STAT 4680:
Statistical Collaboration
Kristopher Pruitt

Final Paper

Marco Codero | Avery Fulton | Greyson Hall | Prue Whaley

Applied Mathematics
UNIVERSITY OF COLORADO **BOULDER**

Abstract

Achieving student success is a multifaceted endeavor, especially in the landscape of higher education. This paper explores the nuanced metrics essential for gauging student success, particularly focusing on major retention and academic performance within the School of Engineering at the University of Colorado - Boulder. While traditional measures like GPA offer insights, they may not encapsulate the full spectrum of student satisfaction and achievement. Through an examination of retention rates for active students and graduation rates for those who have left the institution, we aim to uncover multiple pre-college predictors of success beyond mere academic performance. By delving into these dynamics, this study contributes to a deeper understanding of what constitutes student success in modern educational settings.

1 Introduction

In the pursuit of understanding student success, we will be delving into the metrics that measure it to find the ones most predictive of a successful student. In today's world, the achievement of higher education is often marked by various challenges, and one of the goals of the University of Colorado Boulder (CU) is to ensure student success and retention. While academic achievement is typically reflected in a student's GPA, it is essential to recognize that success encompasses a broader spectrum of outcomes. There is no one definition for what it means for a student to be successful, in some instances, a student can graduate with a high GPA but be unhappy in their chosen field. In another instance, a student can be interested in the material but unhappy with the social climate of their classes and decide to switch majors. In both cases, it would be hard to define if either student was successful with just numerical data. We acknowledge the depth of the problem we are researching and have decided to focus on retention and academic performance as the metrics we will use to determine success. In this study, we delve into the intricate dynamics of student retention and success, focusing on the GPA for those students still pursuing their degree and graduation rates for those who are no longer actively enrolled as primary indicators.

1.1 Background Information

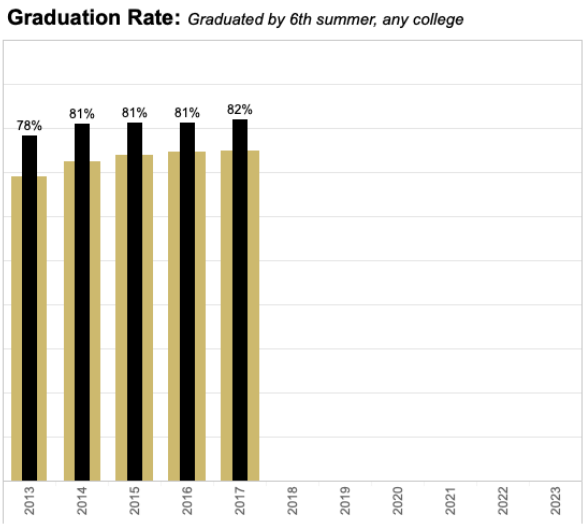

Figure 1: Graduated from any college, black is engineering gold is all of CU. The percentages were calculated by taking the total number of students who started in the year in the x-axis as the denominator with the students who graduated out of those students in the numerator. We are looking at students who graduated in 6 years or less making 2017 the last year we have data for.

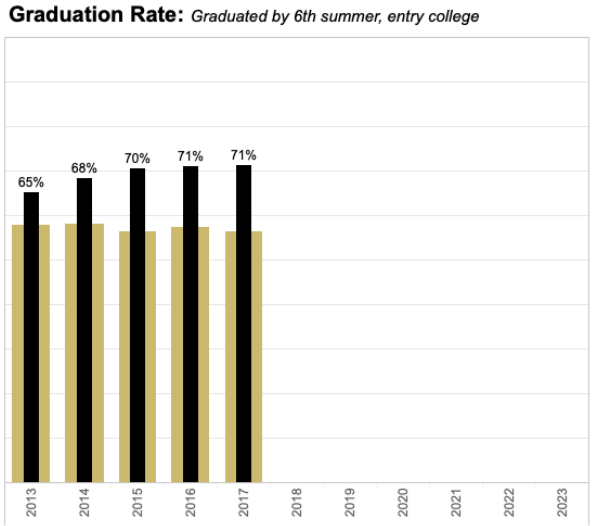

Figure 2: Graduated from their starting college, black is engineering gold is all of CU. The percentages were calculated by taking the total number of students who started in the year in the x-axis as the denominator with the students who graduated out of those students in the numerator. We are looking at students who graduated in 6 years or less making 2017 the last year we have data for.

