Predicting Student Retention via Expectancy Value Theory Using Data Gathered before the Semester Begins

Dr. Pamela Bilo Thomas, University of Louisville

Dr. Pamela Bilo Thomas is an assistant professor at the University of Louisville, where she teaches introductory programming languages courses in Python and C/C++. She has published in a variety of journals in conferences in her subject area of computational social science, and is interested in using data and machine learning techniques to understand human behavior.

Dr. Campbell R. Bego, University of Louisville

Campbell Rightmyer Bego, PhD, PE is a cognitive science and engineering education researcher in the Department of Engineering Fundamentals at the University of Louisville's Speed School of Engineering. She studies engineering learning and engineering retention.

Arinan De Piemonte Dourado, University of Louisville

Arinan Dourado, Ph.D., is an Assistant Professor of Mechanical Engineering at the University of Louisville. Prior to joining UofL, he worked as a Lecturer in his home country (Brazil) for three years, teaching and mentoring low-income, first-generation STEM students from rural communities. Additionally, Dr. Dourado worked as an instructor at the University of Central Florida for two years, primarily serving Hispanic first-generation students. Currently, his working on developing and applying machine learning/artificial intelligence tools to identify and suggest intervention actions to increase student retention and success.

Using Early Data and Machine Learning Methods to Identify Students At Risk of Leaving Engineering

Abstract

This full-length research paper presents results from a machine-learning analysis of engineering student persistence at a large southeastern university. Students leave engineering school for many reasons, ranging from low math preparation to a low sense of belonging in engineering, which can be viewed through the Situated Expectancy Value Theory (EVT) framework of academic decision-making. Prior work has found many strong predictors of persistence from first-semester data, including EVT variables, but when it comes to identifying interventions, it might be better to identify predictors from earlier in the first semester.

In this study, we attempted to predict student persistence using two machine learning techniques, neural networks and decision trees, and only using early data from the beginning of the first semester (EVT survey data, standardized test scores, demographic data, and Pell eligibility).

We found that decision trees were better able to predict retention rates from the beginning of the semester than neural networks, as neural networks struggled to find clear signals that indicated if a student was likely to drop out of engineering school. Grouping students together who are at-risk of leaving engineering school from the beginning of the semester will allow instructors and advisors to focus their attention on those groups, and therefore improve retention rates.

Introduction

Educating the next generation of engineering undergraduates is important for many reasons. From a societal perspective, engineers work together to solve problems and improve the quality of life for many people. For students, an engineering major unlocks the ability to get a job in a growing field with a myriad of opportunities. However, many engineering students do not make it through engineering school. Low first-year persistence in engineering programs has been a problem for institutions of higher education for many years. But retaining engineering students is critical for the modern economy; an increasing number of qualified engineers is needed to create and build the infrastructure that is necessary in a globalized world. Unfortunately, students face unexpected challenges in their freshman year. Researchers have discovered that many different experiences and perceptions lead to attrition, including low grades, feelings of belonging uncertainty, imposter syndrome, financial issues, loss of interest in engineering, and other life stressors [1, 2, 3]. Although it is possible to intervene on any one of these factors and hope to make a difference, interventions that target the most important contributors to student decision-making are likely to make the greatest

improvement in retention rates.

Research has revealed that grades received in the first semester are strong indicators for student persistence [4, 5, 6]. This is especially true in mathematics courses; those who perform well (i.e., receive an A or B) are likely to persist, whereas students who perform poorly (i.e., receiving a D, F, or W) are likely to leave (e.g., [7]).

Understanding the factors that lead to persistence should help stakeholders (e.g., instructors and advisors) develop interventions to counteract these problems. The consequences of low first-semester performance can be many. Poor grades from the first semester can lower students' self-perceptions and restrict their eligibility for financial aid, among other consequences. In some cases, like when a student's GPA drops below "good standing" for financial aid, students might not be able to recover within a semester and therefore will ultimately lose their scholarship, even when they are assisted with something like a growth mindset intervention. Thus, this highly predictive performance variable is available too late to change the outcome.

It is therefore important to look earlier in the engineering program timeline for earlier predictors and earlier opportunities to intervene, such as prior engineering experience, math preparation, standardized test scores, high school GPA, and again, perceptions and beliefs, since interventions received in the first year are important for increasing student grades and retention [8]. A framework that addresses these factors simultaneously is situated expectancy-value theory (EVT [9]). EVT describes academic decision-making from an individual student perspective, with primary components of students' (1) expectations of success and (2) subjective task values. These components can be measured with surveys. The summary of this framework is illustrated in Figure 1.

