that only 13.7% of entering undergraduate students in 2003 intended to major in STEM, which is about 9% decrease compared to the 1995–1996 cohort (Chen and Weko, 2009). From 2003 to 2006, just 50% of students who matriculated as a STEM major earned a STEM degree or continued studying STEM subjects (Chen and Ho, 2012). Similar patterns were found in more recent studies. Chen and Soldner (2013) reported that about half (48%) of bachelor's degree students and more than two-thirds (69%) of associate's degree students who entered STEM fields between 2003

[*] Email: Youngjin.Lee@unt.edu.

Received May 6, 2022; Accepted September 16, 2022

and 2009 had left these fields by spring 2009. The study conducted by National Science Board (2018) indicates that more than 30% of students who started as a science or engineering major left the field within two years. These findings call for studies on STEM persistence and dropout because shortage of STEM graduates will cause a workforce deficit and have a negative effect on economy (Olson and Riordan, 2012).

Researchers in various fields, from higher education to educational data mining, studied several factors that may have an impact on STEM degree completion and dropout, and tried to predict whether students will be able to earn a STEM degree in time. According to Cromley et al. (2015), cognitive (e.g., GPA, ACT/SAT score, STEM achievement in high school, etc.), motivational (e.g., STEM self-efficacy, interest, relevance of STEM materials, etc.) and institutional factors (e.g., course timing and registration policy, career counseling, financial aid policy, etc.) are playing an important role in helping students persist in the STEM degree program. Various analysis methods, such as educational data mining (Berens et al., 2019; Aulck et al., 2019), longitudinal data analysis (Bettencourt et al., 2020), survival analysis (Chen et al., 2018), latent growth model (Dai and Cromley, 2014), and hierarchical model (Le et al., 2014), have been used to investigate why students are dropping out of STEM programs, and to build predictive models of STEM degree completion and dropout.

As summarized in the Related Work section below, these studies analyzed either data from a self-reported survey assessing emotional and motivational factors and/or institutional data such as demographic information and academic performance to explain how predictors can explain or predict the likelihood of successfully completing a STEM degree. However, since these studies used *aggregated* academic performance, such as cumulative GPA, as a predictor, they were not able to identify *individual* courses having a large effect on student success in earning a STEM degree. In the present study, students' timely graduation was predicted based on their performance on individual courses, allowing for identifying important prerequisite courses from data routinely collected by institution of higher education. More specifically, this study seeks to answer the following two research questions: (1) to what extent can we predict timely graduation of college students majoring in biology?; and (2) which freshman year courses had a greater impact, positive or negative, on timely graduation?

In order to answer these questions, predictors reflecting student performance in freshman year courses were generated from registrar records at a large, public research university in the southern US. Then, five machine learning models estimating a likelihood of completing a degree program in time were developed, and their predictive power was compared. Finally, importance of predictors was examined at the degree program and at the student level to identify important prerequisite courses affecting student success in earning a biology undergraduate degree in time.

2 Related Work

Many learning theories, such as Ausubel's meaningful learning theory (Ausubel, 1963), Bruner's spiral theory of learning (Bruner, 1974), Gagné's theory of instruction (Gagné and Briggs, 1974), and Reigeluth's elaboration theory (Reigeluth, 1979), emphasize importance of sequencing and organizing educational contents to facilitate student learning. As a result, identifying "prerequisite relations" by performing a task analysis of a domain has become an important part of designing effective educational materials (Reigeluth et al., 1978). Although some efforts have been made to empirically validate the structure of a target knowledge domain, identifying prerequisite relations among program components have been relying mostly on subject matter experts and not data-driven. This study applied machine learning to registrar records, such as course grade, to identify prerequisite relations among the courses students took in their freshman year.

Based on the focus of data analysis, previous studies can be categorized into two classes: (1) identifying important variables that can explain student success; or (2) developing a model that can predict student success in a course or degree program. Studies focusing on explaining student success tend to utilize conventional statistical methods while studies predicting student success typically take a machine learning approach to develop a predictive model based on self-reported survey, demographic and academic performance data. The following sections summarize the findings from previous studies explaining or predicting student success.

2.1 Explaining Student Success in College

Thompson and Bolin (2011) performed a series of Chi-squared tests on institutional data to investigate student attrition in STEM degree programs. They found that the proportion of students who earned, switched from or dropped out of STEM programs is different depending on major, ethnicity and high school ranking of students. Smith et al. (2012) developed a multiple regression model to examine the impact of various explanatory variables on the grade of students in an accounting course. Although their analysis found several statistically significant explanatory variables (e.g., age, gender, major, English as a second language, and deferring completion of the first accounting unit, etc.), their model explained only 10% of the variance in the data. Cochran et al. (2013) studied relationships between various explanatory variables (e.g., age, gender, ethnicity, GPA, financial aid status, major and previous withdrawal experience, etc.) and student retention in online courses. The results from their LR analysis indicate that GPA in college classes and class standing (senior vs. non-senior) are important characteristics of students regardless of gender, ethnicity and other factors. Patrick et al. (2021) performed a LR analysis to examine relationships between engineering interest, recognition and performance/competence assessed in a self-reported survey and one-year persistence of engineering students after controlling for gender, major, classification and mother's education. Although they found that major, classification, and engineering interest showed a statistically significant relationship with persistence of students, their LR model explained only 6% of variance in observed engineering persistence, suggesting a weak predictive power for future predictions.