The data we have consists of information regarding students from the Engineering school at CU Boulder, as the University provides public graduation rates from Tableau. We can see from both Figure 1 and Figure 2 that the most recent data we have is from 2017, Figure 1 shows that 75% of all CU Boulder graduates by their sixth summer while 82% of all engineering graduates by their sixth summer. Figure 2 shows the students who graduated from the college that they started in, 56% of all CU Boulder students graduated from their entry college by their sixth summer while 71% of students who started in engineering graduated in engineering by their sixth summer. These graduation rates overall for the engineering school are fairly good as the engineering school has better rates than the university as a whole, however, there is always room for improvement. According to research.com, Harvard has an overall graduation rate of 97% as of 2023 showing that increased graduation rates are possible. CU Boulder offers many student support services to aid in student success, such as help centers (for example, the writing and math centers), support recitation for difficult classes like Calculus, and easily accessible advising appointments. According to CU Boulder, in 2023, $196,473,227 dollars out of the $2.3 billion dollar budget was allocated to academic support. The main goal of our project is to help identify groups of students who could benefit from the support CU Boulder offers and drive those resources toward the students to increase retention rates in the College of Engineering and overall graduation rates.

Recognizing the complexity of student retention and success, our study adopts a statistical lens to explore the relationships between various factors influencing these outcomes. By employing statistical analyses, we aim to unravel the underlying patterns and associations that shape students' trajectories within their academic journeys. Further, we seek to identify potential predictors and barriers to student retention and success, offering insights that can inform interventions and initiatives aimed at promoting student achievements. By finding and explaining the determinants of student retention and success, we allow professors and administration alike to improve student outcomes by driving the resources that CU Boulder provides to the students who need them most.

1.2 Problem Statement

In the subsequent sections of this study, we will dive into the data analyzed, methodology employed, results obtained, and the conclusions drawn, providing an examination of the intricacies surrounding student retention and success from a statistical perspective. The purpose of this study is to explore the pre-college data of students in the Engineering school because we want to find metrics that are most predictive of a student who will struggle academically in order to encourage those students to use the academic support services offered by CU Boulder, further increasing retention and graduation rates.

2 Data Sources

Our data has been compiled by the University of Colorado Boulder and given to us for the purpose of studying student success by Vanessa Dunn the Director of Analytics for the Engineering school at University of Colorado Boulder. The database includes information about student demographics, high school GPA, AP scores, nationwide standardized test scores (the SAT and ACT), college GPA, college class information, and grades. We have included a full table of the variables in our data set in the appendix. The information has been collected through both the University's Registrars and Admissions offices. The data contains 61 unique variables with 17,347 unique undergraduate records. The data spans from Spring 2013 to Fall of 2024. We know that student success expands beyond just the data that is contained in this database and we will be using the book *Talking about Leaving Revised* to better inform our understanding. CU was one of the universities studied in this book thus making it a good resource for our research. Given that our database contains information from before, after, and during COVID we must take into consideration the effect that COVID had on student success. To assist our understanding, Vanessa recommended looking at the article "Persistence, Relocation, and Loss in Undergraduate STEM Education."

3 Methodology

For this project, we will be doing exploratory and predictive analysis. These are the best-suited methodologies to find the most predictive factors of a struggling student because we are looking for trends in data we already possess. One of the challenges we ran into early in the explorations of our data was missing data points. The missing data is due to external factors such as the change in the scoring of the SATs, lack of pre-college data for transfer students, optional placement tests, and changing admission expectations due to COVID-19. These factors cause students to have varying amounts of data. To account for these variations we have split the students into four groups based on our available data. Those groups are pre-COVID transfers, pre-COVID freshmen, post-COVID transfers, and post-COVID freshmen.

3.1 One-Hot Encoding

The data needs to have a standard format to do any major analysis across the groups. Due to the uniqueness of each student, missing gaps must be filled across all features. One of the ways this can be done is by imputing the missing data or adding new columns that describe and give new information to the system. This process is called One-Hot Encoding, where columns with binary data are created to add new information, and interaction between model features, without applying assumptions to the imputed data. Within our data-cleaning process, we engineered 14 additional columns to describe the standardized test availability, graduation type, math placement scores, high school credit / GPA information, etc. By introducing these features into the data, we can more accurately and easily distinguish between groups of students that have commonalities between them.