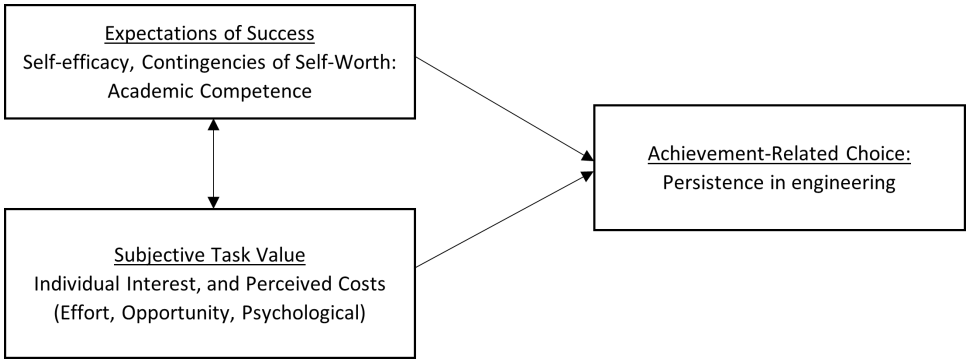

Figure 1: Primary decision-making factors for persistence within the EVT framework.

Previous work has found that the change of EVT values from the beginning to the end of the first semester impacts engineering persistence [4]. For instance, a student for any reason may begin to feel as if they do not belong in an engineering program, and that decrease in sense of belonging indicates that they are more likely to leave the program. But once again, the *change* in student motivation is only available after the first semester, which may be too late to intervene.

Current Study

In the current study, we investigated using early data to predict engineering persistence or attrition to maximize the time available for intervention. We used local data from the University of

Louisville's J. B. Speed School of Engineering, from which fewer than 60 percent of engineering students graduate within six years, and approximately 30% leave within the first year.

We used two different machine learning methods: a neural network and a decision tree. These are the two most commonly used models for predicting persistence (see [10]). The early data included information that was available at the beginning of the school year, which consisted of student demographic data, standardized test scores, and survey data measuring student motivation through the EVT framework, as measured at the beginning of their first semester in engineering.

Our research questions were as follows:

1. Is it possible to predict student persistence using "early" student data with two common machine learning models (neural networks and decision trees)?
2. What are the most important early (i.e., the beginning of first-year) predictors of engineering persistence?

Methods

Participants

This analysis was approved by our institutional review board. Our study utilized retrospective data from two engineering cohorts (students entering in Fall 2018 and 2019, $N = 995$) at the University of Louisville, a large public institution, for whom early EVT survey data was available (i.e., who took the survey during the first week of their freshman engineering course).

Data

De-identified data included:

1. Demographic information, including race, gender, and Pell eligibility,
2. ACT scores (Math, English, and Science/Reading),
3. Financial aid information (category: merit or need based; type: grant, scholarship, loan or workstudy; and source: federal, institutional, private, or state),
4. Responses to EVT surveys, conducted at the beginning of the first semester, including: interest in engineering, perceived costs of studying engineering, self-efficacy, and contingencies of Academic Competence, Academic Competence subscale. Example items and references for each of these scales are provided in Table 1, and
5. Retention, defined as enrollment in the engineering school in the fall of the second year.

Analysis

Two machine learning techniques were investigated in this work: a neural network and a decision tree. A *neural network* works to learn patterns via an iterative process of trial and error to classify data into categorical outputs [11], and the results are black box (it is not possible to tell why a classification was made without the aid of explainable methods). For the neural network analysis,

Table 1: EVT Survey Items

Scale	Response Options	# Items	Example Item
Self-Efficacy [13]	1-Not at All True to 7-Very True	8	I'm certain I can understand the most difficult material presented in this course.
Academic Competence [14]	1-Strongly Disagree to 7-Strongly Agree	5	Doing well in academics gives me a sense of self-respect.
Interest in Engineering [15]	1-Not at All True to 5-Very True	8	Engineering is practical for me to know.
Effort Cost [16]	1-Strongly Disagree to 6-Strongly Agree	4	When I think about the hard work needed to get through engineering school, I am not sure that it will be worth it in the end.
Opportunity Cost [16]	1-Strongly Disagree to 6-Strongly Agree	4	Studying for engineering school takes a lot of time away from other activities that I want to pursue.
Psychological Cost [16]	1-Strongly Disagree to 6-Strongly Agree	3	I'm concerned that my self-esteem will suffer if I am unsuccessful in engineering school.

we split our data into 80 percent training and 20 percent testing data, and trained a multi-layer perceptron to learn patterns that would predict student attrition. We used SMOTE [12] to create balanced classes, since our data was imbalanced with more students staying in engineering after the first semester than leaving. We developed models using the training data and then looked at predictive accuracy values (correct majority prediction, correct minority prediction, and overall accuracy).

