

6-11-2018

Combining Academics and Social Engagement: A Major-Specific Early Alert Method to Counter Student Attrition in Science, Technology, Engineering, and Mathematics

Andrew J. Sage

Iowa State University, ajsage@iastate.edu

Cinzia Cervato

Iowa State University, cinzia@iastate.edu

Ulrike Genschel

Iowa State University, ulrike@iastate.edu

Craig Ogilvie

Follow this and additional works at: https://lib.dr.iastate.edu/ge_at_pubs

Iowa State University, cogilvie@iastate.edu

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Engineering Education Commons](#), [Higher Education Commons](#), [Science and Mathematics Education Commons](#), and the [Statistical Methodology Commons](#)

The complete bibliographic information for this item can be found at https://lib.dr.iastate.edu/ge_at_pubs/255. For information on how to cite this item, please visit <http://lib.dr.iastate.edu/howtocite.html>.

This Article is brought to you for free and open access by the Geological and Atmospheric Sciences at Iowa State University Digital Repository. It has been accepted for inclusion in Geological and Atmospheric Sciences Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Combining Academics and Social Engagement: A Major-Specific Early Alert Method to Counter Student Attrition in Science, Technology, Engineering, and Mathematics

Abstract

Students are most likely to leave science, technology, engineering, and mathematics (STEM) majors during their first year of college. We developed an analytic approach using random forests to identify at-risk students. This method is deployable midway through the first semester and accounts for academic preparation, early engagement in university life, and performance on midterm exams. By accounting for cognitive and noncognitive factors, our method achieves stronger predictive performance than would be possible using cognitive or noncognitive factors alone. We show that it is more difficult to predict whether students will leave STEM than whether they will leave the institution. More factors contribute to STEM retention than to institutional retention. Early academic performance is the strongest predictor of STEM and institution retention. Social engagement is more predictive of institutional retention, while standardized test scores, goals, and interests are stronger predictors of STEM retention. Our approach assists universities to efficiently identify at-risk students and boost STEM retention.

Keywords

analytics, early alert, science, technology, engineering, and mathematics retention

Disciplines

Educational Assessment, Evaluation, and Research | Engineering Education | Higher Education | Science and Mathematics Education | Statistical Methodology

Comments

This is a manuscript of an article published as Sage, Andrew J., Cinzia Cervato, Ulrike Genschel, and Craig A. Ogilvie. "Combining Academics and Social Engagement: A Major-Specific Early Alert Method to Counter Student Attrition in Science, Technology, Engineering, and Mathematics." *Journal of College Student Retention: Research, Theory & Practice* (2018). doi:

Combining academics and social engagement: a major-specific early alert method to counter student attrition in STEM

Andrew J. Sage^{1*}
Cinzia Cervato²
Ulrike Genschel¹
Craig A. Ogilvie³

¹Department of Statistics, Iowa State University, Ames IA, 50011.

²Department of Geological & Atmospheric Sciences, Iowa State University, Ames IA, 50011.

³Department of Physics, Iowa State University, Ames IA, 50011.

*Correspondence to: ajsage@iastate.edu

We would like to acknowledge Dan Nettleton for his suggestions on random forest methodology and Jonathan Compton and Aurelia Kollasch for their efforts in processing and providing the student data. We are grateful for the suggestions of two anonymous reviewers, which helped strengthen the paper. This work was supported by a grant to Iowa State University from the Howard Hughes Medical Institute through the Science Education Program.

Abstract

Students are most likely to leave STEM majors during their first year of college. We developed an analytic approach using random forests to identify at-risk students. This method is deployable midway through the first semester and accounts for academic preparation, early engagement in university life, and performance on midterm exams. By accounting for cognitive and noncognitive factors our method achieves stronger predictive performance than would be possible using cognitive or noncognitive factors alone. We show that it is more difficult to predict whether students will leave STEM than whether they will leave the institution. More factors contribute to STEM retention than to institutional retention. Early academic performance is the strongest predictor of STEM and institution retention. Social engagement is more predictive of institutional retention, while standardized test scores, goals, and interests are stronger predictors of STEM retention. Our approach assists universities to efficiently identify at-risk students and boost STEM retention.

Keywords: analytics, early alert, STEM retention

Introduction

Retaining talented and diverse students in science, technology, engineering, and mathematics (STEM) majors is vital to building the human capital required to solve society's largest, and most complex problems. However, attrition rates are high for many STEM disciplines, and research has shown that the dropout rate is highest during the first year in college (Chen, 2014). Successful transition from high school to college is critical, and interventions that increase student engagement or involve students in learning communities can be effective at increasing students' confidence in their ability to succeed in STEM (Wilson et al., 2012). Identifying at-risk students is especially important for universities with large enrollment so that they might offer additional support and resources to these students beyond regular advising.