3.2 Decision Trees

We'll utilize each of the four groups we've formed to construct decision trees. Decision trees represent non-parametric supervised machine learning algorithms characterized by their tree-like structure. Each internal node of the tree represents a metric of student success that the algorithm has determined is important to the outcome. The terminal nodes represent a classification of whether the student is successful or not. Because of the distinctions between pre and post-COVID students, the response variable varies: for pre-COVID students, it's graduation status in engineering, while for post-COVID students, it's GPA. Furthermore, decision trees have the advantage of producing highly interpretable visualizations.

3.3 Random Forest

The primary drawback of creating a single tree model is its susceptibility to overfitting or underfitting the data, leading to potential unreliability. Using a Random Forest approach solves this issue as it uses bootstrapping and random shuffling to induce artificial variability. This algorithm is made up of multiple decision trees to reach a single result. For our project, we employ the Random Forest approach on the four distinct groups to identify the most influential variables. Subsequently, we will compare the top metrics identified in all four models and deduce the most influential metrics across all groups.

3.4 Linear and Logistic Regression

Using the metrics found from the random forest algorithm we will use linear and logistic regression with the same response variables we used for the decision trees. Both pre-COVID freshman and transfer students will use linear regressions and COVID freshman and transfer students will use logistic regression because of the nature of the response variables used. Linear regression is used for continuous data while logistic regression is used for categorical data. Both types of analysis are used to get a better understanding of the relationship between the response variable and those used to predict it.

4 Exploration

To investigate the landscape of student success in the CU Engineering department, it is important to understand the indicators that not only signal that a student is struggling within the department but also the signals that will make a student stay and succeed within engineering. For students to succeed in the engineering department, they must be able to meet and adapt to the problems and challenges that are present within STEM education, namely course design, pedagogy, and assessment methods (Weston, 2019). This is a difficult metric to quantify but provides good intuition on which factors to explore further and focus our analysis on.

Starting this process, we first transformed the dataset from a CSV file into an R dataframe to more easily manipulate and run models on the data. After speaking with our subject matter expert Vanessa Dunn, she explained that the goal of the engineering department was to ultimately help students succeed and graduate. Therefore, to assess “student success,” we

decided to use “GradEN” as our main response variable, which is a simple binary variable that quantifies whether or not the student succeeded in graduating within the engineering department. Before delving into the rest of the response variables, however, we split the dataset into four categories due to the large difference in the educational experience of all four groups. These groups were determined by two distinct factors. Recognizing the significant impact of the COVID-19 pandemic on the educational experience of all students, we first split the data into two categories: students whose first semester pre- and post-COVID. In addition to the COVID-19 distinctions, we further segmented the data to account for transfer students, whose unique academic trajectories often intersect with varied institutional policies and also had pandemic-related disruptions, thereby constituting a specific group for analysis. We then examined the remaining response variables to determine which of them were the most relevant to predicting if a student would graduate within the department.

Based on our background research, we decided to focus our attention on variables that give insight into how well a student was prepared for a STEM education at CU Boulder as well as variables that give early benchmarks into how they have adjusted to the coursework and environment within their major. Accordingly, we have selected a range of predictive variables that encompass both academic preparedness and early collegiate performance. These include standardized test scores like the SAT and ACT, which serve as high school benchmarks, the number of physics course hours taken, which indicates specific subject readiness, and overall high school GPA, reflecting general academic performance. For insights into college adaptation, we consider the first and last declared major, the number of transfer credits accepted, the timing of the initial engineering course enrollment, and the GPA achieved during the first semester at CU Boulder. When taken together, these variables provide a comprehensive view of a student’s journey from high school through the transition into university-level STEM education, leading to either graduation within the engineering department or another path entirely.

Before beginning our analysis, it is also important to recognize that not every group of students is the same: requirements for pre-college testing, both on the high school level and college level, have changed, as well as the tests themselves. This was the main motivation behind the division of the four groups created to be analyzed separately so each group of students has similar amounts of data.

5 Analysis and Results

As mentioned above, four groups of students at the University of Colorado-Boulder were the main focus of this study. Specifically, pre- and post-COVID transfer students and pre- and post-COVID freshman students. To determine which features are most important for pre-COVID students, the graduation rate was the selected response variable, whereas, for post-COVID students, GPA was the selected response variable. In both cases, a random forest model was employed to predict student success due to its ability to capture complex, non-linear relationships between predictor variables and outcomes, robustness to overfitting, and effectiveness in handling missing data. Below are the generated plots for each group of students.