Decision trees create a top-down sorting procedure that separates data into different categories [17], which can be used to interpret both the category that best divides the data and the best numerical split that explains differences between categories. After being split, similar data is grouped into "leaves" at the bottom of the tree. A leaf, therefore, contains a group of data that all satisfy specific categorical requirements. We trained the decision tree model to stop splitting the data so that the minimum size of a leaf contained 4 percent of the data. Restricting the minimum leaf size avoids overfitting, which results when classification algorithms become too granular. The code used the Python library scikit-learn to generate the decision tree [18], and visualizations were made using the library dtreeviz [19].

We also performed follow-up analyses on the decision tree results. We looked at the top layer of the tree to identify the variables that impacted engineering attrition the most from the early dataset. We also looked at the lowest level of the tree, ranking persistence probability rates for each leaf, and looked for categorical similarities across high- and low-probability-of-persistence groups.

Results

Descriptive Statistics

Tables 2 and 3 below present demographic statistics of our dataset.

Table 2: Demographic Data

Category	n	% of Data	% Retained
Sex			
Male	778	78.19%	68.12%
Female	217	21.81%	80.18%
Race			
White	797	80.10%	70.01%
Asian	60	6.03%	78.33%
Black/African American	46	4.62%	69.56%
Hispanic/Latino	44	4.42%	68.18%
Economic			
Pell-Eligible	253	34.09%	66.40%

Table 3: Financial Aid Data

Category	n	% of Data	% Retained
Category			
Merit	981	98.59%	70.94%
Need	400	40.20%	66.75%
Type			
Grant	283	28.44%	67.84%
Scholarship	965	96.98%	71.08%
Loan	433	43.51%	65.12%
Work-study	29	2.914%	72.41%
Source			
Federal	525	52.76%	66.85%
Institutional	861	86.53%	72.24%
Private	317	31.85%	75.07%
State	810	81.40%	71.08%

Correlations, Illustrated

Figure 2 illustrates persistence with respect to standardized test scores and EVT values. These figures help to describe the context of the University of Louisville's J. B. Speed School of Engineering and show how there are groups of students who leave engineering at every value of every variable - correlations are difficult to uncover in this dataset since many students who look similar at the beginning of the semester have very different outcomes.

