A Decision Tree Analysis of Freshman Students' Completion Rates at The University of Southern Mississippi

Supervised by Dr. Bernd Schroeder

Cole Springer

Anmol Chapagain

An undergraduate research project

Funded by Wright W. and Annie Rea Cross Mathematics Undergraduate Research Scholarship

Department of Mathematics

University of Southern Mississippi

United States

April 23rd, 2021

Contents

List of Figures

1 Problem Statement

As society continues to put a premium on education the pressure that individuals feel to get a college education continues to increase. While the need for an education is felt all throughout society (Pew Research Center, 2014), the ability to successfully complete the requirements for a college degree are not quite as uniform as we would hope. Some students looking to complete a degree face challenges that are not obstacles to other prospective students and sometimes these extra challenges are completely out of the student's control. This in turn puts higher education institutions in a spot where they want to have as many of their accepted students pass and complete their education while only having limited resources to support their student body's need for such support. Therefore, for academic institutions, identifying the factors that hinder or support a student's ability to succeed in their educational pursuit is a step in the direction of allocating resources as efficiently as possible while having the greatest positive impact.

2 Research Outcomes

The research into what factors play a role in academic success or failure is important for schools to know as they figure out how to counteract factors that negatively impact their student's ability to graduate. Currently, institutions have a few main ways of supporting students that include academic tutoring for students that need help in their classes, financial assistance for those that are struggling to cover tuition and necessary expenses, and counseling for those that feel like they need guidance or just someone to talk to.

3 Available Data

For this research we will be working with data that is provided by the Office of Institutional Research (OIR) at the University of Southern Mississippi (USM). The data that is received from the OIR will be sorted into two groups based on whether it was available before or after the student's enrollment. The data supplied that was available before a student's enrollment consists of the following factors: gender, race, state residency status, national residency status, high school GPA, age entering the first term, Pell Grant Eligibility, composite ACT score, Mathematics ACT score, English ACT score, Reading ACT score, and Science ACT score. Data available for factors after a student's enrollment consists of academic college, first term load status, first term GPA, Greek Life Status, Honors College or Luckyday status, last term load status, last term of enrollment, and completion.

4 Decision Trees

Decision trees are a type of model that show a branching "tree" of conditions and their respective outcomes. It consists of a series of nodes, branches, and leaves where nodes represent categorical attributes that we are using to separate the data, branches represent the probability of each potential destination from a node, and leaves are final destination outcomes which represent complete data separation and collection. Each node can be considered as an event that occurs where each node on lower tiers is dependent on the decision outcomes of nodes above them based on the probability of each event occurring. This type of model gives sequential outcomes based on the conditions it has for its attributes and this allows for data that passes through the decision tree to be easily categorized and studied in a convenient way for this research project.

5 Partykit's "ctree" Partitioning

Partykit is a package for the R programming language that contains tools for data analysis. One of these tools is the ctree function and this function attempts to apply a recursive binary partitioning algorithm on data that is provided to it. The algorithm first assigns case weights to each of the $\mathbf{n}$ data points. Then there is an attempt to determine if, for any $\mathbf{j}$, the response $\mathbf{Y}$ and $\mathbf{X}_j$ are not independent. If, for all covariates $\mathbf{X}_j$, independence between $\mathbf{X}_j$ and the response $\mathbf{Y}$ cannot be rejected, then an output of an "empty" tree is given where none of the data is able to be partitioned. If the algorithm is able to find one or more dependent $\mathbf{X}_j$ then it picks the one with the strongest relationship with the response variable $\mathbf{Y}$. Step two of the algorithm splits the chosen covariate $\mathbf{X}_j$ into non overlapping sets where the data in the new sets are given new case weights. Steps one and two of the algorithm are repeated until there are no longer any dependent $\mathbf{X}_j$ remaining at which point the partitioning of the data is complete.

6 Data Processing

The data received from the University of Southern Mississippi for this research project was not completely ready to be used for analysis as given. Some of the data, such as GPA ranges and race, needed to be translated into something that can be used by the analysis functions in R. There were also some categories that contained some empty cells (missing data) which needed to be accounted for as well. To fix these issues, a function was created to assign non-numerical data a number based on the attributes that appear in a given category of data. The function was also used to properly order the attributes in a way that the outcome could easily be read from the decision tree. For example, the data available gave a range that a particular student fell into in regard to their GPA, i.e. 2.50-2.74. For this particular GPA the function would assign 2.74 so that when ran through the ctree partitioning the resulting tree, assuming that 2.74 is considered a good partition point, would partition based on whether or not someone's GPA was greater than 2.74 or less than or equal to 2.74.

7 Results

The findings are reported in graphs and tables as they appear in R.