Lee and Ferrare (2019) analyzed the Beginning Postsecondary Student Longitudinal Study (BPS: 04/09) data, which captures parental information (education and income), academic performance (incoming college credits taken in high school, and GPA in the freshman year), financial context (employment, loans or tuition) and major of 16,680 first-time beginning students, and characteristics of institution they attended (Historically Black Colleges and Universities, Hispanic Serving Institutions or doctoral-granting). Their LR model suggests that students who switched from STEM to non-STEM were more likely to drop out of college, compared to students who stayed in their initially declared major. Bettencourt et al. (2020) analyzed the Education Longitudinal Study (ELS: 2002/12) data to compare first-generation students to continuing-generation students who were enrolled in STEM degree programs. Their multinomial regression analysis indicates that pre-college STEM factors, such as advanced math and science courses taken in high school, have a positive relationship with getting a STEM degree while first generation status is associated with a higher drop-out rate.

These studies found that certain demographic information (e.g., age, gender, ethnicity, etc.) can explain whether college students will be able to graduate in time. However, these demographic factors would not be able to accurately predict students' timely graduation because they account for only a small fraction (< 10%) of variance in the data.

2.2 Predicting Student Success in College

Kovačić (2010) investigated the effect of social-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environments (course program and course block) on attrition of students using decision tree. Of these variables, ethnicity, course program and course block were strongly associated with persistence and drop-out of students, and the decision tree model was able to predict 60.5% cases correctly. By using LR, decision tree and artificial NN, Delen (2011) predicted from several variables, such as age, gender, ethnicity, financial status, major, SAT score, GPA and credit hours earned in college, whether undergraduate students will return in sophomore year. All three models showed a weak predictive power, classification accuracy ranging from 63.31% to 68.55%. Bayer et al. (2012) examined whether social behaviors of students can improve performance of machine learning algorithms predicting drop-out and school failure. When they incorporated social dependencies extracted from email and discussion board conversations into the model, accuracy of machine learning algorithms was improved by more than 10%.

When Chen et al. (2018) performed a survival analysis to predict at-risk students who are likely to drop out of college, they found that it performed better than popular machine learning algorithms, such as LR, decision tree and boosting. Nagy and Molontay (2018) developed gradient boosted tree, naïve bayes, KNN, and deep learning models to predict dropout of 15,825 undergraduate students based on data available at the time of matriculation (e.g., academic performance in secondary school, demographic information, etc.). Berens et al. (2019) used an Adaptive Boosting (AdaBoost) algorithm combining LR, NN and decision tree models to predict at-risk undergraduate students who are likely to drop out of college. Their AdaBoost model, which employs features derived from demographic information (e.g., age, gender, immigration background, etc.) and academic performance (e.g., GPA, average semester credit points earned, number of attempted but failed exams, etc.), was able to predict attrition of students with an accuracy of 82% at the end of first semester. Aulck et al. (2019) trained regularized LR, KNN, RF, support vector machine and gradient boosted tree models on registrar records of more than 60,000 undergraduate students to predict their academic success. Their machine learning models were able to predict students' second year enrollment and eventual graduation with sufficiently high accuracy.

Similar to the results of studies focusing on explaining student success, various demographic information, such as age, gender and ethnicity, was found useful in predicting student success. Also, these studies found that aggregated academic performance, such as cumulative GPA, is an important predictor of student success. However, these studies were not able to identify important prerequisite courses having a large effect on student success because academic performance used as a predictor in their model was averaged over all the courses students took in high school or college.

3 Data

In order to answer the research questions, this study developed five predictive models employing LR (Cox, 1958), KNN (Altman, 1992), NN (McCulloch and Pitts, 1943), RF (Breiman, 2001)

Table 1: Description of predictor variables.

Name	Type	Value
Gender	Categorical	{Female, Male}
Ethnicity	Categorical	{African American, American Indian, Asian/Pacific Islander, Hispanic, Non-resident, Other, White}
ACT/SAT	Numeric	Normalized by the highest possible score
Grade	Categorical	{A/B, C/D, F, Not_Taken}

and XGBoost (Chen and Guestrin, 2016) on data obtained from the office of registrar of a large, public research university in the southern US.

The data set includes demographic information of 569 students majoring in biology (age, gender, and ethnicity), their ACT/SAT score, academic plan submitted in the first semester (Fall 2015), academic plan completed at graduation, and letter grades in freshman year courses. Of these 569 students, only 21% of them (N = 118) were able to earn a biology bachelor's degree in four years.

3.1 Data Pre-processing

As this study seeks to identify freshman year courses that have a large impact on student success in completing a bachelor's degree in biology, the course grades undergraduate students received in their first year of study were used as primary predictor variables. When students took the same course more than once, the grade from the first attempt was used. In the data set analyzed in this study, there were 305 unique courses at least one student majoring in biology took in the freshman year. The median enrollment size of courses in the data set was 2, indicating that a large number of courses contain only one or two students majoring in biology.

Because the data set contains a large number of courses, using letter grades as a predictor causes two problems. First, the data set will contain too many missing values because students did not take the same set of courses. To address this issue, the course grade was treated as a categorical variable with four levels (A or B/C or D/F, and Not_Taken), which allows for encoding performance of students without generating missing values. The other problem is that it will create too many predictors with little predictive power because many courses contain only one or two students. In order to avoid creating too many, non-informative predictor variables from course letter grades, the courses associated with less than 28 students, which corresponds to 5% of the total number of students, were merged into a new course called "other." This pre-processing step reduced the number of unique courses in the data set from 305 to 29.

3.2 Description of Predictors

Table 1 lists the variables that were used in predicting student success in getting a biology degree in twelve semesters (including four summer semesters).