Tinto's student retention model (Tinto, 1975) characterizes factors that influence a student's decision to stay in or leave college. Examples include contextual factors such as current learning environment and experiences, academic factors including past performance, but also noncognitive factors such as self-efficacy, interests, or goals. In this study, we focus on the impact of Tinto's factors on the retention of STEM students. We consider two forms of STEM attrition; 1) STEM students leaving a STEM major, while remaining at the institution, and 2) STEM students leaving an institution altogether.

Many studies have validated different parts of Tinto's model, applied to STEM majors. For example, English and mathematics ACT subtests have been shown to be related to STEM retention (Bettinger, Evans, & Pope, 2013). Noncognitive factors, such as self-efficacy, also play a role in the decision to leave (Chemers, Hu, & Garcia, 2001) even though students' perception of self-doubt and performance avoidance may not be correlated with objective measures of academic performance (Shedlosky-Shoemaker & Fautch, 2015). For current learning

experiences, student dissatisfaction with instructional differences between high school and college faculty has been shown to be higher in STEM than in other majors, contributing to attrition (Daempfle, 2003). Further, the above factors often contribute to low grades in first STEM courses, which in turn are strong predictors of STEM retention (Ehrenberg, 2010; Ost, 2010; Rask, 2010; Aulck, Aras, Li, l'Heureux, Lu, & West, 2017).

In this work, we describe an analytic approach for identifying at-risk STEM students early in their first semester. We use random forest methodology (Breiman, 2001) to assess the predictive value of all explanatory variables. Random forests are a popular machine learning approach, capable of strong predictive performance on high-dimensional data. The use of random forests in the education literature is limited (Mendez, Buskirk, Lohr, & Haag, 2008; Aulck et al., 2017). Although Aulck et al. (2017) found random forests less effective than logistic regression at predicting STEM retention, our results suggest that random forests grown using conditional inference trees (Hothorn, Hornik, & Zeileis, 2006) perform on par with logistic regression. An advantage of random forest methodology is the ability to measure variable importance for complex datasets (Breiman, 2003; Janitza, Strobl, & Boulesteix, 2006; Strobl, Boulesteix, Zeileis, & Hothorn 2007; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008; Hapfelmeier, Hothorn, Ulm, & Strobl, 2014).

Our method combines information from a wide range of factors that serve as proxies for the factors in Tinto's model. We apply this method to estimate the probability of first-year STEM students at a large public research university in the Midwest either switching to a non-STEM major or leaving the institution altogether prior to their second year, allowing advisors to offer interventions aimed at retaining STEM students. Throughout the paper, we use the phrase *institutional retention* to refer to a student staying at the university through the start of the second

year, and the phrase *STEM retention* to refer to a student staying in STEM through the start of the second year, conditional upon the student staying at the university.

We made progress on two challenges: 1) the need for a method taking into account both cognitive and noncognitive factors to predict students at risk, i.e. utilizing academic information along with measures of social engagement, self-efficacy, and student interests and goals, 2) the flexibility for the prediction to account for type of STEM major, gender, and ethnicity to further improve prediction.

Data and Methodology

We have data for 7,459 freshmen STEM students who enrolled at the university for the Fall 2014 and Fall 2015 semesters. Of these, 3,667 enrolled in Fall 2014 and 3,792 enrolled in Fall 2015. Transfer students are not included since the factors contributing to their retention are likely to differ from the factors influencing retention among students enrolling directly from high school. Of these students, 64.6% are men, 95% are U.S. citizens, 75.3% are white (non-Hispanic), 5.1% Hispanic, 3.4% Asian, and 2.5% black (non-Hispanic) American students. The remaining U.S. students are either of other ethnicities or did not provide this information. Ethnicities of international students are not provided. Approximately 12% of students left the institution prior to the start of their second year. An additional 9% switched to a non-STEM major during that time. These departure rates vary by major (Table 1).

Table 1:

Summary of Retention by Major Type

Major Type	First-year STEM students	Students leaving university (%)	Students leaving STEM (%)
Biological Sciences	1,543	185 (12.0%)	107 (6.9%)

Engineering	4,100	399 (9.7%)	309 (7.5%)
Human & Health Sciences	1,171	193 (16.5%)	192 (16.4%)
Mathematical Sciences	405	61 (15.1%)	60 (14.8%)
Physical Sciences	240	37 (15.4%)	25 (10.4%)
Total	7,459	875 (11.7%)	693 (9.3%)

Our approach uses 38 variables that serve as proxies for the factors in Tinto’s retention model and are available in the first eight weeks of a student’s initial semester. Personal inputs include high school academic information, demographics, and standardized test scores. A student’s learning experiences are described through first-semester course enrollment, midterm grades, and learning community participation. Responses to the MAP-Works® survey (now Skyfactor®) (Woosley & Jones, 2011), completed by freshmen in the first 6 weeks of the fall semester, provide insight into a student’s social engagement, proximal environment, and self-efficacy, while the ACT Interest Inventory (Swaney, Lamb, & Prediger, 1995) survey provides information on students’ interests and goals in their junior or senior year of high school. Considering a broad range of cognitive and noncognitive factors potentially allows universities to identify at-risk students who may not be detected by frequently used cognitive measures such as midterm grades alone. Table 2 gives the list of variables we considered, which are grouped into fifteen categories of variables similar to one-another.