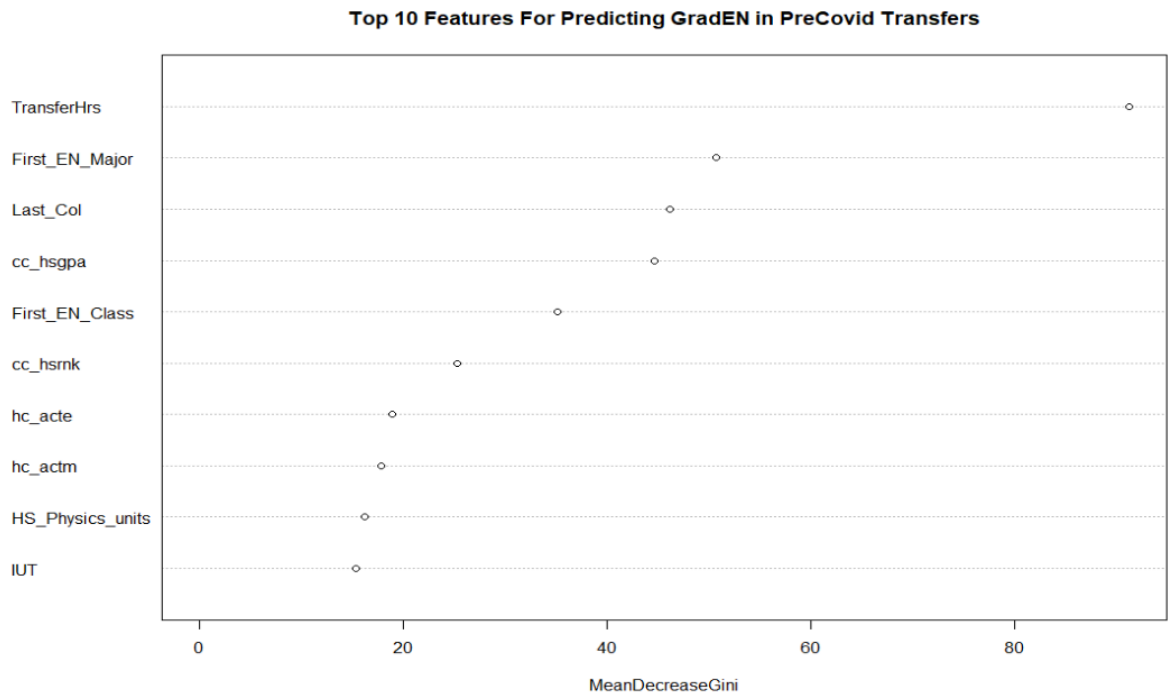

Figure 5: Significant features for pre-COVID transfer students

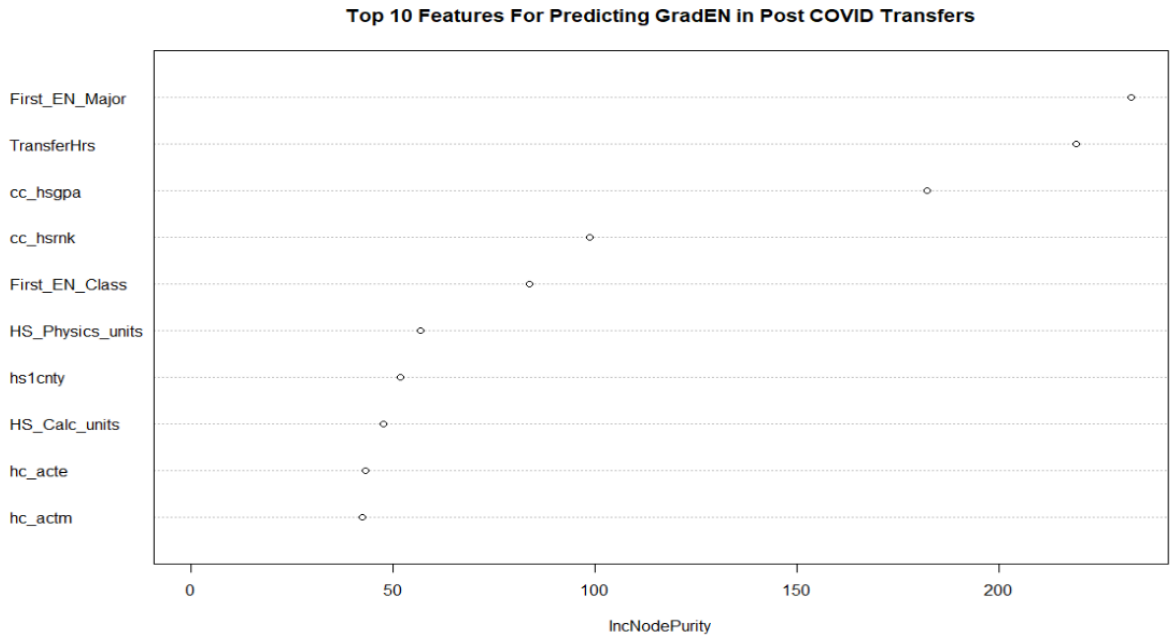

Figure 6: Significant features for post-COVID transfer students

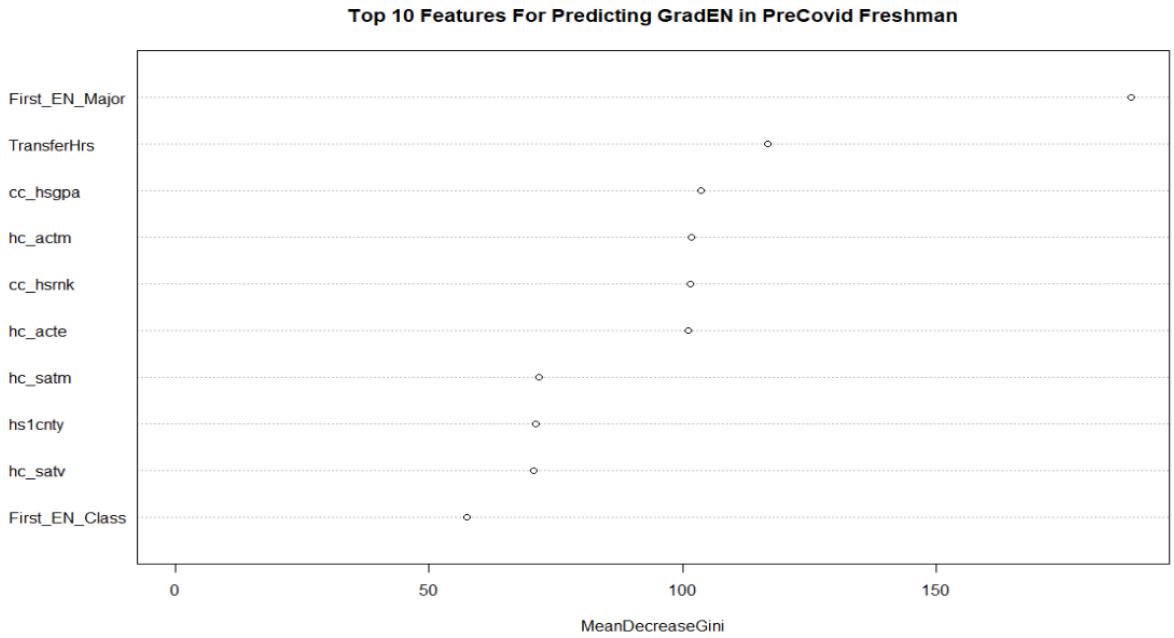

Figure 7: Significant features for pre-COVID freshman students

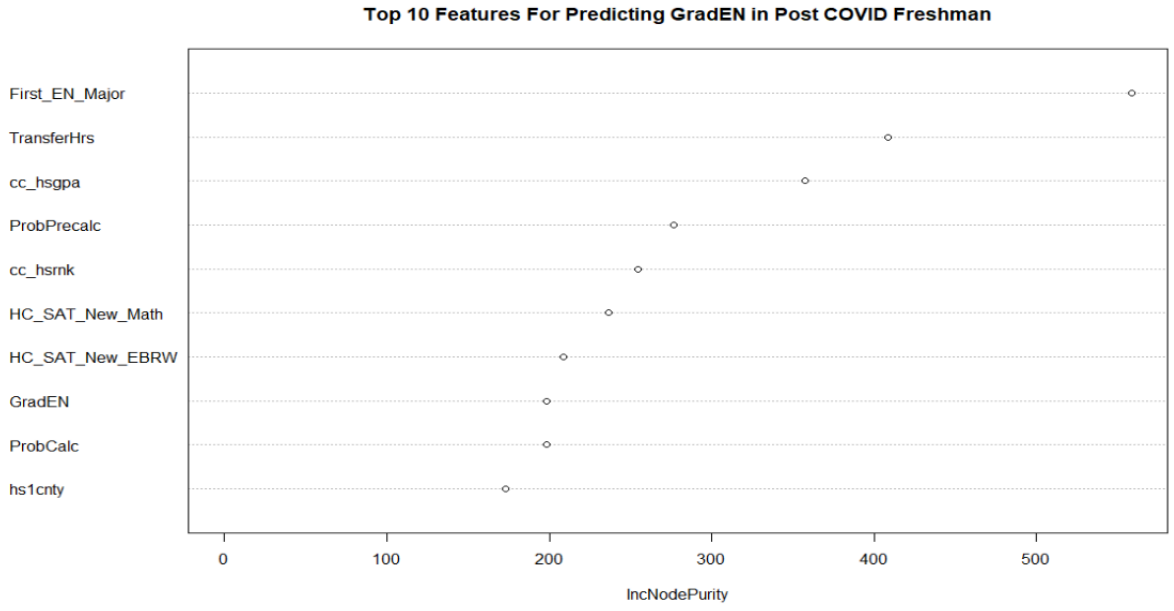

Figure 8: Significant features for post-COVID freshman students

From figures five through eight, it can be seen that the variable "First_EN_Major" is the most predictive factor in every instance except pre-COVID transfers. This variable represents the first major that a student chooses within the engineering discipline. Further, "TransferHrs" and "hsgpa" (a students' transfer hours and high school GPA, respectively) both are significant features. These being the highest weighted features is expected, as they both reflect a student's