Of 569 students matriculated as a biology major in Fall 2015, 363 (64%) of them were female. Table 2 summarizes the proportion of female and male students who were able to get a biology degree by the Fall 2019 semester. Gender is correlated with student success in getting a degree in time; 23% of female students were able to graduate with a degree while only 17% of male students were successful.

Table 2: Graduation vs. gender & ethnicity of students.

Predictor	Value	Graduation	Number of students	Proportion of students who earned the degree
Gender	Female	Yes	82	0.23
	Female	No	281	0.77
	Male	Yes	36	0.17
	Male	No	170	0.83
Ethnicity	African American	Yes	16	0.17
	African American	No	76	0.83
	American Indian	Yes	1	0.14
	American Indian	No	6	0.86
	Asian/Pacific Islander	Yes	21	0.25
	Asian/Pacific Islander	No	64	0.75
	Hispanic	Yes	33	0.20
	Hispanic	No	135	0.80
	Non-resident	Yes	2	0.20
	Non-resident	No	8	0.80
	Other	Yes	0	0.00
	Other	No	3	1.00
	White	Yes	45	0.22
	White	No	159	0.78

White was the largest ethnicity group (N = 204, 36%), followed by Hispanic (N = 168, 30%) and African American (N = 92, 16%). Students in the Asian/Pacific Islander category were most successful in getting a degree in time (graduation rate = 0.25), followed by students in the White category (graduation rate = 0.22). American Indian and African American were two ethnic groups showing a higher rate of drop-out or major-change as summarized in Table 2.

Students who were able to graduate with a biology degree did not have a higher ACT/SAT scores ($M = 0.68$, $SD = 0.09$; t $(180.64) = 1.26$, $p = 0.21$), compared to students who dropped out or switch the major ($M = 0.67$, $SD = 0.09$).

Figure 1 shows the proportion of grades from all courses in the data set analyzed in this study. Students who were able to graduate with a biology degree got more A or B's, compared to students who either dropped out or changed the major. Also, approximately 11.5% of grades given to students who dropped out of the biology program or switched the major were F while only 2.0% of grades were F for students who were able to get the degree in time.

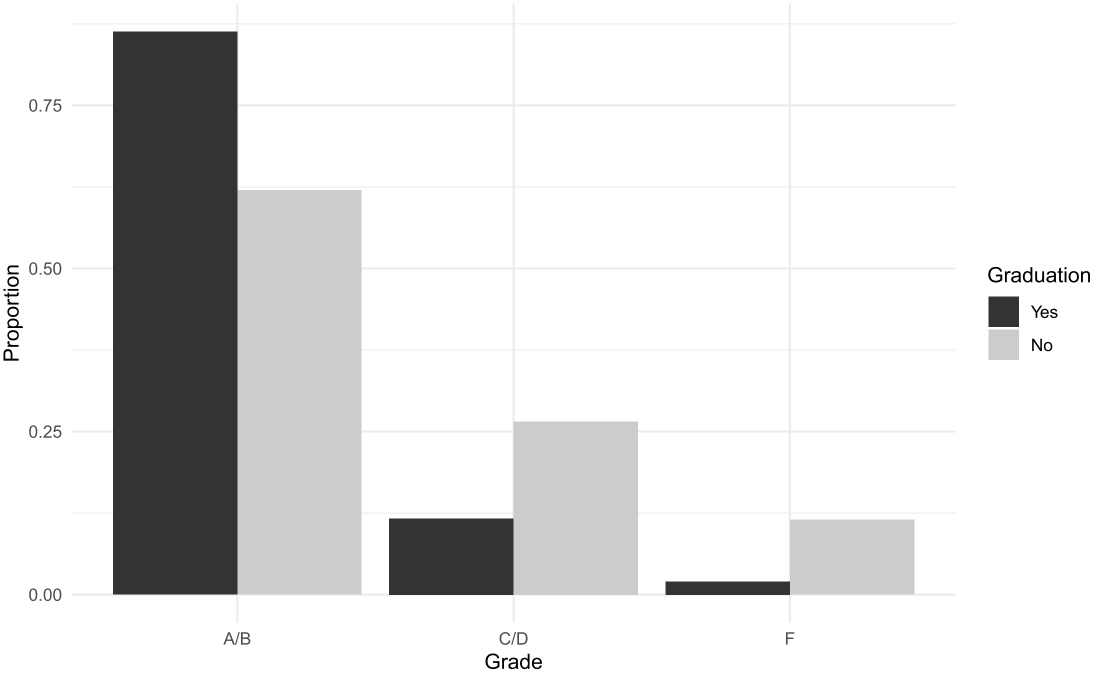

Figure 1: Proportion of grade distribution.

4 Method

In order to predict student success in earning a bachelor's degree in biology, five predictive models based on LR, KNN, NN, RF and XGBoost, were developed using an R package called tidymodels (Khun and Silge, 2022), and performance of these predictive models were compared.

Because using one data set in building and evaluating a predictive model yields overly optimistic results (James et al., 2013), this study used a training/test set approach in which a training set is used to estimate model parameters, and a separate test set is used to evaluate predictive power of models without an opportunistic bias. When creating training and test sets, which consist of 80% and 20% of pre-processed data, a stratified random sampling was used in order to ensure that the ratio of positive (students who were able to earn a biology bachelor's degree in time) and negative instances (students who either dropped out or switched the major) in training and test sets are similar. For KNN, NN, RF and XGBoost models, 10-fold cross validation was used to tune their hyperparameters over a parameter space maximizing entropy (Shewry and Wynn, 1987). Table 3 summarizes the best hyperparameter values of predictive models obtained from the 10-fold cross validation process.