Table 2:

Variables Considered. Abbreviations are provided for variables referenced later.

Group	Variables Included
Academic Skills and Behaviors	MAP-works factor score
ACT Interest Survey	Questions on whether student intended to major in STEM (STEM-INTD) and sureness of major (MAJ.SURE).
College Activities	Athletics, Greek life
Demographics	ethnicity, gender, US citizenship, in-state residency
Financial Concerns	MAP-works factor score
First Semester Classes	Course enrollments in: biology, calculus, chemistry, physics, remedial math, academic study skills course
First Semester Midterm Grades	Number of points (C=1, D=2, F=3) for midterm grades in STEM classes (STEM-MT) and all classes
High School Courses and Grades	GPA (HS GPA), Rank, semesters of math (SEM.MTH), science units, social science units
Learning Community Participation	Is student a member of a learning community? (LC)
Math/Science Self-Efficacy	MAP-works factor score

Satisfaction with Major	Response to single MAP-works question on likelihood of changing major (MAJ.CHG.)
Satisfaction with University	MAP-works factor score
Scholarships	University awarded scholarships
Social Integration	MAP-works factor score
Standardized Tests	ACT composite, ACT English, ACT math (ACT-M), ACT reading, ALEKS math placement

Because many questions from the MAP-Works® survey are similar to one another, a factor analysis was first conducted to group together responses to highly correlated questions. Using the promax method of rotation, responses were aligned to five different factors, namely academic skills, financial concerns, self-efficacy, satisfaction with the university, and social integration. Students' factor scores for each of these were then used as predictor variables for retention. A MAP-Works® question about the student's likelihood of changing majors did not align to any of these factors, but was also included as a predictor variable on its own due to its potentially strong predictive ability. Other MAP-Works® questions that did not align to factors and showed little predictive value were not considered. Because students who completed the MAP-Works® survey occasionally skipped questions, missing values were imputed using the Multivariate Imputation by Chained Equations (MICE) algorithm (Van Buuren & Groothuis-Oudshoorn, 2011), prior to conducting the factor analysis. This was only done in situations where students completed the survey, but skipped particular questions. MAP-Works® scores for students who did not take the survey were left as missing.

Random forests (Breiman, 2001) were grown using conditional inference trees, available in the *party* package (Hothorn, Buehlmann, Dudoit, Molinaro, & Van Der Lann, 2006; Strobl et al., 2007; Strobl et al., 2008) in R (R Development Core Team, 2013). This approach has been shown to yield unbiased measures of variable importance for datasets containing a mix of numeric and categorical predictors (Strobl et al., 2008). Cross-validation showed that optimal prediction performance occurs when the *minsplit* parameter, which determines maximal terminal node size, is set to 75. We therefore used this setting, although we found variable importance measures to be largely insensitive to changes in this parameter. Breiman (2003) suggests growing large forests, consisting of many decision trees, especially when variable importance is of interest. Smaller forests can lead to unstable results, due to the randomness associated with the process of growing trees. Therefore, we used random forests consisting of 1,000 trees, which we found led to highly stable variable importance results.

Our situation is an example of an unbalanced classification problem, as students stay in STEM far more frequently than they leave. In such problems, receiver operating characteristic (ROC) curves have been shown to be useful in assessing variable importance. Janitza et al. (2006) showed that a variable importance measure based on area under the curve (AUC) is preferable to one based on misclassification rate, which is typically used in classification problems. The ROC curve traces the rates of true positives (i.e. at-risk students who left) and false positives (i.e. at-risk students who stayed) as the at-risk threshold varies. AUC provides a measure of prediction quality that is independent of this threshold. A method capable of perfectly separating students who leave from those who stay would achieve an AUC of 1, while a method predicting completely at random is expected to achieve an AUC of 0.5. The fact that AUC does

not depend on the at-risk threshold is advantageous since the appropriate threshold will vary based on the needs and resources of institutions and departments.

Because of the complex interrelationships between many of the variables, we grouped explanatory variables into 15 sets of closely related predictors. For example, the set of variables related to high school academic information includes high school GPA, class rank, and the number of math, science, and social science units. Students were divided into five different groups of STEM majors: biological sciences (BIO), engineering (ENGR), human and health sciences (HH), mathematical sciences (MTH), and physical sciences (PHY). A complete list of the majors classified within each category is available upon request.

Rather than measuring the importance of predictor variables individually, we permuted the values of all variables within a set of closely related predictors at the same time. This provides a measure of the importance of each set of predictor variables. We calculated variable importance across all students, and also within each type of STEM major. When analyzing STEM retention, rather than institutional retention, only the 6,584 students who remained enrolled at the university at the start of their second year were considered.

The algorithm below describes the procedure used to calculate variable importance for sets of predictor variables.

1. Using a simple random sample consisting of 63% of students in the dataset, grow a tree using the *ctree* command in the *party* package with the settings described above.
2. For each student not in the set used to grow the tree, estimate the probability of that student leaving STEM.
3. For each type of STEM major, calculate AUC using the estimates from step 2.

