2021 ASEE ANNUAL CONFERENCE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #34356

A Data-driven Approach for Understanding and Predicting Engineering Student Dropout

Danika M. Dorris, North Carolina State University

Danika Dorris is a Ph.D student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. She received a Bachelor's of Science in Industrial and Systems Engineering from the University of Tennessee. Her work currently focuses on modeling wellbeing of emerging adults and college student attrition

Dr. Julie L. Swann, North Carolina State University

Julie Swann is the department head and A. Doug Allison Distinguished Professor of the Fitts Department of Industrial and Systems Engineering. She is an affiliate faculty in the Joint Department of Biomedical Engineering at both NC State and the University of North Carolina at Chapel Hill. Before joining NC State, Swann was the Harold R. and Mary Anne Nash Professor in the Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. There she co-founded and co-directed the Center for Health and Humanitarian Systems (CHHS), one of the first interdisciplinary research centers on the Georgia Tech campus. Starting with her work with CHHS, Swann has conducted research, outreach and education to improve how health and humanitarian systems operate worldwide.

Julie Ivy, North Carolina State University

Julie Simmons Ivy is a Professor in the Edward P. Fitts Department of Industrial and Systems Engineering and Fitts Faculty Fellow in Health Systems Engineering. She previously spent several years on the faculty of the Stephen M. Ross School of Business at the University of Michigan. She received her B.S. and Ph.D. in Industrial and Operations Engineering at the University of Michigan. She also received her M.S. in Industrial and Systems Engineering with a focus on Operations Research at Georgia Tech. She is President of the Health Systems Engineering Alliance (HSEA) Board of Directors. She is an active member of the Institute of Operations Research and Management Science (INFORMS), Dr. Ivy served as the 2007 Chair (President) of the INFORMS Health Applications Society and is a past President for the INFORMS Minority Issues Forum. Her research interests are mathematical modeling of stochastic dynamic systems with emphasis on statistics and decision analysis as applied to health care, public health, education and humanitarian logistics.

A Data-Driven Approach for Understanding and Predicting Engineering Student Dropout

Abstract

This is a research paper focused on identifying influential factors of student dropout. Students who drop out of college can suffer negative effects on their wellbeing long after they leave college. Many of these dropout students are dropping out in the first two years which highlights the importance of identifying at-risk students early on. We analyzed data for the 1,754 students in the 2014 undergraduate engineering cohort at a large public university. Of these 1,754 students, 12.6 percent dropped out. Ninety-two factors describe each student's application information (e.g., SAT), academic metrics, (e.g. course load, GPA), and demographic information (ethnicity, age). Defining characteristics and significant predictors of at-risk students were identified using three types of analyses: (i) statistical testing for comparisons, (ii) cluster analysis, and (iii) logistic regression predictions. We first identify significant differences between graduate and dropout populations with hypothesis testing. Then, we use clustering to identify subgroups within the cohort and label each group according to a set of defining characteristics. Lastly, significant predictors are extracted from a logistic regression model predicting eventual dropout. Statistical testing for comparisons found that there was a lower proportion of female and full-time students in the dropout population than those who graduated. Most dropout students formed a separate cluster from the rest of the cohort, and the time of dropout influenced the clusters formed within the dropout population. From the regression models, we learn that GPA and passed credits are significant predictors in the first year, and race does not become a significant predictor of dropout until the second year. The factors that influence dropout change over time which emphasize the importance of dynamic dropout prediction models. The findings from each phase of our analysis highlight the complexity of understanding the causes of dropout and the importance of personalizing interventions for specific populations within a cohort.

Introduction

Nearly 20 million students attended American colleges and universities in Fall 2019, and roughly 625,000 of these students were enrolled in an undergraduate engineering program[1], [2]. Thirty percent of engineering students drop out before the second year[3], and more than 60 percent of dropouts occur in the first two years[4]. 10% of college students suffer from anxiety disorders[5], while 20% meet the criteria for alcohol use disorder[6]. Additionally, most college students are not actively seeking the help they need to manage these stressors, which negatively impact graduation, employment, and financial health[5], [7], [8]. Students who drop out of college can suffer negative effects on their wellbeing long after they leave college [9], [10].

In order to reduce the dropout rate, universities need to know which students are at risk and why before they can intervene. Academic performance, admissions data, and student survey responses have commonly been used to better understand dropout [9], [11]–[13]. Logistic regression, neural networks, and decision trees are among the methods that have been used for dropout risk prediction [3], [9], [14], [15]. While some studies have found that academic rigor is a contributor to dropout, several studies have found that non-academic factors such as lack of confidence, lack of teacher-student interaction, poor quality of teaching, and lack of belonging also affect a

student's decision to drop out [10], [13], [16], [17]. Many studies have also found that a higher proportion of ethnic minorities and women are dropping out, though a few studies have found these demographic factors to not be as influential as other non-academic factors [10], [13], [17].

We propose a framework for better understanding dropouts from engineering before graduation, which uses data commonly collected by universities and combines three educational data mining approaches[18]: prediction, clustering, and relationship mining. Through our framework, we focus on answering one primary question: Which students are at-risk of dropping out? To answer this primary question, we can break down this primary question into six foundational questions, which motivate the three phases in our framework: (1) Are there differences between dropouts and graduates? (2) How do dropouts differ from the rest of the cohort? (3) What are the predominant characteristics of the cohort? (4) What are characteristics of the dropout population? (5) What are the dropout predictors? (6) How do dropout predictors change over time?

Our combined approach aims to identify distinct features of the dropout population and to extract significant predictors of eventual dropout using student information systems (SIS) data collected early on in a student's academic career. This framework is built on a three-phase approach involving (i) statistical testing for comparisons, (ii) cluster analysis, and (iii) logistic regression predictions, where the earlier analyses inform the later ones. Specifically, we first identify significant differences between graduate and dropout populations with hypothesis testing. Then, we use clustering to identify subgroups within the cohort and categorize each group according to a set of defining characteristics. Lastly, significant predictors are extracted from a set of logistic regression models predicting eventual dropout with rolling time horizons. We demonstrate this framework using an engineering cohort at a large public university, though this framework is generalizable for other universities or academic years.

Data

This study analyzed data for 1,754 students in the 2014 undergraduate engineering cohort at a large public university in the southeast. This cohort includes any undergraduate student who was admitted to the College of Engineering in either the summer or fall of 2014 and who enrolled in the fall 2014 semester. This year was selected for the beginning of the analysis to allow time for graduation.

Ninety-two factors analyzed include demographic information, academic performance (e.g., GPA), academic program (e.g., major), course load (i.e., number of courses taken in a semester), and academic success (e.g., graduated). Several of the variables are listed in Table 1, with a complete list of factors found in Appendix Table A.1. Table 1 also indicates the phases of analysis in which each variable was included.

Academic factors were updated on census date (ten days after the first day of classes) and on the last day of the semester during each semester a student was enrolled. The census date is the last day for tuition refunds due to dropping a course or changing from credit to audit. It is also the last day for undergraduate students to drop from full-time to part-time and the last day to drop a course without receiving a "W" grade.

Table 1. Abbreviated list of variables. An "x" in each phase column indicates that the variable was included in the corresponding phase of analysis.