In estimating predictive power of machine learning algorithms, accuracy, sensitivity, specificity, and AUC values computed on the test set were used. Each machine learning algorithm estimated a probability for each student to earn a biology degree, and the student was predicted to be able to graduate with a degree in time if the estimated probability value was greater than a pre-determined cutoff value. Because the data set was highly imbalanced, namely that only 21% of students were able to earn a degree, it was not desirable to use 0.5, which has been frequently used in previous studies, as a cutoff probability value. This study, therefore, used a value minimizing an absolute difference between sensitivity and specificity computed on the training set as a cutoff probability value (Fernández et al., 2018).

Table 3: Tuned hyperparameter values.

Model	Hyperparameter	Value
KNN	neighbors	27
RF	mtry	2
	min_n	33
NN	hidden_units	10
	penalty	0.971
	epochs	831
XGBoost	mtry	19
	min_n	4
	tree_depth	8
	learn_rate	9.37×10^{-7}
	loss_reduction	6.11×10^{-10}
	sample_size	0.532

The shortcoming of using accuracy, sensitivity or specificity as performance measurement is that these metrics can change quite significantly when a different cutoff threshold value is used (Kleinbaum and Klein, 2010). In order to address this issue, this study also used AUC in evaluating performance of machine learning models because AUC is computed on a large number of different cutoff threshold values. AUC can vary from 0.5 (predictive power not better than simple guessing) to 1.0 (perfect predictive model), and is known to be equal to a probability that a binary classification model will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006).

When computing importance of predictors, this study employed a method called "permutation feature importance" (Breiman, 2001). In this method, importance of each predictor is assessed by the decrease in the model's predictive power when values of the predictor under consideration is randomly shuffled; a predictor is considered important if shuffling its values significantly decreases the model's predictive power.

Also, in order to explain the model's prediction at the student level, this study adopted an explainable machine learning approach called break-down. The break-down method captures one variable's contribution to model prediction on a particular student by computing a shift in the expected probability of observing an event of interest, graduating with a biology degree in time, while fixing values of other predictors (Biecek and Burzykowski, 2021). As a result, the break-down plot decomposes model prediction into contributions attributed to different predictors in the model.

5 Results

5.1 Predicting Student Success in Earning Bachelor's Degree in Biology

Table 4 lists the probability cutoff values derived from the training set, and the performance metric values (accuracy, sensitivity, specificity, and AUC) computed on the test set. Overall, LR did not perform as well as the other models because it is unable to capture non-linear

Table 4: Cutoff probability and metric values of predictive models.

	LR	KNN	NN	RF	XGBoost
Cutoff probability	0.28	0.38	0.44	0.28	0.50
Accuracy	0.68	0.71	0.71	**0.73**	0.71
Sensitivity	0.60	0.57	0.57	0.63	**0.67**
Specificity	0.70	0.74	0.74	**0.76**	0.73
AUC	0.62	0.71	0.70	0.77	**0.78**

relationships between predictor and outcome variables. RF showed the best performance in terms of accuracy and specificity while XGBoost performed best in terms of sensitivity and AUC. Considering the fact that it would be more important for an institution of higher education to identify students who are likely to drop out of the degree program or change their major, RF could be considered a more meaningful model than XGBoost because it showed a better performance on specificity. Another reason for preferring a RF model over XGBoost is that it allows us to use a letter grade as a predictor without dummy coding; since XGBoost cannot handle a categorical predictor, letter grades from each course had to be “dummy-coded”, which makes it less convenient to assess importance of each course in the degree program.

5.2 Identifying More Important Prerequisite Courses

More important prerequisite courses can be identified at the biology bachelor’s degree program level and at the student level. To find more important courses at the degree program level, this study estimated importance of each variable in predicting student success at the predictive model level. To find more important courses at the student level, this study employed a break-down method (Biecek and Burzykowski, 2021) as explained above. Since the RF model showed the highest performance in terms of accuracy and specificity while showing a comparable performance on AUC, the following subsections describe more importance prerequisite courses identified by the RF model.

5.2.1 More Important Courses at Program Level

Figure 2 shows the variable importance of RF model measured with a permutation importance method (Breiman, 2001) in which an average improvement of AUC values was used as importance value of each predictor. The RF model indicates that Chemistry 1420 and Biology 1720 are the two most important freshman year courses that have a large effect on student success, followed by Chemistry 1410, Chemistry 1440 and Biology 1710.

As summarized in the Related Work section, several previous studies found that demographic information (e.g., gender, ethnicity) and academic performance in high school were important predictors for academic success of college students (Bettencourt et al., 2020; Delen, 2011; Kovačić, 2010; Smith et al., 2012). However, the RF model developed in this study found that gender and ACT/SAT score are not as important as academic performance in the freshman year courses in predicting timely graduation of undergraduate students.