academic preparedness and consistency throughout their high school years, indicating their readiness for college-level coursework. Additionally, in most cases, transfer students' previous experience navigating college environments demonstrates their adaptability, while the relevance of transfer coursework to their chosen field of study provides a potential advantage and reflects in their GPA and commitment to their major of choice.

COVID impacted many students, both before and after the global pandemic. The transition from pre-COVID to post-COVID transfer students introduces many differences in predictive factors for college success. While pre-COVID transfer students navigated traditional learning environments with in-person classes and regular campus activities, post-COVID transfer students faced disruptions due to the sudden shifts to remote or hybrid learning models. These changes could impact academic performance, adaptability, and overall readiness for college-level coursework differently among the two groups. From the four figures above, only *Figure 5*, the pre-COVID transfer students have a different order of top three predictive factors. As opposed to their first engineering major, their transfer hours, and their high school GPA, the order becomes their transfer hours, their first engineering major, and "Last_Col," the most recent college they attended. In context, these predictive factors make a lot of sense, the last college a student attended will strongly influence their experience and overall success at a new university, and the other two factors have been explained above.

6 Discussion

From our analysis, we claim a student's transfer hours, high school GPA, and their first major in the College of Engineering are the top predictive factors of student success across all four groups. The relationships between if a student is successful or not and their transfer hours and GPA is clear: the more transfer hours and the higher the GPA the more likely a student will be successful. The relationship between student success and their first college major is not as simple.