Label	Type	Phase 1	Phase 2	Phase 3
Admitted in Fall 2014	Binary	x	x	x
First term enrolled was Fall 2014	Binary	x	x	x
Cumulative Term GPA for Career at End of Term	Interval	x	x	x
Gender	Binary	x	x	x
Reported Ethnic Group 1 (White)	Binary	x	x	x
Reported Ethnic Group 2 (Black or African American)	Binary	x	x	x
Reported Ethnic Group 3 (Hispanic/Latino)	Binary	x	x	x
Reported Ethnic Group 4 (Asian)	Binary	x	x	x
Reported Ethnic Group 5 (American Indian or Alaska Native)	Binary	x	x	x
Reported Ethnic Group 6 (Not Specified)	Binary	x	x	x
Reported Ethnic Group 7 (Native Hawaiian or Other Pacific Islander)	Binary	x	x	x
Student in Two or More Races	Binary	x	x	x
Tuition Residency	Binary	x	x	x
Academic Level - Term Start	Interval	x	x	x
Engineering First Year Major	Binary	x	x	x
Mechanical Engineering Major	Binary	x	x	x
Age	Interval	x	x	x
Academic Load at Census Date	Nominal		x	x
F - Full-Time at Census Date	Binary	x		
P - Part-time at Census Date (Anyone Less Than Full-Time)	Binary	x		
T - Three Quarter Time at Census Date	Binary	x		
H - Enrolled Half-Time at Census Date	Binary	x		
L - Less than Half-Time at Census Date	Binary	x		
N - No Unit Load at Census Date	Binary	x		
SAT Composite Score	Interval	x	x	x
Dropout occurred after three years	Binary	x	x	
Dropout occurred after two years	Binary	x	x	
Dropout occurred within the second year	Binary	x	x	
Dropout occurred within the first year	Binary	x	x	
Student graduated	Binary	x		
Student dropped out	Binary	x	x	x

New freshmen are enrolled in the Engineering First-Year (EFY) major until they are admitted to a specific engineering major, which is contingent upon them completing a set of general engineering requirements and are assigned to a specific engineering major. New freshmen are typically admitted to a specific engineering major at the end of the second year. The EFY major indicator also informs us about transfer students, since less than three percent of transfer students were EFY majors.

For this study, a “dropout student” is defined as an enrolled student who leaves the university and does not return within six years from the time they first enroll. A “graduated” student is defined as an enrolled student who completes a degree within six years from the time they first enroll. Students who switch programs within the university and eventually graduate from a non-engineering program are still considered a graduate of this cohort. Students who are currently enrolled and still working toward a degree are only included in Phases 2 and 3 of our analysis.

This research received approval from the Institutional Review Board at the university. The results in this article comply with the data management plan for the research including that a minimum number of entries is needed in published results.

Methodology

This framework for systematically classifying students involves a three-phase approach: (i) statistical test for comparisons, (ii) cluster analysis, and (iii) logistic regression predictions of eventual dropout. Figure 1 shows the relationship among each of the phases in the three-phase approach. All analyses in this study were performed with SAS Enterprise Guide.

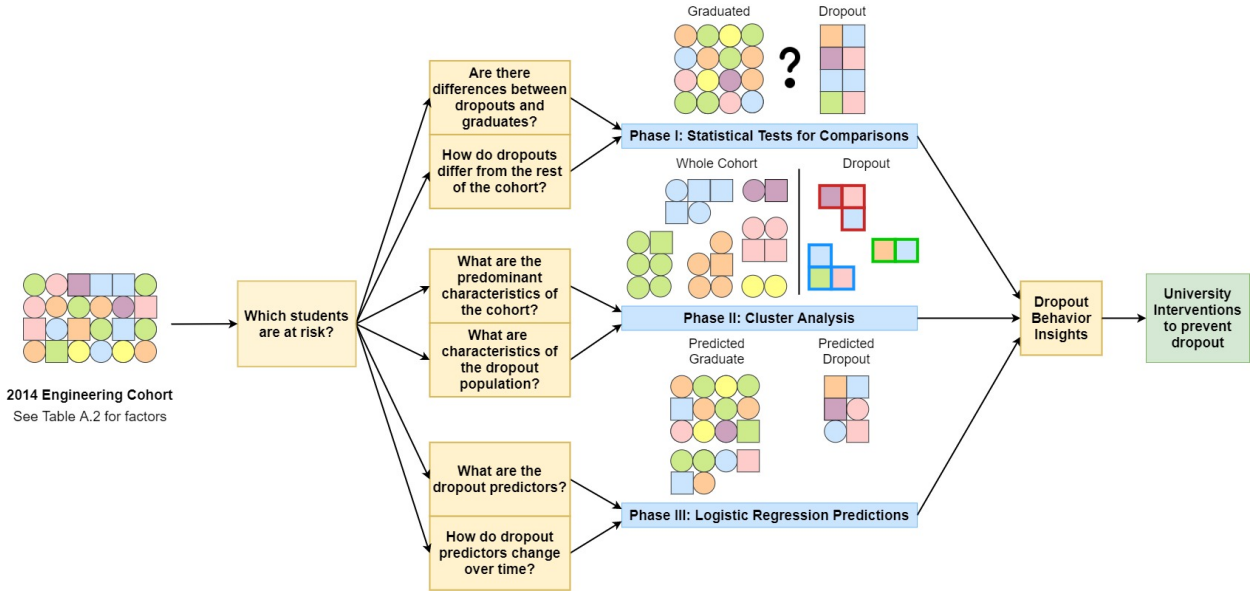


Fig 1. Process flow for the three-phase approach.

Phase 1: Statistical Analysis

Interval variables for the dropout and graduate populations were compared using the Kolmogrov-Smirnov (KS) test, Kruskal-Wallis test, and the two-sample z-test for comparing means. Binary variables were compared using two-sample z-tests for equality of proportions. Phase 1 variables indicated by an “x” in Table A.1 were recorded for the first term (fall 2014). Note that in this phase we are comparing the students who have dropped out to students who have graduated within six years. The students who are still working towards a degree have been excluded from this analysis.

Phase 2: Cluster Analysis

Correlation analysis was used to identify significant multicollinearity. Highly correlated variables were removed and all variables were normalized prior to cluster analysis. Preliminary hierarchical cluster analysis was performed using the variables indicated in the Phase 2 column of Table A.1. The Pseudo t^2 statistic, Pseudo F statistic, and Cubic Clustering Criterion (CCC) were compared to determine the optimal number of clusters to use in K-means cluster analysis. K-means cluster analysis was performed on both the entire cohort and the dropout population, creating seven clusters within the cohort and six clusters within the dropout population.

Phase 3: Logistic Regression

Stepwise logistic regression models were developed to predict probability of dropout within six years using five-fold cross validation. A significance level of 0.05 was used as the criteria for variables both entering and staying in the model. Three types of prediction models were built using the variables indicated by the Phase 3 column in Table A.1. Prediction model 1 includes all the Phase 3 variables recorded for the first term (first-year fall or fall 2014). Prediction model 2 includes all the variables from regression model 1 plus those same variables recorded for the second term (first-year spring or spring 2015). Likewise, Prediction model 3 includes all the variables from the first-year fall and first-year spring terms plus those variables recorded for the second-year fall term (fall 2015). All variables were normalized prior to building the models, and all models predicted the same response (dropout within six years). Figure 2 depicts the three predictions and the data used for each prediction.

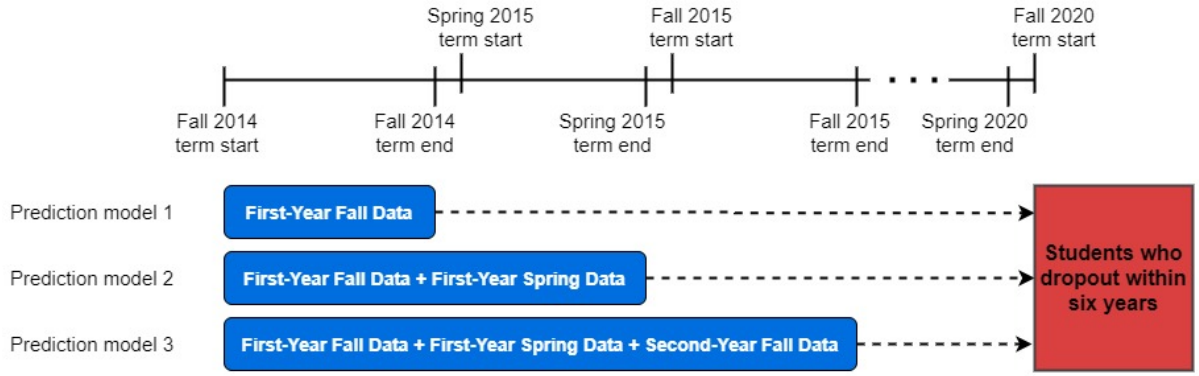


Fig 2. Data used in each of the three logistic regression models to predict dropout within six years. The data used for each term is listed in Table A.1.