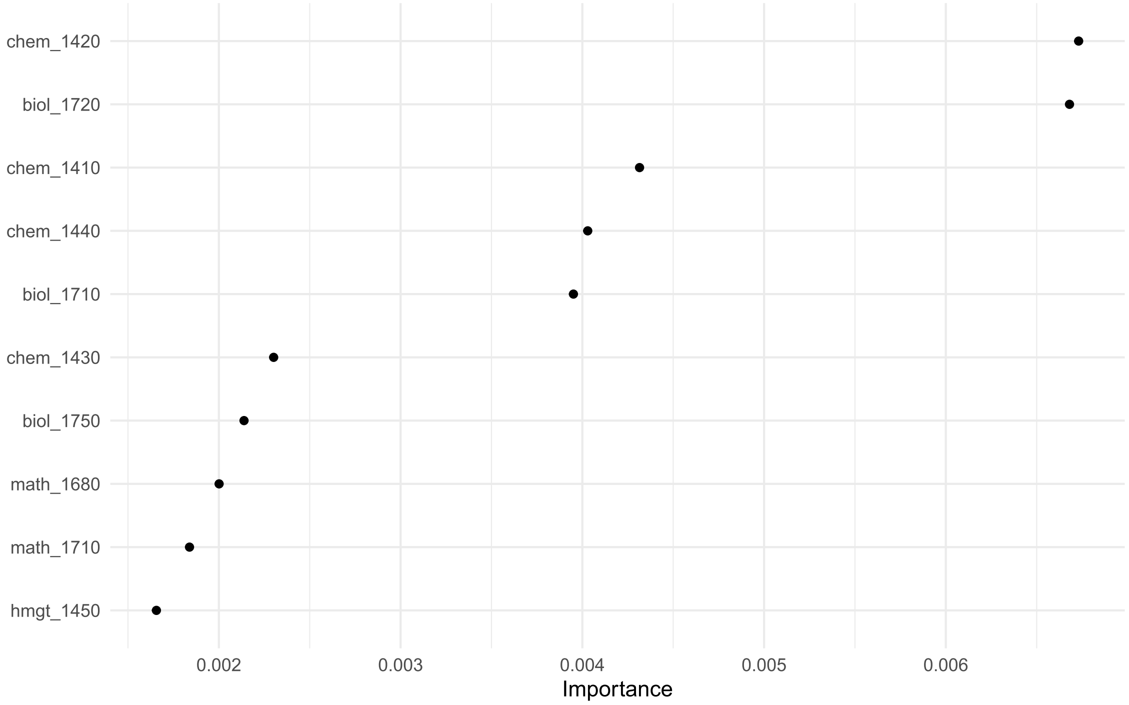

Figure 2: Variable importance of RF model.

5.2.2 More Important Courses at Student Level

In order to identify courses that have a large effect at the student level, it is crucial to understand how a predictive model makes a decision on each student. Compared to linear models such as LR, fully non-linear models are based on highly flexible approaches that are much more difficult to interpret (James et al., 2013). In order to better understand how KNN, NN, RF and XGBoost models make a prediction on a specific instance (i.e., individual student), this study employed a break-down method.

The break-down method is a model-agnostic approach which decomposes the model's prediction into separate contributions from each predictor, allowing for visualizing which predictor is more important when making a decision on a particular instance (Biecek and Burzykowski, 2021). Figure 3 shows a break-down plot of a student who failed to earn a biology degree in time. The dashed lines mark the cutoff probability value estimated from the training data. Because this student got an A/B in Chemistry 1440 (chem_1440 = 1), the fourth important predictor at the model level, the probability of graduation increased by 0.026. However, the probability of success decreased because this student received an F on Chemistry 1420 (chem_1420 = 3), the most important predictor at the model level, and did not take Biology 1720 (biol_1720 = 4), the second important predictor at the model level. Although this student received A's on Mathematics 1650 (math_1650 = 1), Chemistry 1410 (chem_1410 = 1) and Political Science 1040 (psci_1040 = 1), which improved the likelihood of success, the RF model correctly predicted that this student would not be able to graduate in time because those courses are not as important as Chemistry 1420 and Biology 1720.

Figure 4 shows a break-down plot of a successful student. When this student did not take Chemistry 1420 (chem_1420 = 4) and Chemistry 1440 (chem_1440), the probability of success