4. For each set of predictor variables, one at a time, randomly permute the values of all variables in a given set.
5. Repeat steps 2 and 3 using the permuted values.
6. For each set of variables, calculate the change in AUC when permuted values are used in place of the actual values.
7. Repeat steps 1-6 to create 1,000 trees and average the change in AUC corresponding to permutation of each group of variables. These represent variable importance scores for each set of predictor variables.

Results

First, we used random forests grown from data on all 2014 first-year STEM students, to estimate the probability of each 2015 first-year STEM student leaving either STEM or the institution during their first year. Establishing a criterion for designating a student as at-risk requires consideration of the benefits of correctly identifying those students who will leave, against the costs of falsely designating students as at-risk when they will actually stay. Furthermore, designating too many at-risk students might make it difficult for advisors to provide helpful interventions for those most in need. We provide alerts to advisors of the 15% of students with the highest estimated probabilities of leaving the institution or STEM. Table 3 provides a breakdown of the number of students designated as at-risk and the actual retention results. We see that random forests enable us to identify a subpopulation of 567 students who go on to leave STEM at a rate more than three times higher than other students.

Table 3:

Results when 15% of students are classified as at-risk.

	Left STEM or Institution	Stayed in STEM at Institution	Total	Percentage Leaving
Classified At-Risk	293	274	567	51.7%
Not Classified At-Risk	505	2720	3225	15.7%
Total	798	2994	3792	21.0%

Although we believe that the “highest 15%” threshold provides a reasonable criterion for designating at-risk students, taking into account the cost/benefit analysis and advising resources available at the institution, individual institutions or departments may wish to establish their own thresholds. Designating a higher number of students as at risk would increase the number of students correctly identified who go on to leave, but also increase the false positive rate. For example, if we had designated 20% of students as at risk, we would have correctly identified an additional 72 students who went on to leave, but would have also identified 119 students who stayed, lowering the percentage of the at-risk group who actually left to 48.2%. Because it is independent of threshold, AUC provides an objective way to measure the impact of academic and social factors on STEM retention, as well as institutional retention of STEM students.

Figure 1 displays ROC curves for predicting STEM retention for 2015 first-year students, using random forests trained on retention data from 2014 first-year students. When cognitive and noncognitive factors were used to predict STEM retention, the random forest approach achieved an AUC of 0.7326, compared to 0.7154 when only cognitive factors are used, and 0.6637 for only noncognitive factors. Thus, when both cognitive and noncognitive information is taken into account, we achieve an AUC that is 0.2326 higher than what would be expected of predictions made purely at random, compared to differences of 0.2154 and 0.1637 when only cognitive or

noncognitive factors are used. For institutional retention, the random forest achieves an AUC equal to 0.7695 when both cognitive and noncognitive factors are used, compared to 0.7513 and 0.6289 for only cognitive or noncognitive factors respectively. The fact that higher AUC's are obtained for institutional retention than STEM retention indicates that it is easier to identify students most likely to leave the institution than those most likely to leave STEM, conditional on staying at the institution. By comparison, logistic regression models using the same cognitive and noncognitive factors achieve AUCs of 0.7267 and 0.7782 for STEM and institutional retention, respectively. Therefore, the random forest approach slightly outperforms logistic regression for STEM retention and logistic regression is slightly preferable for institutional retention.

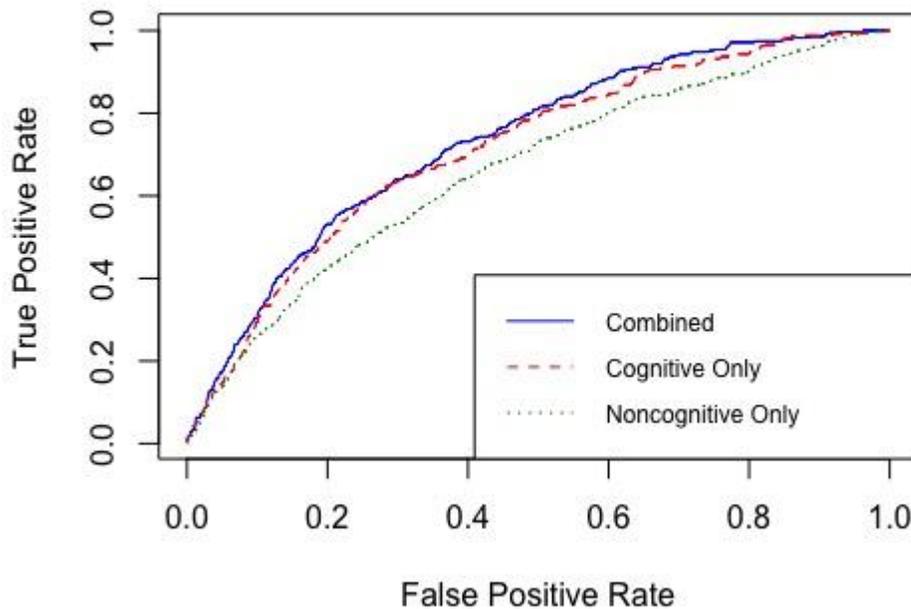


Figure 1: Receiver operating curve for predicting STEM retention combining cognitive and noncognitive factors, along with curves obtained using each type of information alone.