To account for an imbalance in the low dropout response, the training set was balanced by under sampling the students who did not drop out. Training sets with dropout proportions of 20, 30, 40, and 50 percent were tested for each prediction model. All models were validated through five-fold cross-validation with validation sets representative of the dropout proportion in the original data. Five-fold cross validation produced five separate regression models for each training set. Dropout prediction rate and average validation area under the ROC curve (AUC) across these five models were used to select the best dropout proportion to use for each prediction. For the best model in each prediction, factors that appeared in at least three of the five models were presented as most important.

Results

Descriptive Statistical Analysis

A summary of the demographics and academic information collected for this cohort is shown in Table 2. These variables are referred to by the labels as they appeared in the data. American Indian/Alaska Native students and Native Hawaiian/Other Pacific Islander students were combined in Table 2 for the purposes of reporting. Each of these three groups was considered separately in analyses.

Table 2. Overview of demographics and academic information.

Variable	Cohort (n=1754)
Age at beginning of Fall 2014 term (years)	18.9 (±0.15)
Gender	
Male	76.7%
Female	23.3%
Race	
White	78.5%
Black or African American	3.8%
Asian	9.0%
American Indian/Alaska Native, Native Hawaiian/Other Pacific Islander, or Not specified	4.9%
Two or More Races	3.8%
Ethnicity	
Hispanic/Latino	4.9%
Student Admit Type	
Transfer	14.5%
New Freshman	85.5%
Residency	
In-State	81.4%
Out-of-State	18.6%
First Semester Enrolled	
Fall 2014	92.6%
Summer 2014	7.4%
Student Status	
Graduated within six years	86.1%
Dropout within six years	12.6%

A similar summary of demographics and academic information collected for the dropout population can be found in the Appendix (Table A.2). Table 3 displays a summary of the timing of dropout among the dropout population in the cohort. Most students drop out during the spring 2015 term, and students who drop out before year three compose approximately 63% of the entire dropout population.

Table 3. Number of dropouts in each term.

Term	Number of Dropouts
Fall 2014	22
Spring or Summer 2015	48
Fall 2015	29
Spring or Summer 2016	37
Fall 2016	23
Spring or Summer 2017	22
Fall 2017 to Summer 2018	18
Fall 2018 to Summer 2020	23

Phase 1: Statistical Testing for Comparisons

Factors with significant differences are shown in Table 4. Values denoted by an "x" were removed due to small sample size (less than eleven). All the factors with redacted values had a significantly higher proportion in the dropout population than in the graduated population. Note that part-time students include students who are enrolled in less than 12 credit hours (full-time credit hour requirement) at the specified time of data collection.

Table 4. Summary of results from KS, Kruskal-Wallis, two-sample z-test for means, and two-sample z-test for proportions, where "x" indicates values that cannot be shared due to small sample size. Descriptive percentages are shown with significance indicated next to the variable name. Variables are listed in decreasing order of statistical significance. Significance of interval variables was reported based on the maximum p-value among the three tests.

Variable	Dropout (n=222)	Graduated (n=1510)
First Term GPA***	2.50 (±.13)	3.36 (±.04)
Full-Time at End of Term***	89.40%	96.40%
Part-Time at End of Term***	10.60%	3.60%
Less Than Half at End of Term***	x	x
No Load at End of Term***	x	x
Full-Time at Census**	93.40%	97.00%
Part-Time at Census**	6.60%	3.00%
Less Than Half-Time at Census**	x	x
Male**	84.10%	75.40%
Female**	15.90%	24.60%
Two or More Races**	6.20%	3.20%
Half-Time at Census*	x	1.10%
African American*	7.00%	4.20%

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

First term GPA and academic load at the end of the term are the most significant differences between graduated and dropout populations. Moreover, first term GPA, proportion of full-time students, and proportion of females are significantly lower in the dropout population compared to the graduate population, whereas proportions of part-time students, male, multiracial and African American students are higher in the dropout population.

Phase 2: Cluster Analysis

From hierarchical cluster analysis, the optimal number of clusters to use in K-means cluster analysis was found to be seven. The seven clusters from the K-means clustering are represented by nine defining characteristics, which were ultimately used to label each cluster. The importance of academic load, GPA, gender, and race in the clusters is consistent with the findings in Phase 1. In Figure 3, GPA, academic load, and academic level are represented by their standardized values. Transfer students cluster separately from the new freshman students, as indicated by the low proportion of EFY students. Within new freshman and transfer clusters, students who start in the summer cluster separately from those who start in the fall. Moreover, out-of-state students and non-White students cluster separately among the new freshman students who start in the fall.

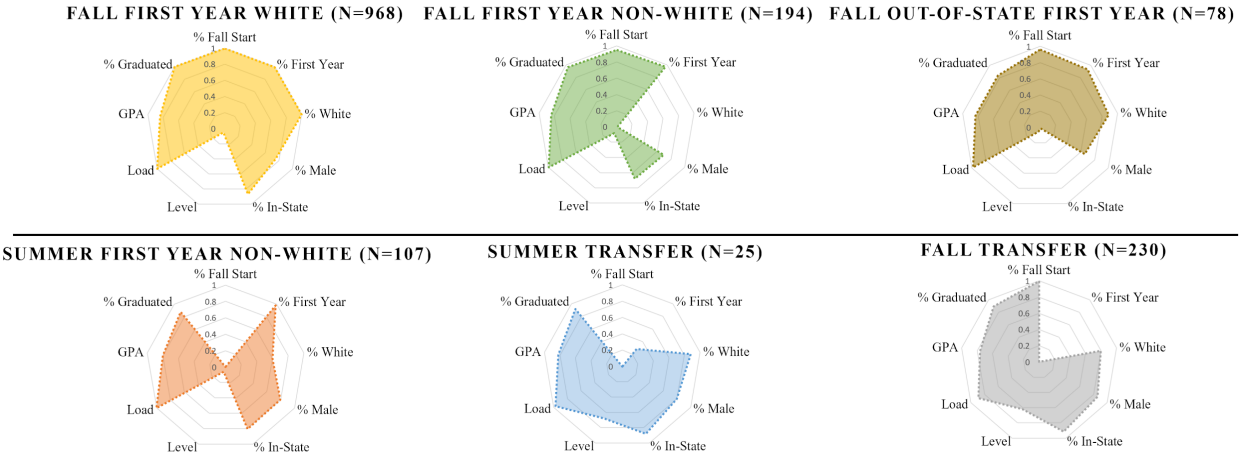


Fig 3a. Six of the seven clusters from the K-means cluster analysis for the entire cohort

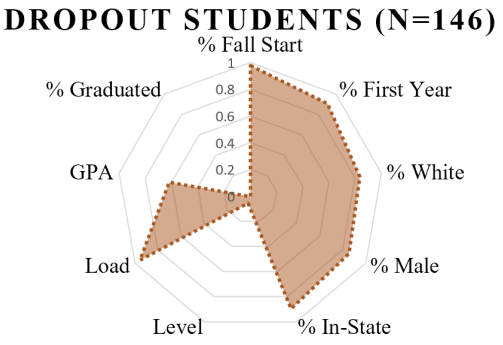


Fig 3b. Dropout cluster formed from the K-means cluster analysis for the entire cohort

While six of the seven clusters were dominated by graduated students, the "dropout cluster" consists only of students who dropout and 71 percent of all dropout students in the cohort. This cluster also has a lower average GPA compared to the other clusters.