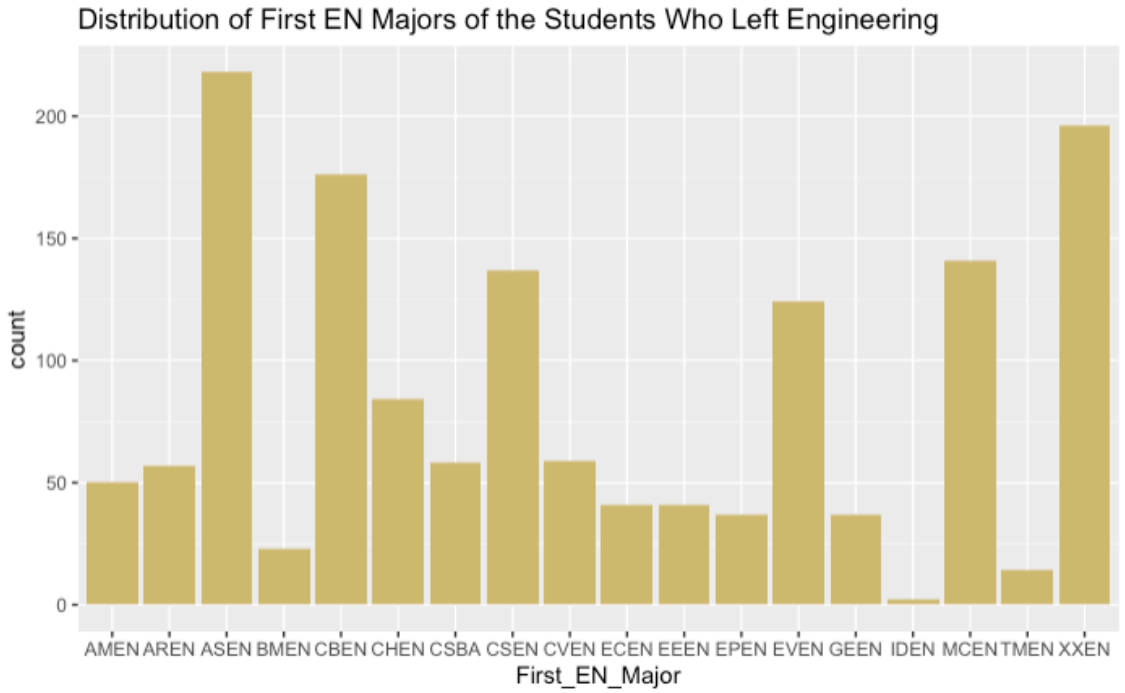

Figure 9: Bar chart of the first Engineering major for the students who have left the College of Engineering.

To further interpret how a student's first engineering major can be predictive of success, Figure 9 shows the amount of students who leave engineering to either transfer to another college or drop out of the university. The top three majors that are losing students are Aerospace Engineering (ASEN), Exploratory Studies (XXEN), and Chemical and Biological Engineering (CBEN). The majority of students who leave the College of Engineering transfer into the College of Arts and Sciences. These results match our prior intuition about what majors are struggling because Aerospace and BioChem are considered two of the hardest engineering majors. On the other hand, Exploratory Studies is the route students who got into the engineering school but have not yet picked a major take, so it makes sense that a portion of them would decide engineering as a whole is not for them. The next step for the University based on our findings is to advertise the resources and offer more help to the students in these three majors to help reduce the amount of students who drop the major.

Another interesting finding of our study is that SAT scores and ACT scores only appear in our top ten predictive factors in the post COVID transfer group. This tells us that standardized test scores are not as helpful in determining the success of a student as other factors in our study. This provides evidence to support the University of Colorado Boulder decision to no longer require students to submit their standardized test scores. Initially, the university decided to stop requiring test scores because of challenges taking the tests due to COVID. Currently, the university has decided to continue the practice because they believe the value of a student lies in more than just a test score.

7 Conclusion

Our research has demonstrated that the interpretation and prediction of student success is governed by the unique journeys that each student takes. Looking at the predictors that contributed most to our response variables, it's clear that all of them acted in tandem to detail what path a particular student chose and how they responded to the challenges along their chosen path. Whether it be how far along in their journey the students are, which major they chose, or how well their high school prepared them, we see that across all of our student groups, it is clear to see that student success is not a one-size-fits-all equation. However, we have been successful in identifying key variables that are better indicators for the story of success than some others.

While our analysis did allow us to isolate key variables, it was limited in the sense that of the many variables that were present within our dataset, only a few of them were able to tell complete stories for every student within the analysis. As our data was composed of every previous study conducted by the University data analysis department, some of the variables were subset in a way that made them completely non-informative. This is compounded by inconsistent policy by the University, which over the course of the time scale that our data examined, frequently changed the requirements for what students should submit when applying.

Further, while it has been shown that indicators such as race and gender are extremely significant in the experience of a student (especially within the department of engineering), we were unable to examine the influence of these predictors, as we were working with an anonymized dataset in order to protect student privacy. However, while these would be good additions to model student success, care would have to be used if such predictors were included in further analysis, as building in such bias into a model can often lead to the perpetuation of the same bias.

In light of what our investigation was able to uncover as well as what we were limited by, we urge that future research that includes a broader spectrum of the student experience, and are able to more directly interpret a student's experience when they arrive at the university itself. Metrics such as connections between classmates, interactions with professors, and engagement with campus resources/programs would contribute greatly to improving the model of a student experience, and therefore improve our model for predicting student success. It is also clear that a more standardized approach to data collection is instrumental in giving future data scientists consistent data from which to draw more nuanced conclusions. This would also pave the way for new questions and investigations within the department, taking our findings and creating new metrics and models to ensure that every student is allowed to thrive and prosper to their fullest potential.