The structure of an individual decision tree offers insight on which explanatory variables are most important in the prediction. Figure 2 gives the first splits in a single decision tree for

predicting STEM retention. Summary information on the number of students (n) and the proportion leaving STEM (y) are provided for the deepest nodes displayed in the figure. Since the first split is performed using information on a student's first semester midterm grades in STEM courses, we can infer that these midterm grades are an important predictor of retention. Other important factors include a student's major, response to a MAP-Works® question about likelihood of changing major, learning community participation, semesters of high school math, high school GPA, ACT math score, and ACT interest questions about whether the student intends to major in STEM, as well as sureness of major.

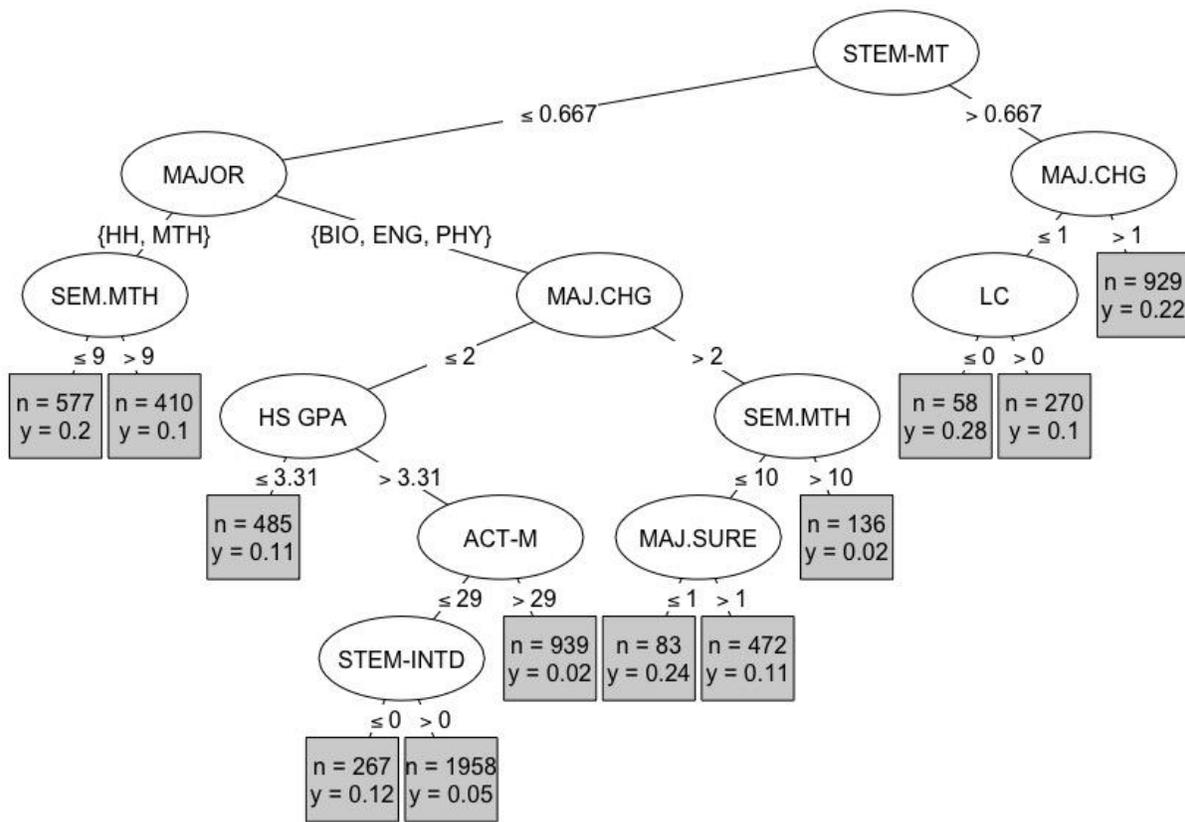


Figure 2: A single decision tree. See Table 1 for variable abbreviations.

Because random forests consist of a large number of trees, grown to great depths, it is difficult to discern the importance of explanatory variables visually. The algorithm discussed in the *Data and Methodology* section can be used to measure the importance of these factors. If a variable is rarely used for splitting nodes, then randomly permuting its values in step 4 of the algorithm will have little impact. On the other hand, randomly permuting values of variables associated with variables used heavily in splitting will lead to a loss of important information and damage AUC. Table 4 gives variable importance scores that reflect the loss in AUC resulting from permutation of each group of variables divided by the total AUC loss over all groups of variables. Summaries are provided for both institutional retention and STEM retention. Figure 3 displays the results graphically, and also breaks down importance scores by type of STEM major.