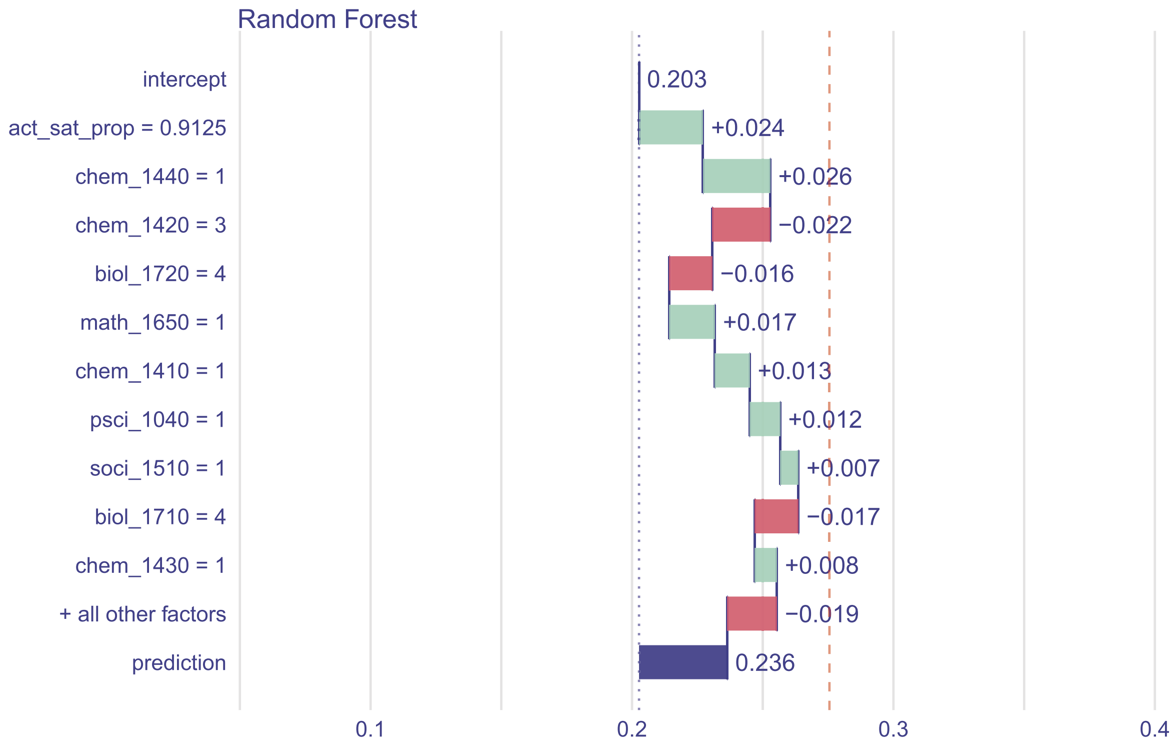

Figure 3: Break-down plot of unsuccessful student.

decreased by 0.017 and 0.015, respectively. However, based on good performance on other courses (e.g., Biology 1720 (biol_1720 = 1) and Biology 1710 (biol_1710 = 1)) and gender (1: Female), the RF model correctly predicted that this student would be able to earn a biology bachelor's degree in time. Note that even though gender and ethnicity were not important predictors at the model level, the break-down method indicates that these factors have some effect on this particular student.

6 Discussion

This study applied five data mining algorithms, LR, KNN, NN, RF, and XGBoost, to information available in registrar records (gender, ethnicity, ACT/SAT score, and letter grades from freshman year courses) to predict whether students will be able to earn a biology bachelor's degree in time. Of these five algorithms, RF and XGBoost showed a better performance than the other algorithms. Although RF and XGBoost showed a similar performance, RF could be considered more appropriate because it was better at identifying students who are likely to drop out of the degree program or change their major, which could allow an institution of higher education to provide proactive advice to those students. Also, this study was able to identify several freshman year courses that have a larger impact on student success at the biology bachelor's degree program level by calculating variable importance values at the predictive model level. Finally, this study demonstrated that more important courses for a particular student can be identified by breaking down the model's prediction into separate contributions from each predictor in the model. Practitioners who are not familiar with machine learning prefer a simple linear model such

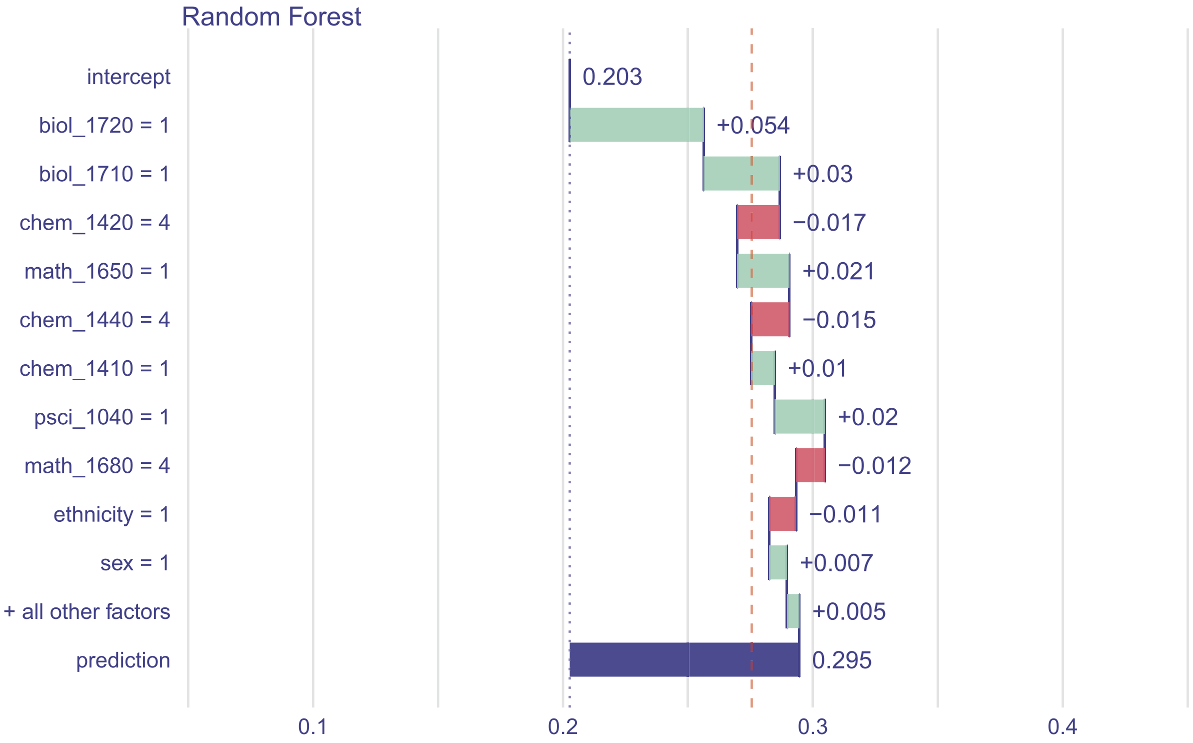

Figure 4: Break-down plot of successful student.

as LR because it is easy to interpret (James et al., 2013). This study showed that an explainable machine learning approach, such as break-down, can help interpret how fully non-linear models make predictions, allowing practitioners to use more powerful non-linear models.

It is important to recognize that this is an observational study that analyzed a secondary data set. Therefore, it is possible that there are important confounding variables and factors, such as pre-existing student ability or parental background, not available in the data set analyzed in the study. Consequently, when interpreting the model's prediction, it is crucial to acknowledge that importance of courses could change when additional predictors are added to the model in the future analysis.