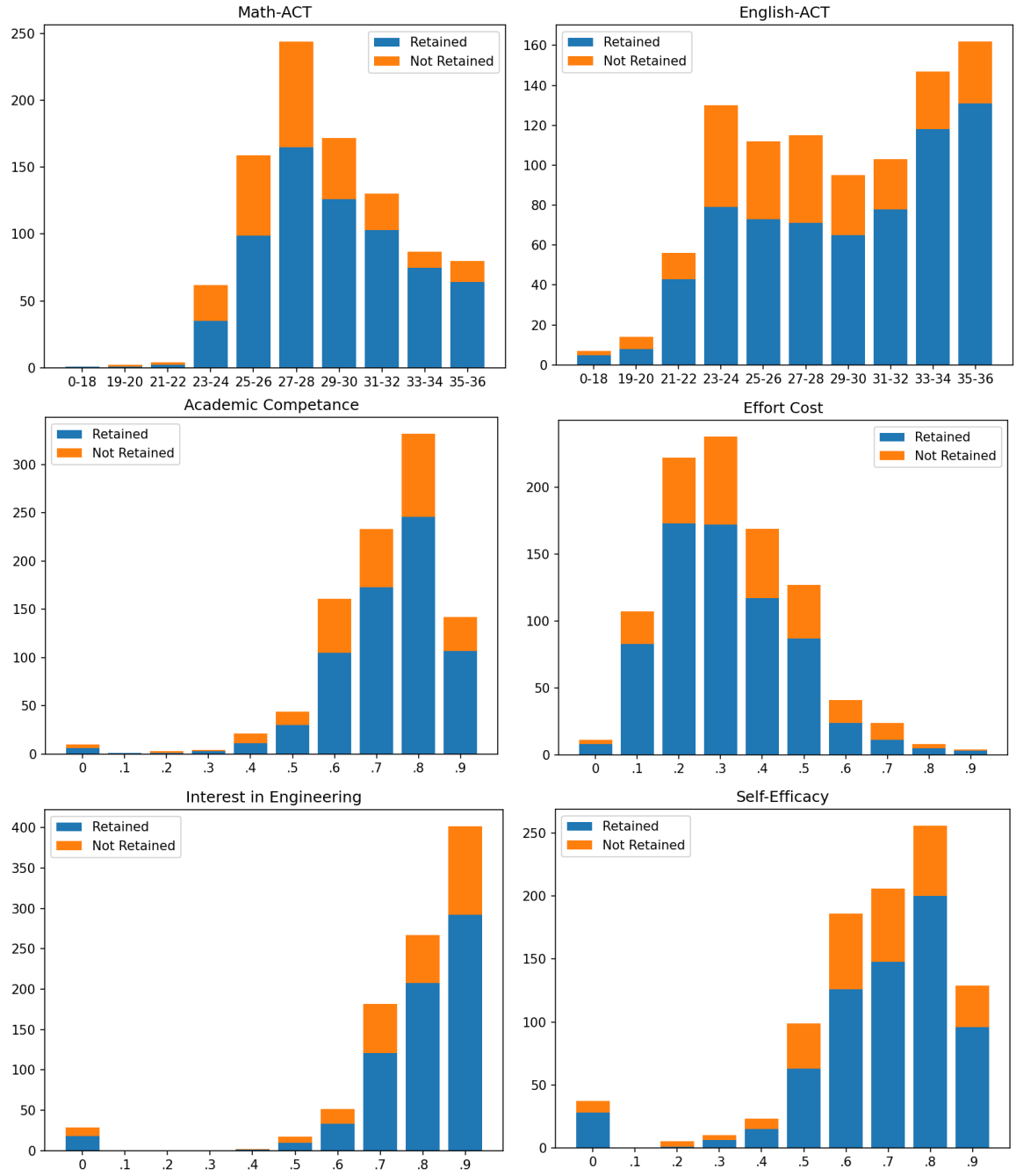

Figure 2: The distribution of student performance on the Math and English sections of the ACT as well as answers to EVT surveys. The stacked histogram also shows the number of students that persist based on their scores and survey responses.

Neural Network

The neural network model performed poorly when it attempted to predict which students were at risk of dropping out using early data. The models failed to converge, meaning that they could not find meaningful patterns that predicted student retention. Despite the application of many

optimization techniques (e.g., varying the size of the dropout parameters, changing the number of nodes included in each layer, stratified sampling, and synthetic data balancing), none of the models that we created were able to create consistent results. No model was able to attain an overall accuracy score of greater than 70 percent (i.e., greater than chance).

Decision Tree

The decision tree results are shown in Figure 3. The highest level of the tree is ACT-English score, indicating that this is the most predictive variable within our dataset. Students who did well on ACT-English ($>$ 31.5) were more likely to persist, whereas students who did poorly ($\leq$ 31.5) were more likely to leave engineering. However, it should be noted that many students who did well on ACT-English did not persist, and many students who had low ACT-English scores stayed in engineering. The distribution of the students who are retained in engineering based on their ACT-English grade is given at the top level of Figure 3.

The second most predictive factor is different for students who performed higher and lower than 31.5 on ACT-English. For students who performed lower on ACT-English, the next most important factor was students' perceptions of the effort required in engineering school (cutoff at 0.52). For students who did well on the ACT-English, interest in engineering (cutoff at 0.74) was the next most important factor.

Based on two splits, the probability of students persisting or leaving engineering is shown in Table 4.

Table 4 shows that persistence is less than 50 percent for students who (1) scored low on the ACT-English and (2) perceived engineering school as requiring a lot of effort. Meanwhile, over 85 percent of students who did well on the ACT-English and have a high interest in engineering stay in the program. The other groups of students have retention rates of 65 and 68 percent, which is little better than chance, since 70 percent of our students persisted into year 2.

Table 4: Persistence rates based upon the first two levels of the decision tree given in Figure 3.

ACT-English $\leq$ 31.5		
EVT Metric	Number of Students	Percent Retained
Effort Cost $\leq$ 0.52	548	67.51
Effort Cost $\geq$ 0.52	90	48.88
ACT-English $\geq$ 31.5		
EVT Metric	Number of Students	Percent Retained
Interest $\leq$ 0.74	77	64.93
Interest $\geq$ 0.74	280	85.17

Ranked leaf results shown in Table 5 reveal additional patterns of results. Several groups have higher retention rates than the average (70 percent), and several groups show retention rates well below the average.

Table 5: Each of the “leaves" generated by the decision tree ordered by percent retained.