Preliminary hierarchical cluster analysis on the dropout population (all 222 dropouts) identified six clusters as the optimal number of clusters to use in K-means cluster analysis for the dropout population. The K-means cluster analysis on the dropout population yielded the six clusters in Figure 2 represented by nine defining characteristics. Similar to the cohort clusters, GPA, academic level, gender, and race are also used to define the dropout clusters. Academic load, however, is not a defining characteristic among the dropout clusters. Instead, the time the dropout occurs relative to the third year characterizes these clusters. Within the students who dropped out before the third year, transfer students (as denoted by the low proportion of EFY), students who start in the summer, and out-of-state non-White students create separate clusters from the in-state White students.

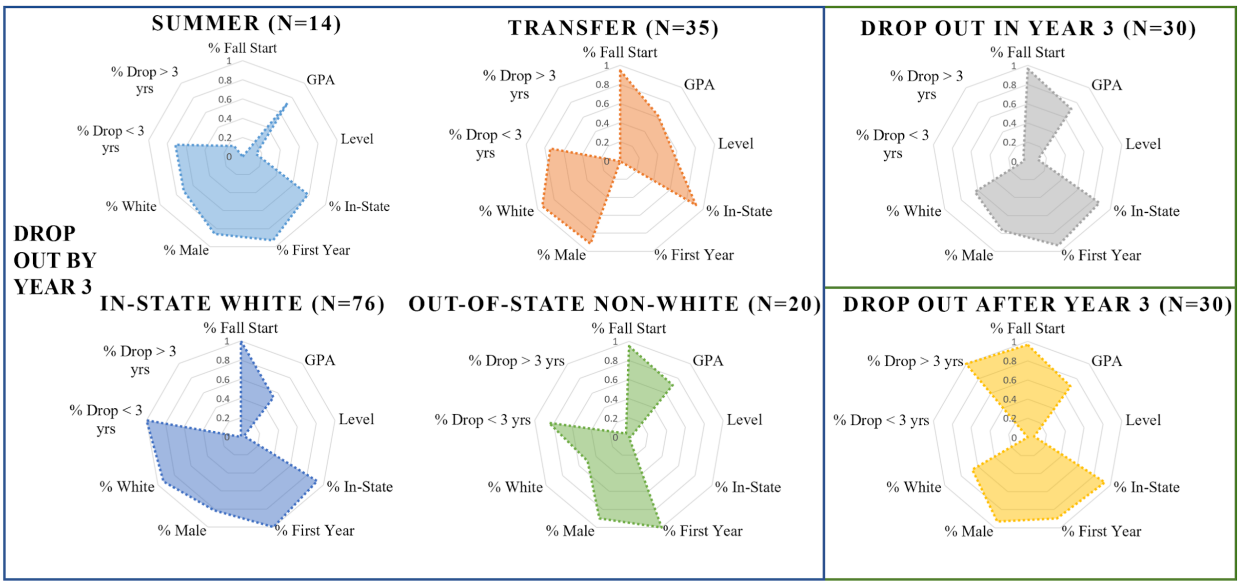


Fig 4. K-means clusters for dropout population

Phase 3: Regression Analysis

Recall that Prediction model 1 includes all the Phase 3 variables recorded for the first term. From Table 5, we see for prediction 1 the fully balanced model (50%) has a higher validation AUC than the other balanced models and a higher dropout prediction rate than any other model. This result for prediction 1 was similar for the first-year spring and second-year fall predictions (Tables A.3 and A.4 in Appendix). Thus, the fully balanced models were selected for all three predictions.

Table 5. Dropout prediction rate and validation AUC of unbalanced and balanced models for prediction 1.

Training Set	Validation AUC	% Dropout Predicted
Unbalanced	0.763	13.6
20% Dropout	0.665	16
30% Dropout	0.663	19.5
40% Dropout	0.639	22
50% Dropout	0.761	63.4

The distribution of significant factors that appeared in any of the 5 fully balanced models for the first-year fall prediction is shown in Figure 5a. Figures 5b-c show similar figures for predictions 2 and 3. Out of ten significant predictors that appeared across five models in prediction 1, the mechanical engineering indicator appears in a majority of the models, and GPA and passed credits appear in all models. Eleven significant predictors appeared in prediction 2, but none of the factors appeared in all the models. Enrolled credits in first term and second term GPA appeared in four of the models and passed credits in first term appeared in three models. In prediction 3, 8 predictors were significant. Of these eight, second term GPA appeared in all models, and indicators for EFY major, Hispanic, and Asian students appeared in most of the models.

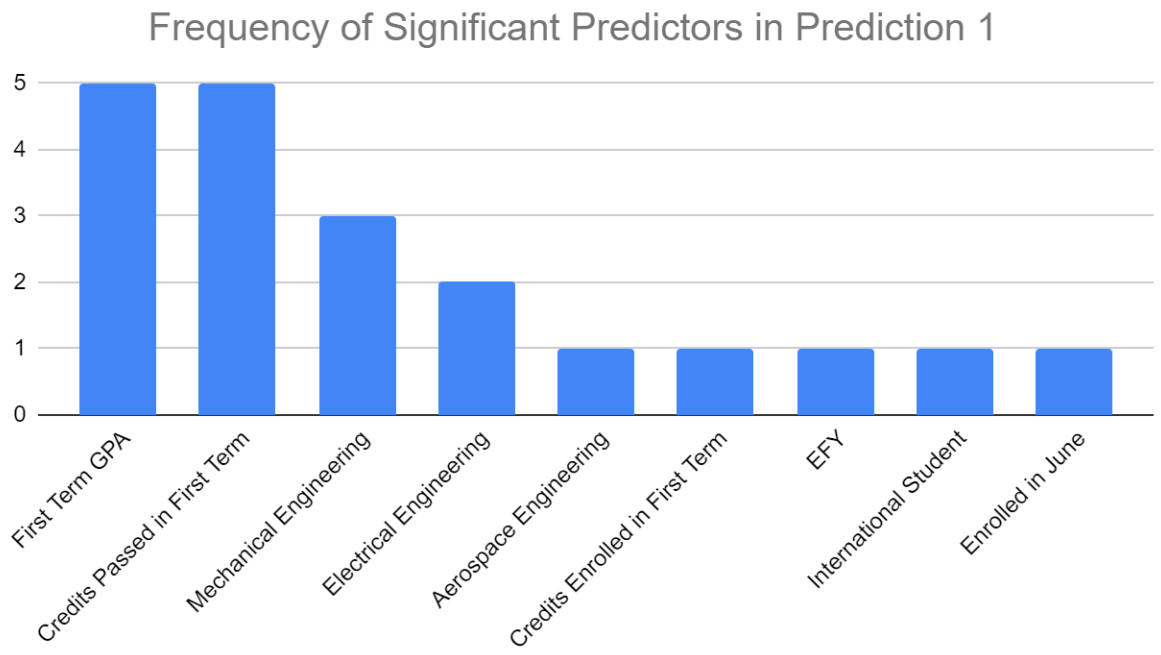


Fig 5a. Frequency of all significant predictors across five fully balanced models for prediction 1.