References

Seymour, E., & Hunter, A.-B. (Eds.). (2019, December). *Talking about Leaving Revisited*. Springer Cham.

Burdman, P., Baker, M., & Henderson, F. (2021). *Charting a new course: Investigating barriers on the calculus pathway to STEM*. California Education Learning Lab.

Filatova, E., & Chen, Y., & Li, H. (2023, June), *Analysis of the COVID-19 Impact on Students' Enrollment, Performance, and Retention* Paper presented at 2023 ASEE Annual Conference & Exposition, Baltimore , Maryland. 10.18260/1-2--42660

"CU Boulder Undergraduate Retention & Graduation Rates." *Public.Tableau.Com*, public.tableau.com/app/profile/university.of.colorado.boulder.ir/viz/GradRetention_Rates/RetentionGraduationRates. Accessed 13 Apr. 2024.

The University of Colorado Boulder - Boarddocs, boarddocs, www.boarddocs.com/co/cu/Board.nsf/files/9E3S9T650A2C/$file/Colorado%20Football%20Ops%20Program%20Plan%2011_18_13%20Final%20with%20GIPF%20as%20Supplement.pdf. Accessed 13 Apr. 2024.

"Harvard University Graduation Rate & Career Outcomes 2023 | Research.Com." *Research.Com*, research.com/best-colleges/harvard-university/graduation-rate-and-career. Accessed 13 Apr. 2024.

GfG. "Decision Tree." *GeeksforGeeks*, GeeksforGeeks, 20 Aug. 2023, www.geeksforgeeks.org/decision-tree/#:~:text=A%20decision%20tree%20is%20a,both%20classification%20and%20regression%20problems.

[7] Appendix A: Variables Used

NAME	LABEL	Format	Notes
First_EN_MajorAREN	First engineering major was Architectural Engineering	Binary	0 = no, 1 = yes
First_EN_MajorASEN	First engineering major was Aerospace Engineering	Binary	0 = no, 1 = yes
First_EN_MajorCBEN	First engineering major was Chemical and Biological Engineering	Binary	0 = no, 1 = yes
First_EN_MajorCHEN	First engineering major was Chemical Engineering	Binary	0 = no, 1 = yes
First_EN_MajorCSEN	First engineering major was Computer Science	Binary	0 = no, 1 = yes
First_EN_MajorCVEN	First engineering major was Civil Engineering	Binary	0 = no, 1 = yes
First_EN_MajorECEN	First engineering major was Electrical and Computer Engineering	Binary	0 = no, 1 = yes
First_EN_MajorEEEN	First engineering major was Electrical, Computer & Energy Engineering	Binary	0 = no, 1 = yes
First_EN_MajorEPEN	First engineering major was Engineering Physics	Binary	0 = no, 1 = yes
First_EN_MajorEVEN	First engineering major was Environmental Engineering	Binary	0 = no, 1 = yes
First_EN_MajorGEEN	First engineering major was General Engineering	Binary	0 = no, 1 = yes
First_EN_MajorMCEN	First engineering major was Mechanical Engineering	Binary	0 = no, 1 = yes

First_EN_MajorTMEN	First engineering major was TMEN	Binary	0 = no, 1 = yes
First_EN_MajorXXEN	First engineering major was XXEN	Binary	0 = no, 1 = yes
TransferHrs	Number of transfer credits	Integer	Calculated by subtracting the earned hours in the first term from the TOTAL earned hours
cc_hsgpa	High School GPA	Float	
First_EN_ClassJR	First engineering course was in junior year	Binary	0 = no, 1 = yes
First_EN_ClassSO	First engineering course was in sophmore year	Binary	0 = no, 1 = yes
First_EN_ClassSR	First engineering course was in senior year	Binary	0 = no, 1 = yes
First_EN_ClassSR5	First engineering course was in super senior year	Binary	0 = no, 1 = yes
hc_acte	ACT english score	Float	
hc_actm	ACT math score	Float	
HS_Physics_units	High school physics units	Integer	1 unit = 1 full year in HS
IUT	Intra-University Transfer	Binary	0 = no, 1 = yes
GradEN	Student earned a Bachelor's degree in the engineering department	Binary	0 = no, 1 = yes
hs1cnty	High School County	3-Digit Code	