N	% Retained	Qualities
54	98.14	ACT-English $\geq$ 31.5, Interest $\geq$ 0.74, Federal-Source Loan, Academic Competence $\leq$ 0.87, ACT-Science/Reading $\leq$ 32.5
46	91.30	27.5 $\leq$ ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, Academic Competence $\geq$ 0.7, ACT-Math $\geq$ 29.5
66	89.39	ACT-English $\geq$ 31.5, Interest $\geq$ 0.74 Federal-Source Loan, Academic Competence $\leq$ 0.87, ACT-Science/Reading $\geq$ 32.5
64	87.50	ACT-English $\geq$ 31.5, 0.74 $\leq$ Interest $\leq$ 0.89, no Federal-Source Loan
40	85.00	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, Academic Competence $\leq$ 0.73, ACT-Math $\leq$ 29.5
43	81.39	ACT-English $\geq$ 31.5, Interest $\geq$ 0.74, Federal-Source Loan, Self Worth $\geq$ 0.87
56	78.57	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, 0.7 Academic Competence $\leq$ 0.73, ACT-Math $\leq$ 29.5, Female
46	78.26	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, Academic Competence $\leq$ 0.7, No Private-Source Loan
41	75.60	ACT-English $\leq$ 27.5, Effort Cost $\leq$ 0.52, Academic Competence $\geq$ 0.7, ACT-Math $\geq$ 29.5
69	71.01	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, 0.7 $\leq$ Academic Competence $\leq$ 0.73, ACT-Math $\leq$ 29.5, Male, Self Efficacy $\geq$ 0.81
53	69.81	ACT-English $\geq$ 31.5, 0.74 $\leq$ Interest $\leq$ 0.89, no Federal-Source Loan
41	68.29	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, 0.7 $\leq$ Academic Competence $\leq$ 0.73, ACT-Math $\leq$ 29.5, Male, Self Efficacy $\leq$ 0.81, Effort Cost $\geq$ 0.4
44	68.18	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, Academic Competence $\leq$ 0.7, Private-Source Loan, Self Efficacy $\geq$ 0.81
77	64.93	ACT-English $\geq$ 31.5, Interest $\leq$ 0.74
48	64.58	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, federal loan
39	51.28	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, 0.7 Academic Competence $\leq$ 0.73, ACT-Math $\leq$ 29.5, Male, Self Efficacy $\leq$ 0.81, 0.31 Effort Cost $\leq$ 0.4
43	51.16	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, Academic Competence $\leq$ 0.7, Private-Source Loan, 0.62 $\leq$ Self Efficacy $\leq$ 0.81
40	45.00	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, 0.7 $\leq$ Academic Competence $\leq$ 0.73, ACT-Math $\leq$ 29.5, Male, Self Efficacy $\leq$ 0.81, Effort Cost $\leq$ 0.31
43	37.20	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, Academic Competence $\leq$ 0.7, Private-Source Loan, Self Efficacy $\leq$ 0.62
42	30.95	ACT-English $\leq$ 31.5, Effort Cost $\leq$ 0.52, no Federal-Source Loan