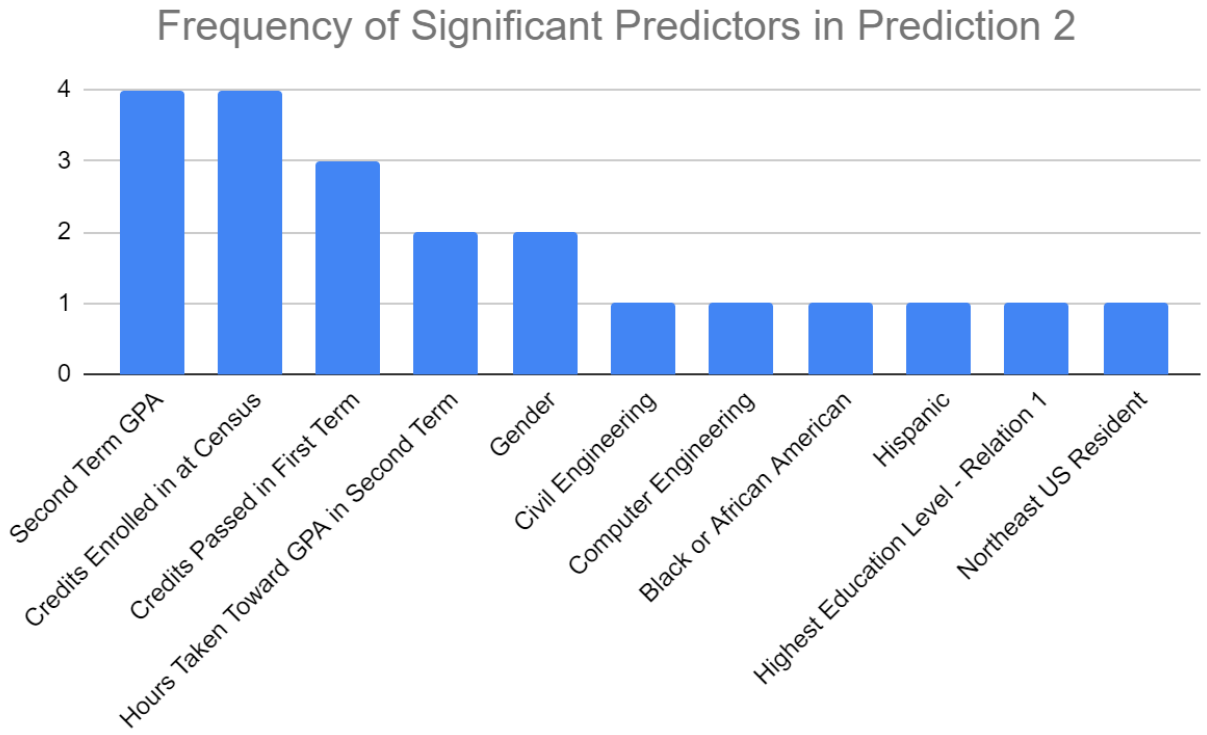


Fig 5b. Frequency of all significant predictors across five fully balanced models for prediction 2.

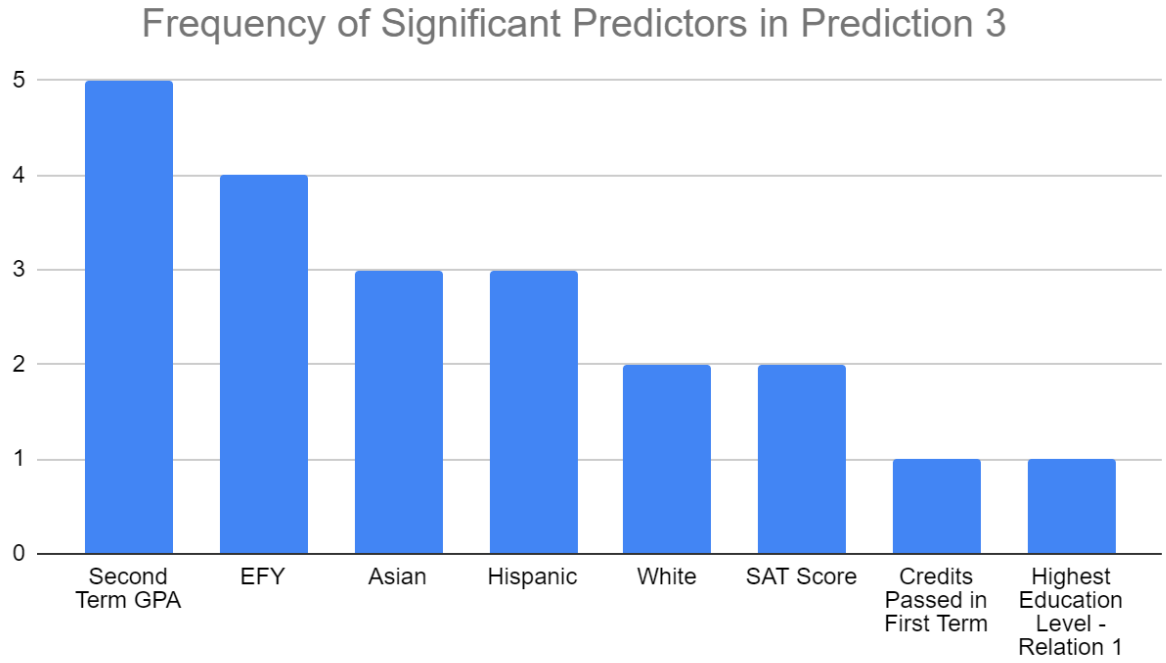
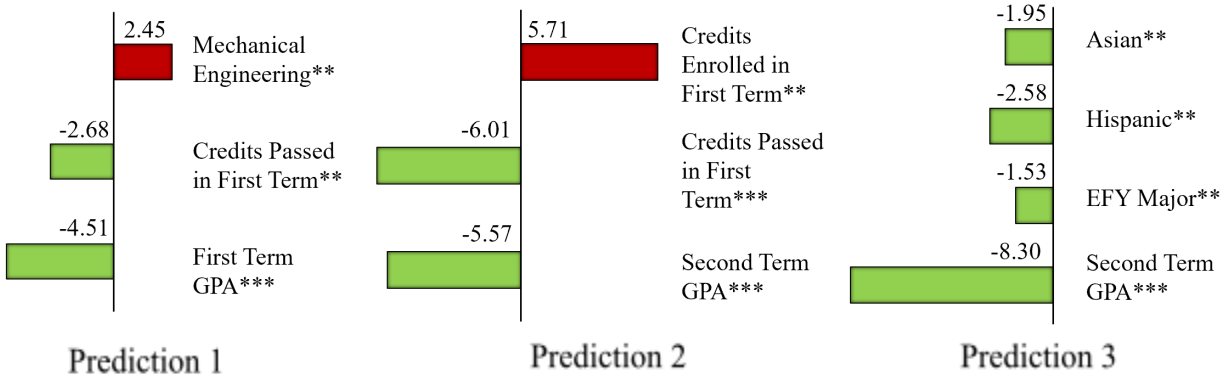


Fig 5c. Frequency of all significant predictors across five fully balanced models for prediction 3.

The average coefficient values for the most important factors across all five fully balanced models in each prediction are shown in Figure 6. GPA is highly significant for all three models. The number of credit hours passed in the first term significantly reduces the likelihood of dropping out for predictions 1 and 2. The number of credit hours a student is enrolled in on the census date of the first term also has a significant positive relationship with the likelihood of dropout in prediction 2. In prediction 1, mechanical engineering majors were more likely to drop out than other engineering majors. In prediction 3, Asian students and Hispanic students are less likely to drop out than other students. Furthermore, EFY majors were less likely to drop out than other engineering majors.



* p < 0.1, ** p <0.05, *** p <0.001

Figure 6. Average logistic regression coefficients for the most important predictors for the three fully balanced regression predictions.

Tables 6a-c present the contingency matrices for the three regression models. With only first term information, we were able to correctly predict 63 percent of the dropout students. Adding two more terms of information to the model increased the dropout prediction accuracy to 69 percent. Sensitivity to dropout increases as more information is used for predicting, while specificity is similar for the first two prediction models but lower for the third prediction. Prediction 2 has the highest AUC value among the models.