Table 4:

Relative importance scores for each category of predictor variables
 (* indicates variable is from responses to MAP-Works® survey)

Variable Category	Institutional Retention Relative Importance	STEM Retention Relative Importance
First Semester Midterm Grades	0.5740	0.2794
High School Courses and Grades	0.2053	0.2031
Standardized Tests	0.0650	0.2497
Social Integration*	0.0591	0.0005
College Activities	0.0418	0.0472
First Semester Classes	0.0137	0.0283

Demographics	0.0124	0.0166
Satisfaction with University*	0.0121	0.0014
Academic Skills/Behaviors*	0.0066	0.0011
Learning Community Participation	0.0041	0.0296
Satisfaction with Major*	0.0025	0.0670
ACT Interest Survey	0.0021	0.0481
Math/Science Self Efficacy*	0.0013	0.0281
Scholarships	0.0000	0.0001
Financial Concerns*	0.0000	0.0000

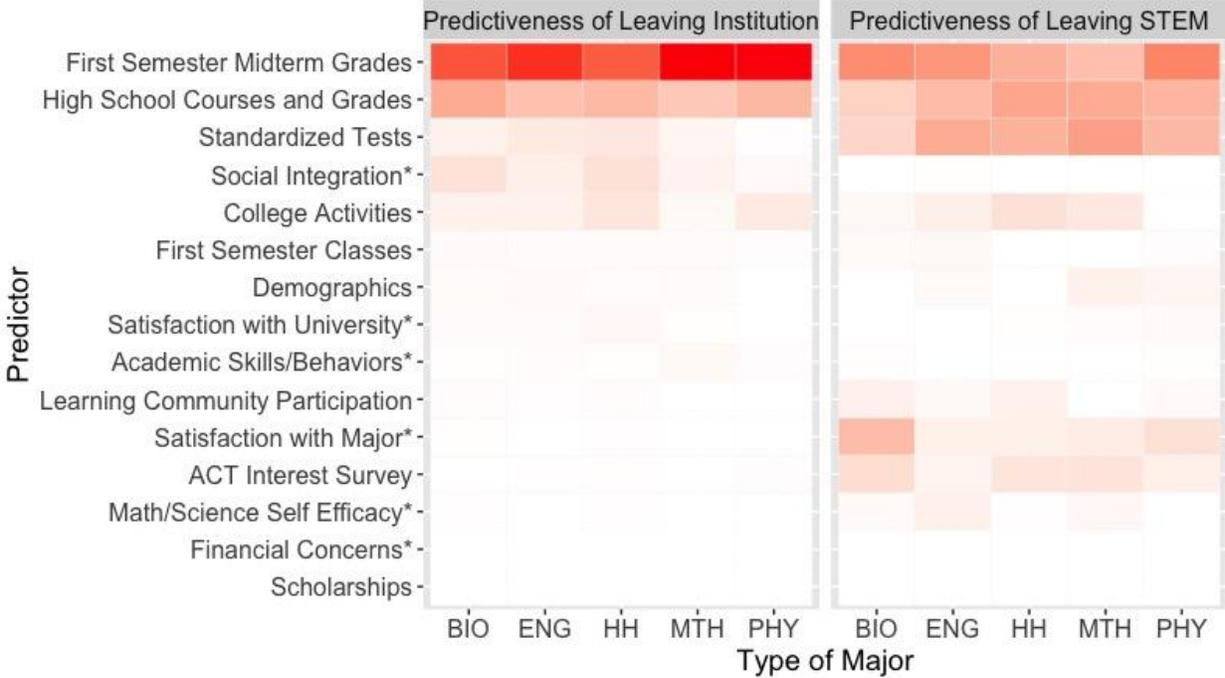


Figure 3: The importance of each category of predictor variable for each type of STEM major (* indicates variable is from responses to MAP-Works® survey).

For institution retention, the predictive importance of midterm grades is more than twice that of any other group of variables, with high school academics consistently ranking second. The importance of factors in the next tier of predictors, i.e. social integration, participation in activities, and standardized test scores, vary somewhat across the majors, but differences are small. Similarly for STEM retention, midterm grades have the most predictive importance, but a larger number of explanatory variables carry predictive value. Goals and interests, learning community participation, self-efficacy, and the MAP-Works® question pertaining to the likelihood of changing major are all stronger predictors of STEM retention than of institution retention. Table 4 and Figure 3 do not tell the direction of the impact that each type of variable has on STEM retention, only whether these sets of variables contain valuable predictive information.

We find that each aspect of Tinto's retention model has some impact on the prediction. The variable groups of the current learning experiences (midterm grades), personal input (high school grades, test scores) and proximal environment (college activities and social integration) provide nearly all of the information needed for predicting institution retention. A student's sense of social integration, as measured through MAP-Works®, has very little value in predicting STEM retention. The fact that a greater number of variables are valuable in predicting STEM retention than institutional retention is consistent with students changing majors for a wide variety of reasons, while students leaving the university commonly share important characteristics, making institutional retention easier to predict.