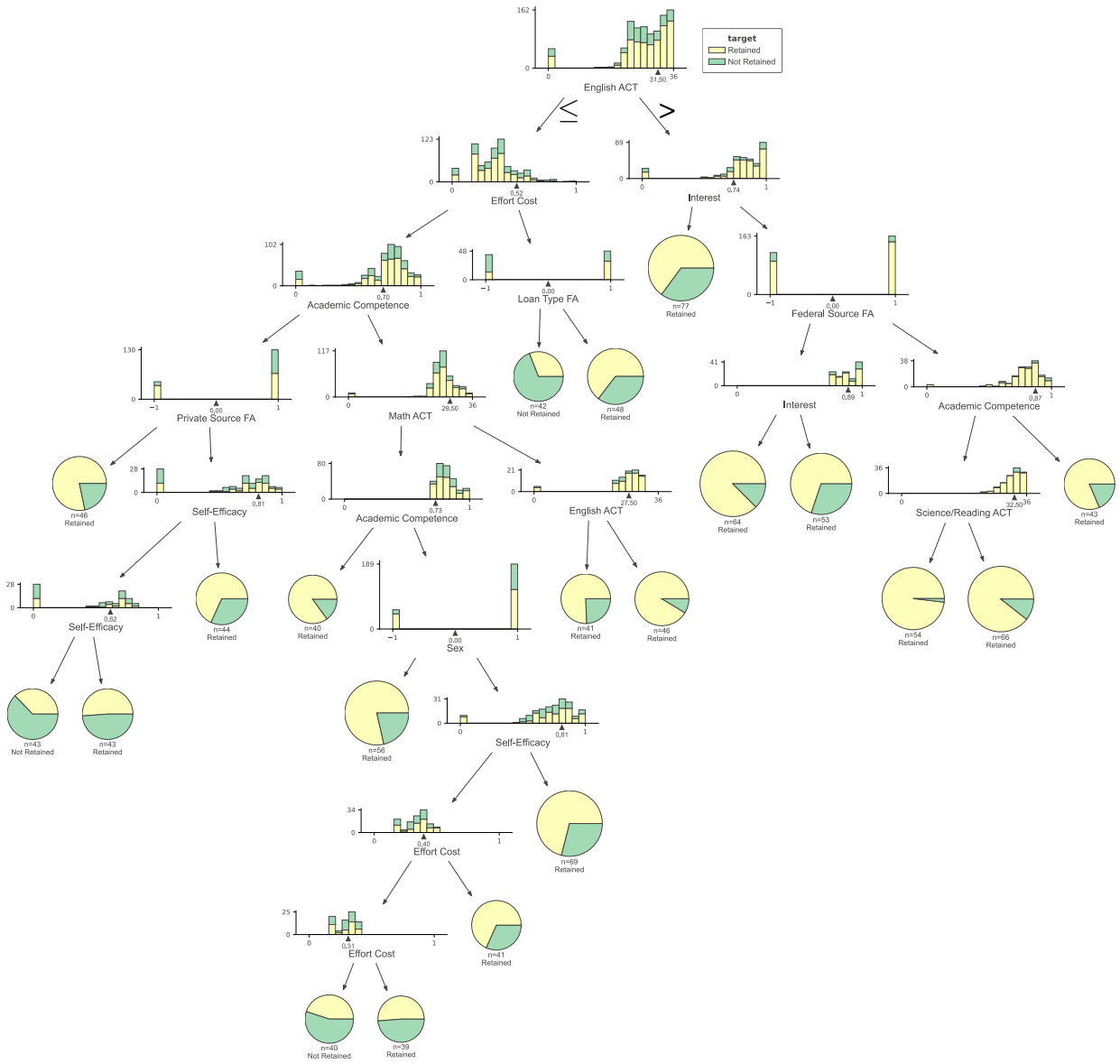

Figure 3: In this decision tree, we see the groupings of students that are more at risk of attrition than others based on data available at the beginning of the first semester. Each leaf contains at least 4 percent of all students so that the model does not risk overfitting. We can see that as the top-level node, a student's ACT-English score is the most important factor that results in students persisting.

Discussion

This study investigated two methods of predicting students at-risk of leaving engineering school using data collected at the beginning of the first semester. It was anticipated that accuracy levels

would be below that of predictive models based on later data.

Neural networks could not predict engineering persistence from early data

Neural networks were not capable of detecting patterns in the early data that consistently led to attrition. Researchers reflected on the limitations of neural networks, and realized that many general issues apply here.

For example, a known difficulty in machine learning is **predicting the minority class**. Due to the intrinsic ratio nature of probability, it is naturally easier to predict the majority class with higher accuracy than the minority class. Predicting the minority class is therefore a known difficulty in machine learning. As we are attempting to predict attrition, and attrition is the minority class, the neural network was having difficulty. In addition, there is no **strong signal** in the available data. As shown in Figure 2, all groups for most independent variables are retained. A neural network required to make a binary prediction based on any individual variable would be inclined to predict "persist" as opposed to "not persist." This would generate a large number of false positives, but would still be greater than 50% in all cases. Simply put, the neural network was unable to learn patterns to predict student attrition with high accuracy when only given information present at the beginning of the semester.

This limitation is logical, because the decision to persist is based not only on student characteristics but also on their experiences in engineering school and the interaction between their characteristics and experiences. Early data only includes pre-college characteristics and preparedness, which is only part of the equation. For this reason, it is difficult to make a binary prediction that will alert us to which students will be retained and which will drop out, since many students who appear to have similar backgrounds when enrolling in school will have different experiences, and therefore different probabilities of retention.

Decision Tree

Although a neural network could not provide accurate classifications, we were able to gain some important information from the decision tree analysis. First, looking only at the top of the tree, we found two groups of students that we could predict as retained or not retained much better than chance. Students who scored lower than 31.5 on the English section of the ACT and also perceived engineering school as requiring a lot of effort left at a rate of 50%, which is much higher than the "chance" value of 30%. Similarly, students who scored higher than 31.5 on ACT-English and higher than 74% interest in engineering were retained at 85%.

The ACT-English split is worth some further discussion. Students who perform very well in English might have an easier time interpreting the word problems in engineering courses, which might allow them to focus on different aspects of the first-year engineering experience (e.g., their interest in the topic) than those who are also struggling with the language. The added load of interpreting English in word problems and in lectures might be too much for those with lower skill levels, which could explain why those with a lower ACT-English score and who perceived a high effort for engineering school were less likely to persist. Secondly, it is important to remember that although the decision tree identified ACT-English at the highest level, ACT-English does not necessarily explain the greatest amount of variance, which could be a conclusion from a regression analysis. It

simply means that this variable at the given value of 31.5 makes the clearest categorical split within the sample at hand. Lastly, it is worth mentioning that our sample has a relatively high distribution of ACT-English scores (35% of our sample scored higher than 31.5; see Figure 2) compared to the national distribution (a score of 32 is the 93rd percentile). This split therefore divides our students into groups of 1/3rd and 2/3rds that have different secondary factors related to persistence.

Looking at Table 5, which shows the variable splits and probability of retention for each leaf, we see that many students were grouped into leaves that had persistence rates similar to the dataset as a whole. These students do not have particularly strong signals that can be used to determine if they stay or leave, which reinforces why predicting student attrition from early data is a challenging problem.