Table 6a. Contingency matrix for prediction 1 (n = 1745)

	Predicted Dropout	Predicted Graduate
Actual Dropout	130	75
Actual Graduate	330	1210
AUC = 0.761	Sensitivity ≈ 63%	Specificity ≈ 79%

Table 6b. Contingency matrix for prediction 2 (n = 1685)

	Predicted Dropout	Predicted Graduate
Actual Dropout	117	58
Actual Graduate	338	1172
AUC = 0.798	Sensitivity ≈ 67%	Specificity ≈ 78%

Table 6c. Contingency matrix for prediction 3 (n = 1510)

	Predicted Dropout	Predicted Graduate
Actual Dropout	79	35
Actual Graduate	414	982
AUC = 0.761	Sensitivity ≈ 69%	Specificity ≈ 70%

From Phase 2, it is apparent that the timing of the dropout relative to the third year is important. Table 7 compares the predictive performance for "early" dropout (dropout before the third year) and "late" dropout (dropout in the third year or later). The ability to predict early dropout increases from prediction 1 to prediction 2 but decreases from prediction 2 to the prediction 3. The ability to predict late dropout steadily increases from prediction 1 to prediction 3.

Table 7. Comparison of prediction accuracy for dropout before the third year and prediction accuracy for dropout in the third year or later.

Prediction Model	% Dropout before Year 3 Predicted Correctly	% Dropout Year 3 or Later Predicted Correctly
Prediction 1	70.8	50.7
Prediction 2	71.6	61.1
Prediction 3	67.3	71.2

Discussion

Are there differences between dropouts and graduates?

In Phase 1, we see that dropouts and graduates differ significantly in terms of first term GPA and academic load. The dropout population also had a higher proportion of male students than female students, which suggests that male engineering students are dropping out at a higher rate than their female peers.

How do dropouts differ from the rest of the cohort?

Academic load appears throughout the three phases of analysis. In the first phase, we learn that a higher percentage of part-time students, particularly those who are less-than-half-time at census or at the end of the term, dropout at a higher rate. Phase 2 also identifies academic load as a defining characteristic. In the third phase, we also see that as more credit hours are taken, particularly above 12 credit hours, students are more likely to drop out. This could suggest a non-linear relationship between academic load and likelihood of dropping out and differences in reasons for dropping out between full-time and part-time students. Part-time students typically have a weaker sense of belonging to the university and require better time management for balancing academics with other commitments (e.g., work, family) than full-time students which has been shown to affect dropout risk [19]–[21]. Simultaneously, costly education has driven more full-time students to work while in school, which can negatively affect the number of credit hours passed during a term and increase dropout risk if a student is working over 20 hours a

week [22]. Additionally, students who receive federal financial aid (loan, grants, work study) tend to finish sooner than students who are only receiving financial support from family or a private scholarship [23]. While we have not included financial aid status in our study, our findings may have implications for understanding dropout in students as a function of financial aid. This is an area for future study.

What are the predominant characteristics of the cohort?

Aside from dropping out, race, transfer status, and whether a student starts in the fall or summer appear to be the predominant characteristics defining student segments within the cohort. Hence, universities should consider these characteristics when personalizing interventions for different segments of the student population.

What are characteristics of the dropout population?

The importance of the timing of dropout is highlighted by the cluster analysis on the dropout population. Students who drop out during the third year cluster separately from those who drop out before the third year and from those who drop out after the third year. The significance of the third year may be related to the challenging curriculum often presented in the third year as students transition from first-year engineering to an engineering major taking more classes within their major. This third-year effect could also explain why the second-year prediction model finds that transfer students are more likely to drop out. Since many transfer students must be at least at a sophomore academic level in terms of completed credits, these students would be enrolled in a "third-year" curriculum in their second year as part of this cohort.

What are the dropout predictors?

Consistent with the literature, GPA is a significant predictor of dropout across all three phases[3], [9]. Moreover, the effect of the first term GPA (and later, the second-term GPA) continues to persist over time. Similarly, the credits passed in the first term has an important relationship with dropout risk, and this also persists over time. These two results together emphasize the importance of the first semesters in the program and of early intervention.

Race appears in both the statistical testing for comparison and cluster analysis as an important factor. However, race only appears as an important factor in the second-year logistic regression models (prediction 3). This suggests that other variables may be picking up some of the importance in the regression models. Our findings do show that African Americans may be at higher risk for dropout than other populations in the analysis such as Whites and Asians.

In contrast to Phase 1 and Phase 2, mechanical engineering was an important predictor in the first-year fall prediction in Phase 3. Many reasons could explain the importance of this predictor. For instance, students may initially choose mechanical engineering with limited knowledge of the field (or other engineering majors) and then find that mechanical engineering does not fit with their interests. The importance of mechanical engineering should be analyzed across other cohorts and universities to get a better understanding of its general influence on dropout.

How do the dropout predictors change over time?

Unlike predictions 1 and 2, all the important predictors in prediction 3 indicate students who are less likely to drop out. The importance of the EFY indicator suggests that transfer students are more likely to drop out. Race and ethnicity are only important factors in prediction 3 and highlight those students who are less likely to drop out. However, as shown in Figure 5b, race and ethnicity did appear as significant factors in one of the prediction 2 models. First year dropout is influenced by the number of credits passed in the first term and GPA, but race and ethnicity become more influential once students reach the second year.

Since most of the students who eventually drop out tend to drop out before the third year, prediction in the first two years is critical. Our prediction models further suggest that prediction in the first term is especially critical. The number of credit hours passed in the first term has a greater influence on dropout risk after adding data from the first-year spring term. Similarly, the number of credits enrolled in for the first-year fall term was not important in prediction but is important in prediction 2 after adding data for another term. This relationship could be a surrogate for non-academic factors such as self-efficacy or confidence, which have been found to be significant predictors of dropout in other studies [14], [15]. For instance, students who pass more credits in the first term might feel more confident in their academic performance in the second term than students who enroll in a higher number of credits but do not perform as well or pass as many. The importance of these first-term factors suggest that behavior exhibited in the first term still affects a student's dropout risk in later terms.

According to Table 6b, prediction 2 is best at predicting dropout. With nearly 50 dropouts occurring in the second term (Table 3), the number of dropouts that occurred in the third term or later is much smaller and thus reduces the fully balanced training set size for prediction 3. This could explain the decrease in overall prediction accuracy for prediction 3. Moreover, as the proportion of dropout students in the cohort decreases with each term, the fully balanced model becomes more prone to overpredicting the dropout response. This could explain the decrease in specificity for prediction 3 and the higher AUC for the unbalanced model compared to all of the balanced models. Though the fully balanced training sets tend to overpredict the dropout response, underpredicting the dropout response could prevent students at risk of dropping out from being identified as someone in need of help and from eventually receiving the help they need.

Limitations and Future Work

While this study presents a case study of a single cohort of students from a single institution, we believe the framework is generalizable and may be applied to additional cohorts and institutions. The performance of the framework in this study may be limited by the information that had been accessible at the time and the size of the sample. Financial aid information will be included in future work, as well as other data (housing, high school information, and psychological factors, etc.) as they become available. Additional cohorts can be added to increase the sample size and training set for the predictions, which is especially needed for prediction 3 when almost a third of dropouts have already occurred. Since this study only presents results from a single cohort, the results may not be generalizable to engineering cohorts at other universities or from other

academic years. However, many of our findings (e.g., the importance of GPA and the timing of dropout) agree with other findings in the literature and are likely generalizable.

Conclusion

With the framework described in this paper, we can identify factors that distinguish dropout students from those who graduate only using data from the first term. We then use a set of prediction models-- each updated with additional data from the next term-- to verify which of these characteristics are significant predictors of dropout and to understand how the significance of these factors changes over time.

Surprisingly, we find that the information available from the first semester enables identification of a large proportion of the at-risk students. Most dropouts occur in the first-year spring term, and students who dropout around the same time relative to the third year behave similarly. Academic load and GPA are consistent predictors of dropout; however, race does not become a significant factor until the second year which suggests influential factors of dropout can change over time. Furthermore, first-term GPA and credits passed in the first-year fall term are still significant after adding information from the first-year spring term. Mechanical engineering is a significant predictor of dropout for this cohort, but this factor should be explored across other cohorts and universities. The overall power of prediction will likely be enhanced further as additional data sources are obtained for analysis. These findings also show the importance of activities that occur during the first year at a university.