We note that variables with a large number of missing values are likely to be ranked as less important than they would be if full information was available. This is especially relevant for variables from the MAP-Works® (31% missing) and ACT Interest (39.5% missing) surveys. When two or more predictor variables provide similar information, the random forest's perceived importance of some of these variables is diminished. For example, if gender and ethnicity are strongly associated with other variables that are used, then removing both variables has little impact on predictive accuracy, even if both variables are associated with retention. Furthermore, it is possible that factors associated with retention will differ between different subgroups of students within each type of major.

Because the importance of factors differs between groups of majors, a one-size-fits-all approach to consider all STEM majors as a single group lowers predictive performance. For example, university activities contain more predictive information for health and human science majors than other kinds of STEM majors, while interests and goals are most prominent in predictions involving biological science majors (Figure 3).

Implications for Future Research

Our research is consistent with the belief that early academic performance is a strong predictor of retention, both at an institution and within STEM. Future research might focus on the impact of student engagement early in their first semester of college. Measures of engagement, such as the number of times a student visits a course webpage, or the degree to which active learning strategies are used in a student's courses might provide valuable information capable of improving the accuracy of at-risk predictions. Information about the number of times a student has had a one-on-one conversation with an advisor early in the first semester might also improve predictions.

Since our method is capable of identifying students who are at risk for leaving an institution, or leaving STEM, future work might investigate whether certain interventions are more helpful for one form of retention than the other. A similar approach could also be used to estimate the risk of students leaving specific majors, rather than the broader categories of majors considered here, or to predict graduate student retention.

The degree to which factors contributing to retention vary between particular subpopulations of students is an important question that requires future research. If certain factors are shown to carry particular importance in the retention of underrepresented students, then universities might use this information to improve diversity in STEM. In order to adequately address these questions, researchers should consider measures of integration and engagement in specific STEM courses in addition to the measures of social integration at the university overall, which were available in this study. Future research might also consider which interventions are most helpful for underrepresented students.

Conclusions

Our study sheds light on the differences between a student's decision to leave STEM versus the institution, while highlighting the importance of early academic experiences in both of these decisions. Our findings also point to places where universities could most effectively allocate resources, e.g. on early academic success for both STEM and university-wide retention. Example programs that foster skills for academic success include freshman orientation courses, learning communities, and student-centered teaching methods designed to keep students engaged (Crosling, Heagney, & Thomas, 2009; Graham, Frederick, Byars-Winston, Hunter, & Handelsman, 2013). Such programs have proven effective in increasing retention and improving diversity in STEM (Dagley, Georgiopoulos, Reece, & Young, 2016).

The alerts are an opportunity for proactive advising, which research has shown to be especially beneficial for students underrepresented in STEM (Rodgers, Blunt, & Tribble, 2014). The alert provides advisers with each student's individual risk factors, e.g. a midterm grade of C- or below. Given the list of individual factors, advisors can meet with a student, guide possible interventions and provide assurance that the student's struggles are not unnoticed or out of the ordinary. If a student decides to switch majors, advisors can initiate a joint discussion to identify a different major better matching their interests and goals either inside or outside of STEM.

While our results are specific to the university at which the data were collected, they may be of greatest interest to similar large public universities emphasizing science and technology. The random forest approach can be tailored to other institutions with different proxies for the factors in Tinto's model.

While early academic performance is the strongest predictor of retention both in STEM and at the university overall, social engagement, interests, and goals are also important in STEM retention. Our findings highlight the importance of offering academic support to freshmen early in their first semester. Providing advisors with major-specific information on the highest ranking factors that triggered an alert for individual students may be the most significant benefit of this early alert approach. This in turn may lead to increased STEM retention, as well as institutional retention.

References

- Aulck, L., Aras, R., Li, L., L'Heureux, C., Lu, P., & West, J. (2017). Stemming the Tide: Predicting STEM attrition using student transcript data. *arXiv preprint arXiv:1708.09344*.
- Bettinger, E. P., Evans, B. J., & Pope, D. G. (2013). Improving college performance and retention the easy way: Unpacking the ACT exam. *American Economic Journal: Economic Policy*, 5(2), 26-52. doi:[10.1257/pol.5.2.26](https://doi.org/10.1257/pol.5.2.26)
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
doi:[10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)
- Breiman, L. (2003). Manual—setting up, using, and understanding random forests V4.0.
Retrieved from https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf
- Chemers, M. M., Hu, L. T., & Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational psychology*, 93(1), 55.
doi:[10.1037/0022-0663.93.1.55](https://doi.org/10.1037/0022-0663.93.1.55)
- Chen, X. (2013). STEM Attrition: College Students' Paths into and out of STEM Fields. Statistical Analysis Report. NCES 2014-001. *National Center for Education Statistics*.
Retrieved from <https://eric.ed.gov/?id=ED544470>.
- Crosling, G., Heagney, M., & Thomas, L. (2009). Improving student retention in higher education: Improving teaching and learning. *Australian Universities' Review*, 51(2), 9.
Retrieved from <https://files.eric.ed.gov/fulltext/EJ864028.pdf>
- Daempfle, P. A. (2003). An analysis of the high attrition rates among first year college science, math, and engineering majors. *Journal of College Student Retention: Research, Theory & Practice*, 5(1), 37-52. doi:[10.2190/dwqt-tya4-t20w-rcwh](https://doi.org/10.2190/dwqt-tya4-t20w-rcwh)