Therefore, a more appropriate way to describe our students may be on a level of risk as opposed to a binary "at-risk" or "not-at-risk" categorization. Just by thinking about students on a scale of risk, it is easier to consider whether they might be able to handle challenges that come their way during their freshman year. For instance, Student A might be more academically prepared for college with EVT response that reflect a high level of self-confidence and interest in engineering. Student B might have lower standardized test scores and doubt their own academic ability. Both students A and B have the potential to persist to year 2 - however, if they both undergo similar stresses in their personal lives, Student A as a low-risk student might be able to recover and persist more easily than Student B. Therefore, Student B's persistence is not deterministic - rather, viewing Student B as higher risk of dropping out alerts their professor or academic advisor to monitor Student B's performance in the class, and step in if their grades begin to slip.

These risk levels also help explain why we had such high false positive and negative groups when we used the neural networks. For the middle groups, giving the students in these groups a binary label would result in high false positive and negative rates. We found it to be more appropriate to use the attrition percentage for each of these groups to describe the likelihood of a student staying in engineering, and chose to use a risk-based approach instead of a binary categorization. Additionally, we did find groups of students, such as those with low ACT-English scores and those that view engineering school to be of high cost, at a higher risk of dropping out.

Areas of Focus for Advisors and Professors

Interestingly, we see that all four facets of EVT, and all three standardized test scores (English, math, and science/reading), were included at some level in the decision tree, while financial aid data and demographic information were less important in predicting student retention. This is encouraging, as EVT parameters are malleable, whereas demographic variables are not. With more individualized information, we may be able to intervene to alter students' risk levels, and even possibly their persistence.

Future Work

As universities move to test-optional admissions, we will lose information that will help us understand students' background knowledge. Future work will have to explore other variable information, including high school GPA and school rankings, among other characteristics. We also see that information can be gathered from EVT scores to understand student persistence and attrition.

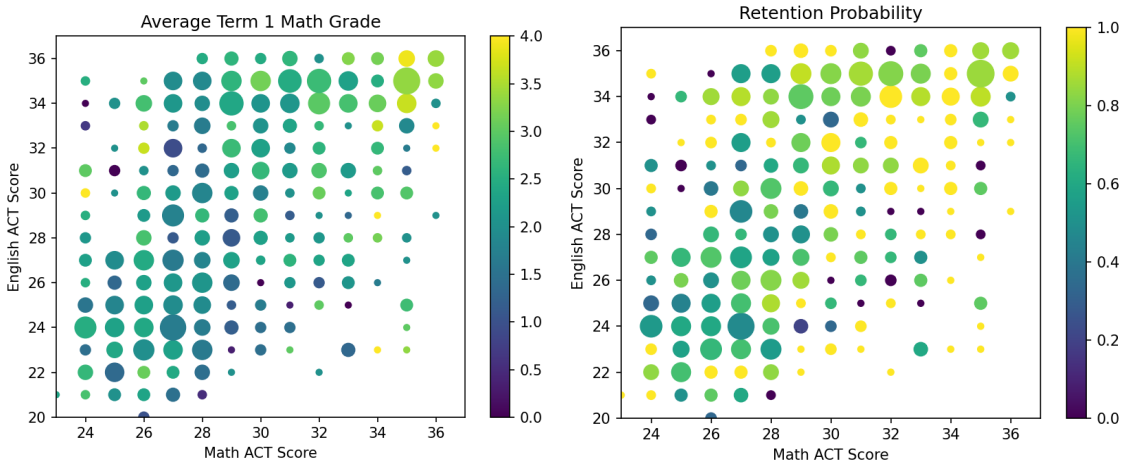

Figure 4: The relationship between ACT score and a student's first term math grade (left) is much clearer than the relationship between ACT score and a student's likelihood of being retained (right). However, ACT scores by themselves not a good way to predict student retention. The size of the dot in the above plots corresponds to the number of students with those respective ACT scores.

Additionally, we expect that as a semester gets underway and grades begin to accumulate, student retention prediction will become more accurate. However, in this case, time and accuracy are trade-offs. Models that incorporate more data will be more accurate, but advisors and professors will lose valuable time which can be used to support and intervene with students who are at risk of leaving the engineering program. Further work can take this information and add in data from early in the semester, such as attendance in the first month, completion of assignments, and midterm grades. This model can alert professors to the students that are struggling during the middle of the semester, when there is still hopefully time to intervene before they fail the class.

In Figure 4, we see that we ACT scores are associated with student's first-term math grade. One future research question would be if we can predict a student's first-term math grade from data gathered at the beginning of the semester. If we can identify which students are at risk of performing poorly in their first math course, interventions can be given to students to help them from the beginning of the class. It would then be our hope that raising a student's first-term math grade will then result in higher retention rates for students.