While influential factors identified in the three phases are consistent with the current literature, our findings highlight *when* these factors are significant. This further suggests the potential for dynamic dropout prediction models and highlights the importance of fine-tuning dropout models for specific populations within a cohort and for specific time periods. This work also sets the foundation for the deeper work where we look at specific types of interventions and simulate the results of applying interventions and personalized approaches.

Acknowledgements

We would like to acknowledge the Matriculation and Well-Being Under Emergent Events (MWEE) NSF RAPID collaborative (NSF Award Number: 2040072) for sponsoring this work.

References

[1] "The NCES Fast Facts Tool provides quick answers to many education questions (National Center for Education Statistics)," Accessed: Apr. 18, 2021. [Online].

[2] J. Roy, "EnginEEring by thE numbErs." Accessed: Mar. 06, 2021. [Online]. Available: www.asee.org/colleges.

[3] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting Student Dropout in Higher Education," 2017. Accessed: Jan. 13, 2021. [Online].

[4] Y. Chen, A. Johri, and H. Rangwala, "Running out of STEM: A Comparative Study across STEM Majors of College Students At-Risk of Dropping Out Early," p. 10, doi: 10.1145/3170358.3170410.

[5] C. Blanco *et al.*, "Mental health of college students and their non-college-attending peers: Results from the national epidemiologic study on alcohol and related conditions," *Archives of General Psychiatry*, vol. 65, no. 12, pp. 1429–1437, Dec. 2008, doi: 10.1001/archpsyc.65.12.1429.

[6] W. S. Slutske, "Alcohol use disorders among US college students and their non-college-attending peers," *Archives of General Psychiatry*, vol. 62, no. 3, pp. 321–327, Mar. 2005, doi: 10.1001/archpsyc.62.3.321.

[7] L. Seamster and R. Charron-Chénier, "Predatory inclusion and education debt: Rethinking the racial wealth gap," *Social Currents*, vol. 4, no. 3, pp. 199–207, Jun. 2017, doi: 10.1177/2329496516686620.

[8] A. B. Abad, "Paying the Price: College Costs, Financial Aid, and the Betrayal of the American Dream by Sara Goldrick-Rab," *The Review of Higher Education*, vol. 42, no. 1, p. E-7-E-10, 2018, doi: 10.1353/rhe.2018.0041.

[9] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in *Procedia Computer Science*, Jan. 2015, vol. 72, pp. 414–422, doi: 10.1016/j.procs.2015.12.157.

[10] N. Kronberger and I. Horwath, "The Ironic Costs of Performing Well: Grades Differentially Predict Male and Female Dropout From Engineering," *Basic and Applied Social Psychology*, vol. 35, no. 6, pp. 534–546, Nov. 2013, doi: 10.1080/01973533.2013.840629.

[11] G. Dewantoro and N. Ardisa, "A Decision Support System for Undergraduate Students Admissions using Educational Data Mining," in *7th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2020 - Proceedings*, Sep. 2020, pp. 105–109, doi: 10.1109/ICITACEE50144.2020.9239244.

[12] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa, "University student retention: Best time and data to identify undergraduate students at risk of dropout," *Innovations in Education and Teaching International*, vol. 57, no. 1, pp. 74–85, Jan. 2020, doi: 10.1080/14703297.2018.1502090.

[13] R. M. Marra, K. A. Rodgers, D. Shen, and B. Bogue, "Leaving Engineering: A Multi-Year Single Institution Study," *Journal of Engineering Education*, vol. 101, no. 1, pp. 6–27, Jan. 2012, doi: 10.1002/j.2168-9830.2012.tb00039.x.

[14] T. Blaetz, “Paper 788-2017 Examining Higher Education Performance Metrics with SAS ® Enterprise Miner ™ and SAS ® Visual Analytics ™,” 2017. Accessed: Mar. 02, 2021. [Online].

[15] M. Hedayetul, I. Shovon, and M. Haque, “An Approach of Improving Student’s Academic Performance by using K-means clustering algorithm and Decision tree,” 2012. Accessed: Mar. 03, 2021. [Online]. Available: www.ijacsa.thesai.org.

[16] J. Watkins and E. Mazur, “Retaining Students in Science, Technology, Engineering, and Mathematics (STEM) Majors,” *Journal of College Science Teaching*, vol. 42, no. 5, pp. 36–41, 2013, Accessed: Mar. 06, 2021. [Online]. Available: https://www.jstor.org/stable/43631580?seq=2#metadata_info_tab_contents.

[17] B. N. Geisinger and D. R. Raman, “Why They Leave: Understanding Student Attrition from Engineering Majors Why They Leave: Understanding Student Attrition from Engineering Majors*,” 2013. Accessed: Mar. 06, 2021. [Online]. Available: http://lib.dr.iastate.edu/abe_eng_pubs.

[18] C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker, *Handbook of Educational Data Mining*. Boca Raton: CRC Press, Taylor & Francis Group, 2011.

[19] S. Quaye, S. Harper, and S. Pendakur, *Student Engagement in Higher Education: Theoretical Perspectives and Practical Approaches for Diverse Populations*, Third. New York: Routledge, Taylor & Francis, 2019.

[20] J. C. K. Yum, D. Kember, and I. Siaw, “Coping mechanisms of part‐time students,” *International Journal of Lifelong Education*, vol. 24, no. 4, pp. 303–317, Jul. 2005, doi: 10.1080/02601370500169194.

[21] C. MacCann, G. J. Fogarty, and R. D. Roberts, “Strategies for success in education: Time management is more important for part-time than full-time community college students,” *Learning and Individual Differences*, vol. 22, no. 5, pp. 618–623, Oct. 2012, doi: 10.1016/j.lindif.2011.09.015.

[22] E. Hovdhaugen, N. Frølich, and P. O. Aamodt, “Informing Institutional Management: institutional strategies and student retention,” *European Journal of Education*, vol. 48, no. 1, pp. 165–177, Mar. 2013, doi: 10.1111/ejed.12002.

[23] D. Glocker, “The effect of student aid on the duration of study,” *Economics of Education Review*, vol. 30, no. 1, pp. 177–190, Feb. 2011, doi: 10.1016/j.econedurev.2010.08.005.

Appendix

Table A.1. Complete list of variables in original dataset.

Variable	Label	Type	Phase 1	Phase 2	Phase 3
ADMIT_2146	Admitted in June 2014	Binary	x	x	x
ADMIT_2147	Admitted in July 2014	Binary	x	x	x
ADMIT_2148	Admitted in Fall 2014	Binary	x	x	x
STRM_FIRST_2146	First term enrolled was June 2014	Binary	x	x	x
STRM_FIRST_2147	First term enrolled was July 2014	Binary	x	x	x
STRM_FIRST_2148	First term enrolled was Fall 2014	Binary	x	x	x
EOT_NC_REG_STATUS	Registration Enrolled Status	Nominal	x	x	x
EOT_CUM_GPA	Cumulative Term GPA for Career.	Interval	x	x	
EOT_CUR_GPA	Current Term GPA	Interval			x
NC_GENDER	Gender	Binary	x	x	x
NC_EG1	Reported Ethnic Group 1 (White)	Binary	x	x	x
NC_EG2	Reported Ethnic Group 2 (Black or African American)	Binary	x	x	x
NC_EG3	Reported Ethnic Group 3 (Hispanic/Latino)	Binary	x	x	x
NC_EG4	Reported Ethnic Group 4 (Asian)	Binary	x	x	x
NC_EG5	Reported Ethnic Group 5 (American Indian or Alaska Native)	Binary	x	x	x
NC_EG6	Reported Ethnic Group 6 (Not Specified)	Binary	x	x	x
NC_EG7	Reported Ethnic Group 7 (Native Hawaiian or Other Pacific Islander)	Binary	x	x	x
TUITION_RES	Tuition Residency	Binary	x	x	x
ACAD_SUBPLAN_FLG	Student has academic subplan	Binary	x	x	x
EOT_TOT_CUMULATIVE	Total Cumulative Units at End of Term	Interval	x	x	x
EOT_TOT_PASSD_GPA	Total Passed Toward GPA at End of Term	Interval	x	x	x
EOT_TOT_PASSD_NOGPA	Total Passed Not Toward GPA at End of Term	Interval	x		
EOT_TOT_PASSD_PRGRSS	Total Passed Not Toward GPA at End of Term	Interval	x		
EOT_TOT_TAKEN_GPA	Total Taken Toward GPA at End of Term	Interval	x	x	x
EOT_TOT_TAKEN_NOGPA	Total Taken Not Toward GPA at End of Term	Interval	x		
EOT_TOT_TAKEN_PRGRSS	Total Taken for Progress at End of Term	Interval	x		