- Dagley, M., Georgiopoulos, M., Reece, A., & Young, C. (2016). Increasing retention and graduation rates through a STEM learning community. *Journal of College Student Retention: Research, Theory & Practice*, 18(2), 167-182.
doi:[10.1177/1521025115584746](https://doi.org/10.1177/1521025115584746)
- Ehrenberg, R. G. (2010). Analyzing the factors that influence persistence rates in STEM field, majors: Introduction to the symposium. *Economics of Education Review*, 29(6), 888-891.
doi:[10.1016/j.econedurev.2010.06.012](https://doi.org/10.1016/j.econedurev.2010.06.012)
- Hapfelmeier, A., Hothorn, T., Ulm, K., & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing*, 24(1), 21-34.
doi:[10.1007/s11222-012-9349-1](https://doi.org/10.1007/s11222-012-9349-1)
- Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2005). Survival ensembles. *Biostatistics*, 7(3), 355-373. doi:[10.1093/biostatistics/kxj011](https://doi.org/10.1093/biostatistics/kxj011)
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3), 651-674.
doi:[10.1198/106186006X133933](https://doi.org/10.1198/106186006X133933)
- Janitza, S., Strobl, C., & Boulesteix, A. L. (2013). An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 14(1), 119.
doi:[10.1186/1471-2105-14-119](https://doi.org/10.1186/1471-2105-14-119)
- Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education*, 97(1), 57-70. doi:[10.1002/j.2168-9830.2008.tb00954.x](https://doi.org/10.1002/j.2168-9830.2008.tb00954.x)

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences.

Economics of Education Review, 29(6), 923-934. doi:[10.1016/j.econedurev.2010.06.011](https://doi.org/10.1016/j.econedurev.2010.06.011)

R Development Core Team (2013). R: A language environment for statistical computing Vienna,

Austria: R Foundation for Statistical Computing. Retrieved from [http://www.R-](http://www.R-project.org)

[project.org](http://www.R-project.org)

Rask, K. (2010). Attrition in STEM fields at a liberal arts college: The importance of grades and

pre-collegiate preferences. *Economics of Education Review*, 29(6), 892-900.

doi:[10.1016/j.econedurev.2010.06.013](https://doi.org/10.1016/j.econedurev.2010.06.013)

Rodgers, K., Blunt, S., & Tribble, L. (2014). A real PLUS: An intrusive advising

program for underprepared STEM students. *NACADA Journal*, 34(1), 35-42.

doi:[10.12930/nacada-13-002](https://doi.org/10.12930/nacada-13-002)

Shedlosky-Shoemaker, R., & Fautch, J. M. (2015). Who leaves, who stays? Psychological

predictors of undergraduate chemistry students' persistence. *Journal of Chemical*

Education, 92(3), 408-414. doi:[10.1021/ed500571j](https://doi.org/10.1021/ed500571j)

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable

importance for random forests. *BMC Bioinformatics*, 9(1), 307. doi:[10.1186/1471-2105-](https://doi.org/10.1186/1471-2105-9-307)

[9-307](https://doi.org/10.1186/1471-2105-9-307)

Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in Random Forest Variable

Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8(1),

25. doi:[10.1186/1471-2105-8-25](https://doi.org/10.1186/1471-2105-8-25)

Swaney, K. B., Lamb, R., Prediger, D., & American College Testing Program (1995). *Technical*

manual: Revised unisex edition of the ACT Interest Inventory (UNIACT). ACT. Retrieved

from

<https://www.act.org/content/dam/act/unsecured/documents/ACTInterestInventoryTechnicalManual.pdf>

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research.

Review of Educational Research, 45(1), 89-125. doi:[10.3102/00346543045001089](https://doi.org/10.3102/00346543045001089)

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1-68. doi:[10.18637/jss.v045.i03](https://doi.org/10.18637/jss.v045.i03)

Wilson, Z. S., Holmes, L., Sylvain, M. R., Batiste, L., Johnson, M., McGuire, S. Y., ... &

Warner, I. M. (2012). Hierarchical mentoring: A transformative strategy for improving diversity and retention in undergraduate STEM disciplines. *Journal of Science Education and Technology*, 21(1), 148-156. doi:[10.1007/s10956-011-9292-5](https://doi.org/10.1007/s10956-011-9292-5)

Woodsley, S., & Jones, D. (2011). The foundations of MAP-Works. *Educational Benchmarking*

(EBI). Retrieved from <https://www2.indstate.edu/studentsuccess/pdf/MAP-Works%20Foundation%20Oct%202012%20.pdf>