In this work, the neural network failed to make binary predictions of stay/leave. However, in future work, we can experiment with other neural network architectures to output a user's probability of retention instead of the binary output. We can also experiment with other machine learning models as well, such as Bayesian networks, to see if we can attain our goal of creating an accurate binary classifier that prioritizes students who are at risk of leaving.

Conclusion

In this paper, we showed how we can flag groups of students as at-risk based upon information that can be gathered at the beginning of the school year. Three of the most important variables that were identified as a result of this work was ACT-English scores, the perceived costs of engineering

school, and interest in engineering. In particular, students that do relatively poorly on the ACT-English and perceive engineering as high cost are more likely to drop out, and efforts to help those students should take precedence.

While investigating this problem, we found that this is a hard task, and it is difficult to find clear signals that indicate whether or not a student will leave engineering school at the beginning of the semester with near absolute certainty. However, we do find that we can find certain groups of at-risk students by looking at survey results, demographic data, and standardized test scores. This is important as finding a way to identify at-risk students at the beginning of the semester, such that they can be pointed towards resources that will improve their chances of academic success before they begin to encounter issues and other academic setbacks will ultimately lead to more successful engineering students, graduates, and productive careers. We therefore succeed in finding attributes that identify which students need more attention from professors and academic advisors.

References

[1] K. L. Lewis, J. G. Stout, N. D. Finkelstein, S. J. Pollock, A. Miyake, G. L. Cohen, and T. A. Ito, "Fitting in to move forward: Belonging, gender, and persistence in the physical sciences, technology, engineering, and mathematics (pstem)," *Psychology of Women Quarterly*, vol. 41, no. 4, pp. 420–436, 2017.

[2] E. Ramsey and D. Brown, "Feeling like a fraud: Helping students renegotiate their academic identities," *College & Undergraduate Libraries*, vol. 25, no. 1, pp. 86–90, 2018.

[3] A. M. Gloria, *Psychosocial factors influencing the academic persistence of Chicano/a undergraduates*. Arizona State University, 1993.

[4] C. Bego, P. Thomas, X. Wang, and A. Dourado, "Investigating engineering persistence through expectancy value theory and machine learning techniques," in *2022 ASEE Annual Conference & Exposition*, 2022.

[5] J. Van Dyken, L. Benson, and P. Gerard, "Persistence in engineering: does initial mathematics course matter?" in *2015 ASEE Annual Conference & Exposition*, 2015, pp. 26–1225.

[6] J. A. Middleton, S. Krause, S. Maass, K. Beeley, J. Collofello, and R. Culbertson, "Early course and grade predictors of persistence in undergraduate engineering majors," in *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. IEEE, 2014, pp. 1–7.

[7] C. Bego, J. Hieb, and P. Ralston, "Barriers and bottlenecks in engineering mathematics: How performance throughout a math sequence affects retention and persistence to graduation," in *October 2019 IEEE Frontiers in Education Conference (FIE)*, 2019.

[8] M. Syed, T. Anggara, A. Lanski, X. Duan, G. A. Ambrose, and N. V. Chawla, "Integrated closed-loop learning analytics scheme in a first year experience course," in *Proceedings of the 9th international conference on learning analytics & knowledge*, 2019, pp. 521–530.

[9] J. S. Eccles and A. Wigfield, "From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation," *Contemporary Educational Psychology*, vol. 61, p. 101859, 2020.

[10] F. Agrusti, G. Bonavolontà, and M. Mezzini, "University dropout prediction through educational data mining techniques: A systematic review," *Journal of e-learning and knowledge society*, vol. 15, no. 3, pp. 161–182, 2019.

[11] C. M. Bishop, "Neural networks and their applications," *Review of scientific instruments*, vol. 65, no. 6, pp. 1803–1832, 1994.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[13] P. R. Pintrich *et al.*, "A manual for the use of the motivated strategies for learning questionnaire (mslq)." 1991.

[14] J. Crocker, A. Karpinski, D. M. Quinn, and S. K. Chase, "When grades determine self-worth: consequences of contingent self-worth for male and female engineering and psychology majors." *Journal of personality and social psychology*, vol. 85, no. 3, p. 507, 2003.

[15] Linnenbrink-Garcia *et al.*, "Measuring situational interest in academic domains," *Educational and psychological measurement*, vol. 70, no. 4, pp. 647–671, 2010.

[16] T. Perez, J. G. Cromley, and A. Kaplan, "The role of identity development, values, and costs in college stem retention." *Journal of educational psychology*, vol. 106, no. 1, p. 315, 2014.

[17] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.

[18] 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," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] T. Parr and P. Grover, "dtreeviz: Decision tree visualization," 2020.