Researchers have been emphasizing importance of so-called gateway courses, the courses college students take in the first two years in college, in improving retention of students in STEM. Previous studies found that gateway courses can become an obstacle in getting a STEM degree in time (Alexander et al., 2009; Gasiewski et al., 2012; Malcom and Feder, 2016; Suresh, 2007). Computing variable importance at the predictive model level can be used as a data-driven approach for quantifying importance of gateway courses. This approach can enable an institution of higher education to operate more efficiently because it can allow administrators to allocate limited resources to more important gateway courses first.

In order to support student success and improve student retention, many institutions of higher education provide advising service to students (Schwebel et al., 2012). According to Kuhn et al. (2006), course selection is one of the most frequently discussed topics in what they called prescriptive advising. More important courses identified at the student level can enable an institution of higher education to provide more personalized advising. For example, with a predictive model of student success developed on the freshman and sophomore year data from previous years, academic advisors can provide guidance highly customized to rising sophomore students. The break-down plot can help academic advisors and students to better understand the importance of each course, making their conversation much more useful and meaningful.

It is important to recognize that the role of instructors and advisors becomes even more crucial when advanced learning technologies, such as machine learning, are used in education. For instance, when advisors are recommending courses, they must consider first the need and interest of students, instead of simply following recommendations from a predictive model. Similarly, predictive model of student success should not be used as a punitive tool weeding out students who may drop out or switch the major. Instead, it should be used as a means to providing personalized support and guidance to help students be more successful in their academic endeavor.

This study has several limitations stemming from sample size. This study estimated a cutoff probability value for making a prediction on student success (i.e., graduate with a biology degree in time) from a training set that was used to fit the prediction model. Ideally, a cutoff probability value should be estimated from a separate validation set, which was not possible due to modest sample size. Also, this study did not distinguish students who dropped out from students who switched their major. Of 455 students who did not earn a biology degree in time, 174 of them (38%) graduated with a non-biology degree. It would be meaningful to develop a multi-class classification model, which requires a larger data set, predicting three outcomes (i.e., graduate with a biology degree, drop out, and graduate with a non-biology degree).

Another area of improvement is in estimating importance of variables. Although the permutation feature importance method has been widely used in machine learning research, it was found that that permutation feature importance scores do not always agree with those based on traditional methods, and are biased toward certain variable types (Ctrobl et al., 2007). It would be desirable to use a variable importance measure that has a bias-correction step such as GUIDE (Loh and Zhou, 2021).

7 Conclusion

As digital technologies become an integral part of our society, big data is emerging as a new way of transforming education by allowing for data-driven decisions. According to National Academy of Education (2017), “in the educational context, big data typically take the form of administrative data and learning process data, with each offering their own promise for educational research” (p. 4). Using five machine learning algorithms, this study developed predictive models of student success in earning a bachelor’s degree in biology. These models were able to identify more important courses having a larger impact on student success based on administrative data readily available in institutions of higher education. The findings from this study suggest that predictive modeling can be used for improving student persistence and retention by allowing an institution of higher education to prioritize their effort for improving curriculum and by enabling academic advisors to provide more personalized guidance to students.

Supplementary Material

This includes the data file containing the training and test sets analyzed in the study, and all R code used in the analysis along with an explanatory README.txt file.

References

Alexander C, Chen E, Grumbach K (2009). How leaky is the health career pipeline? Minority student achievement in college gateway courses. *Academic Medicine*, 84(6): 797–802.

Altman NS (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician*, 46(3): 175–185.

Aulck L, Nambi D, Velagapudi N, Blumenstock J, West J (2019). Mining university registrar records to predict first-year undergraduate attrition. In: *Proceedings of the 12th International Conference on Educational Data Mining* (CF Lynch, A Merceron, M Desmarais, R Nkambou, eds.), 9–18. International Educational Data Mining Society.

Ausubel DP (1963). *The Psychology of Meaningful Verbal Learning.* Grune & Stratton, New York, NY.

Bayer J, Bydzovská H, Géryk J, Obsivac T, Popelinsky L (2012). Predicting drop-out from social behavior of students. In: *Proceedings of the 5th International Conference on Educational Data Mining* (K Yacef, O Zaïane, A Hershkovitz, M Yudelson, J Stamper, eds.), 103–109. International Educational Data Mining Society.

Berens J, Schneider K, Görtz S, Oster S, Burghoff J (2019). Early detection of students at risk – predicting student dropouts using administrative student data and machine learning methods. *Journal of Educational Data Mining*, 11(3): 1–41.

Bettencourt GM, Manly CA, Kimball E, Wells RS (2020). STEM degree completion and first-generation college students: A cumulative disadvantage approach to the outcomes gap. *Review of Higher Education*, 43(3): 753–779.

Biecek P, Burzykowski T (2021). *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models.* Chapman & Hall/CRC, Boca Raton, FL.

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

Bruner J (1974). *Toward a Theory of Instruction.* Belknap Press, Cambridge, MA.

Chen T, Guestrin C (2016). Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (B Krishnapuram, M Shah, AJ Smola, C Aggarwal, D Shen, R Rastogi, eds.), 785–794. Association for Computing Machinery.