EOT_UNT_AUDIT	Units Audited at End of Term	Interval	x		
EOT_UNT_TAKEN_PRGRSS	Units Taken for Progress for Term at End of Term	Interval	x		
EOT_TRF_PASSED_GPA	Transfer Passed for GPA at End of Term	Interval	x		
EOT_TRF_PASSED_NOGPA	Transfer Passed Not for GPA at End of Term	Interval	x	x	x
EOT_UNT_TAKEN_GPA	Units Taken Toward GPA for Term at End of Term	Interval	x		
EOT_UNT_TAKEN_NOGPA	Units Taken Not Toward GPA for Term at End of Term	Interval	x		
NC_EXT_TRF_NOGPA	External Transfer Credit Taken not for GPA	Interval	x		
ACAD_LEVEL_BOT	Academic Level - Term Start	Interval	x	x	x
AEBS	Aerospace Engineering Major	Binary	x	x	x
BMEBS	Biomedical Engineering Major	Binary	x	x	x
CEBS	Civil Engineering Major	Binary	x	x	x
CEMBS	Construction Engineering and Management Major	Binary	x	x	x
CHEBS	Chemical Engineering Major	Binary	x	x	x
CPEBS	Computer Engineering Major	Binary	x	x	x
EEBS	Electrical Engineering Major	Binary	x	x	x
EFY	Engineering First Year	Binary	x	x	x
EGRBS	Engineering Major	Binary	x	x	x
ENEBS	Environmental Engineering Major	Binary	x	x	x
IEBS	Industrial and Systems Engineering Major	Binary	x	x	x
MEBS	Mechanical Engineering Major	Binary	x	x	x
MSEBS	Materials Science and Engineering Major	Binary	x	x	x
NEBS	Nuclear Engineering Major	Binary	x	x	x
AGE	Age	Interval	x	x	x
NC_HIGH_ED_REL1	Highest Education Level - Relationship 1	Interval	x	x	x
NC_HIGH_ED_REL2	Highest Education Level - Relationship 2	Interval	x	x	x
NORTHEAST_US	Permanent Address in Northeast US Region	Binary	x	x	x
MIDWEST_US	Permanent Address in Midwest US Region	Binary	x	x	x
SOUTH_US	Permanent Address in South US Region	Binary	x	x	x
WEST_US	Permanent Address in West US Region	Binary	x	x	x
OUTSIDE_US	Permanent Address Outside US	Binary	x	x	x
CNSS_TOT_CUMULATIVE	Total Cumulative Units at Census Date	Interval	x		

CNSS_TOT_PASSD_GPA	Total Passed Toward GPA at Census Date	Interval	x		
CNSS_TOT_PASSD_NOGPA	Total Passed Not Toward GPA at Census Date	Interval	x		
CNSS_TOT_PASSD_PRGRSS	Total Passed for Progress at Census Date	Interval	x		
CNSS_TOT_TAKEN_GPA	Total Taken Toward GPA at Census Date	Interval	x		
CNSS_TOT_TAKEN_NOGPA	Total Taken Not Toward GPA at Census Date	Interval	x		
CNSS_TOT_TAKEN_PRGRSS	Total Taken for Progress at Census Date	Interval	x		
CNSS_UNT_AUDIT	Units Audited at Census Date	Interval	x		
CNSS_UNT_TAKEN_PRGRSS	Units Taken for Progress for Term at Census Date	Interval	x		
CNSS_ACADEMIC_LOAD	Academic Load at Census Date	Nominal	x	x	x
CNSS_ACADEMIC_LOAD_F	F - Enrolled Full-Time at Census Date	Binary	x		
CNSS_ACADEMIC_LOAD_T	T - Three Quarter Time at Census Date	Binary	x		
CNSS_ACADEMIC_LOAD_H	H - Enrolled Half-Time at Census Date	Binary	x		
CNSS_ACADEMIC_LOAD_L	L - Less than Half-Time at Census Date	Binary	x		
CNSS_ACADEMIC_LOAD_N	N - No Unit Load at Census Date	Binary	x		
EOT_ACADEMIC_LOAD	Academic Load at End of Term	Nominal	x	x	x
EOT_ACADEMIC_LOAD_F	F - Enrolled Full-Time at End of Term	Binary	x		
EOT_ACADEMIC_LOAD_T	T - Three Quarter Time at End of Term	Binary	x		
EOT_ACADEMIC_LOAD_H	H - Enrolled Half-Time at End of Term	Binary	x		
EOT_ACADEMIC_LOAD_L	L - Less than Half-Time at End of Term	Binary	x		
EOT_ACADEMIC_LOAD_N	N - No Unit Load at End of Term	Binary	x		
EOT_UNT_PASSD_PRGRSS	Units Passed for Progress for Term at End of Term	Interval	x		
NC_SAT_TOTAL	SAT Composite Score	Interval	x	x	x
DEGREE_CHANGE	Student changed Degree Programs	Binary	x	x	x
MULTIRACIAL	Student in Two or More Races	Binary	x	x	x
DC_IN_SPRING	Dropout occurred in Spring Term	Binary	x	x	
DC_AFTER_4YRS	Dropout occurred after four years	Binary	x	x	

DC_AFTER_3YRS	Dropout occurred after three years	Binary	x	x	
DC_AFTER_2YRS	Dropout occurred after two years	Binary	x	x	
DC_BY_1YR	Dropout occurred within the first year	Binary	x	x	
DC_BY_2YR	Dropout occurred within the second year	Binary	x	x	
GRAD_FLG	Student graduated	Binary	x		
DROPOUT_FLG	Student dropped out	Binary	x	x	x
EOT_PASSD_GPA	Units Passed Toward GPA at End of Term	Interval	x	x	x
EOT_PASSD_NOGPA	Units Passed Not Toward GPA at End of Term	Interval	x	x	x

Table A.2. Overview of demographics and academic information for the dropout population.

Variable	Dropout(n=222)
Age	19.4(.59)
Gender	
Male	84.1%
Female	15.9%
Student Admit Type	
Transfer	18.9%
New Freshman	81.1%
Race	
White	72.7%
Black or African American	5.3%
Asian	8.4%
Native Hawaiian or Alaska Native, American Indian, Other or Not specified	7.0%
Two or More Races	6.6%
Ethnicity	
Hispanic/Latino	6.6%
Residency	
In-State	79.7%
Out-of-State	20.3%
First Semester Enrolled	
Fall	93.0%
Summer	7.0%

Table A.3. Dropout prediction rate and validation AUC of unbalanced and balanced models for the first-year spring prediction.

Training Set	Validation AUC	% Dropout Predicted
Unbalanced	0.814	21.2
20% Dropout	0.762	26.3
30% Dropout	0.736	33.0
40% Dropout	0.750	45.3
50% Dropout	0.798	65.4

Table A.4. Dropout prediction rate and validation AUC of unbalanced and balanced models for the second-year fall prediction.

Training Set	Validation AUC	% Dropout Predicted
Unbalanced	0.783	11.9
20% Dropout	0.702	12.7
30% Dropout	0.670	11.0
40% Dropout	0.659	12.7
50% Dropout	0.761	66.9