Chen X, Ho P (2012). STEM in postsecondary education: Entrance, attrition, and coursetaking among 2003–2004 beginning postsecondary students. *NCES Report no. 2013-152.*

Chen X, Soldner M (2013). STEM attrition: College students' paths into and out of STEM fields. *NCES Report 2014-001.*

Chen X, Weko T (2009). Students who study science, technology, engineering, and mathematics (STEM) in postsecondary education. *NCES Report no. 2009-161.*

Chen Y, Johri A, Rangwala H (2018). Running out of STEM: A comparative study across STEM majors of college students at-risk of dropping out early. In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (A Pardo, K Bartimote, G Lynch, S Buckingham Shum, R Ferguson, A Merceron, X Ochoa, eds.), 270–279. Society for Learning Analytics Research.

Cochran JD, Campbell SM, Baker HM, Leeds EM (2013). The role of student characteristics in predicting retention in online courses. *Research in Higher Education*, 55(1): 27–48.

Cox DR (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B, Methodological*, 20(2): 215–232.

Cromley JG, Perez T, Kaplan A (2015). Undergraduate stem achievement and retention: Cognitive, motivational, and institutional factors and solutions. *Policy Insights From the Behavioral and Brain Sciences*, 3(1): 4–11.

Ctrobl C, Boulesteix A, Zeileis A, Hothorn T (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8: 25.

Dai T, Cromley JG (2014). Changes in implicit theories of ability in biology and dropout from stem majors: A latent growth curve approach. *Contemporary Educational Psychology*, 39(3): 233–247.

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

Ehrenberg RG (2010). Analyzing the factors that influence persistence rates in STEM field, majors: Introduction to the symposium. *Economics of Education Review*, 29: 888–891.

Fawcett T (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874.

Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018). *Learning from Imbalanced Data Sets*. Springer, Cham, Switzerland.

Gagné RM, Briggs LJ (1974). *Principles of Instructional Design*. Holt, Rinehart & Winston, New York, NY.

Gasiewski JA, Eagan MK, Garcia GA, Hurtado S, Chang M (2012). From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory stem courses. *Research in Higher Education*, 53(2): 229–261.

James G, Witten D, Hastie T, Tibshirani R (2013). *An Introduction to Statistical Learning: With Application in R*. Springer, New York, NY.

Khun M, Silge J (2022). *Tidy Modeling with R: A Framework for Modeling in the Tidyverse*. O'reilly, Sebastopol, CA.

Kleinbaum DG, Klein M (2010). *Logistic Regression: A Self-Learning Text*. Springer, New York, NY.

Kovačić ZJ (2010). Early prediction of student success: Mining students enrolment data. In: *Proceedings of Informing Science IT Education Conference* (E Cohen, ed.), 647–665. Informing Science Institute.

Kuhn TK, Gordon VN, Webber J (2006). The advising and counseling continuum: Triggers for referral. *NACADA Journal*, 26: 24–31.

Le H, Robbins SB, Westrick P (2014). Predicting student enrollment and persistence in college stem fields using an expanded P-E fit framework: A large-scale multilevel study. *Journal of Applied Psychology*, 99(5): 915–947.

Lee YG, Ferrare JJ (2019). Finding one's place or losing the race? The consequences of stem departure for college dropout and degree completion. *Review of Higher Education*, 43(1): 221–261.

Loh WY, Zhou P (2021). Variable importance scores. *Journal of Data Science*, 19(4): 569–592.

Malcom S, Feder M (2016). *Barriers and Opportunities for 2-year and 4-year STEM Degree: Systematic Change to Support Students' Diverse Pathways*. The National Academies Press, Washington, DC.

McCulloch WS, Pitts W (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4): 115–133.

Nagy M, Molontay R (2018). Predicting dropout in higher education based on secondary school performance. In: *Proceedings of the 22nd IEEE International Conference on Intelligent Engi-*

neering Systems, 389–394. Institute of Electrical and Electronics Engineers. Paper is available in IEEE Xplore: https://ieeexplore.ieee.org/abstract/document/8523888.

National Academy of Education (2017). *Big data in education: Balancing the benefits of educational research and student privacy: A workshop summary.* National Academy of Education, Washington, DC.

National Science Board (2018). Science & engineering indicators 2018. *NSB Report 2018-1.*

Olson S, Riordan DG (2012). Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. *Report to the President.* Washington, DC.

Patrick AD, Prybutok AN, Borrego M (2021). Predicting persistence in engineering through an engineering identity scale. *International Journal of Engineering Education*, 34(2A): 351–363.

Reigeluth CM (1979). In search of a better way to organize instruction: The elaboration theory. *Journal of Instructional Development*, 2(3): 8–15.

Reigeluth CM, Merrill MD, Bunderson CV (1978). The structure of subject matter content and its instructional design implications. *Instructional Science*, 7: 107–126.

Schwebel D, Walburn N, Klyce K, Jerrolds K (2012). Efficacy of advising outreach on student retention, academic progress and achievement, and frequency of advising contacts: A longitudinal randomized trial. *NACADA Journal*, 32: 36–43.

Shewry MC, Wynn HP (1987). Maximum entropy sampling. *Journal of Applied Statistics*, 14(2): 165–170.

Smith M, Therry L, Whale J (2012). Developing a model for identifying students at risk of failure in a first year accounting unit. *Higher Education Studies*, 2(4): 91–102.

Sullivan JF (2006). Broadening engineering's participation-a call for K-16 engineering education. *The Bridge*, 36(2): 17–24.

Suresh R (2007). The relationship between barrier courses and persistence in engineering. *Journal of College Student Retention*, 8(2): 215–239.

Thompson R, Bolin G (2011). Indicators of success in stem majors: A cohort study. *Journal of College Admission*, 212: 18–24.

Xie Y, Killewald AA (2012). *Is American Science in Decline?.* Harvard University Press, Cambridge, MA.