7.1 Factors before freshmen year

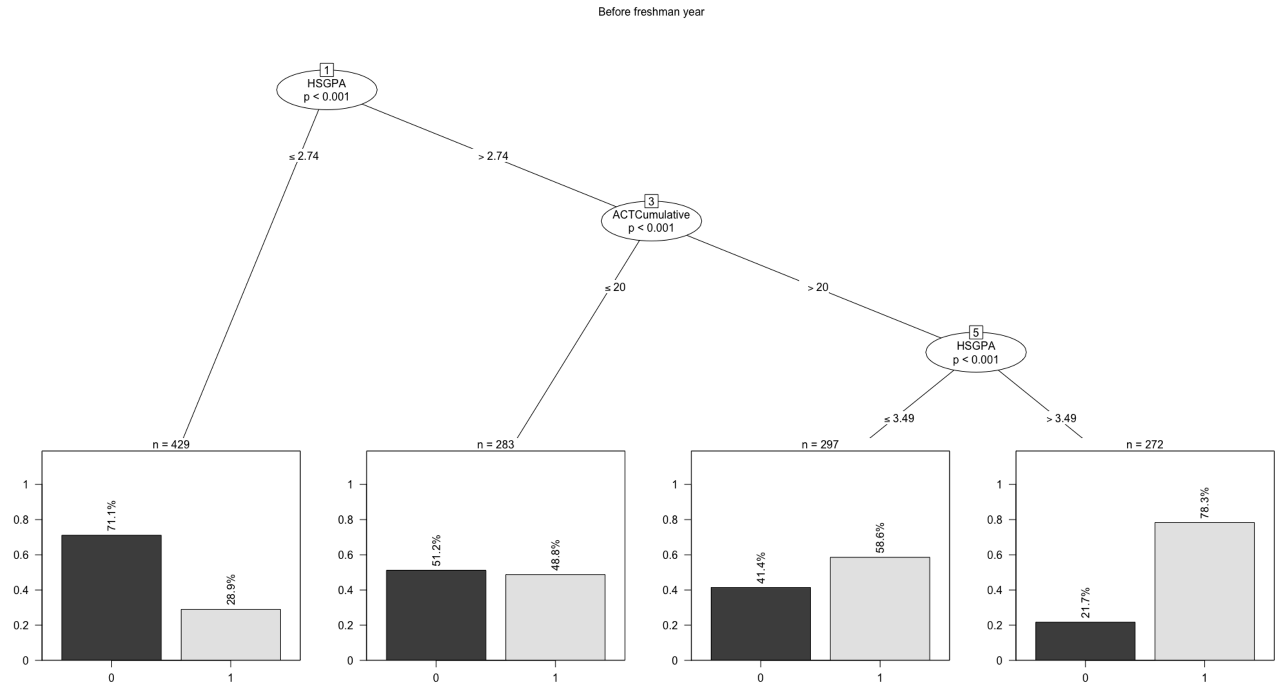
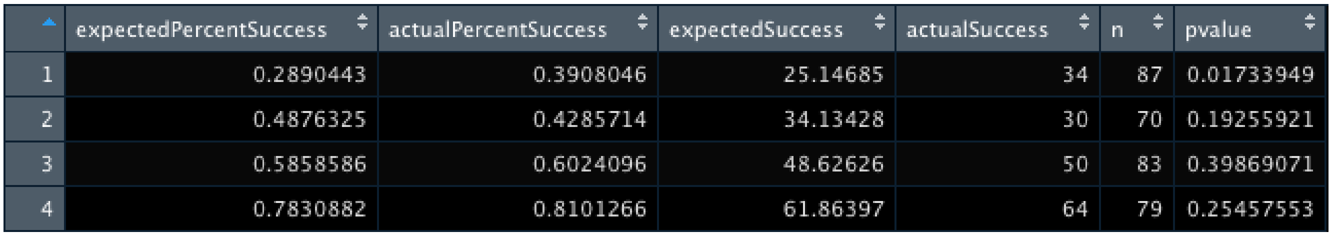


Figure 1: Decision tree formed by the attributes of Gender, High School GPA, ACT Sci, ACT Eng, ACT Math, ACT Reading, ACT Cumulative, Ethnicity, In State Residency, Age, Citizenship

The factors that were the part of the analysis were gender, high school GPA, all types of ACT scores, ethnicity, status of residency, age, and citizenship status. Among those factors only HSGPA and ACT Cumulative were included in the tree due to their significance. HSGPA is included in the root node which shows that it is the factor with most influence on graduation.



	expectedPercentSuccess	actualPercentSuccess	expectedSuccess	actualSuccess	n	pvalue
1	0.2890443	0.3908046	25.14685	34	87	0.01733949
2	0.4876325	0.4285714	34.13428	30	70	0.19255921
3	0.5858586	0.6024096	48.62626	50	83	0.39869071
4	0.7830882	0.8101266	61.86397	64	79	0.25457553

Figure 2: Validation table of before freshman year decision tree

The validation table shows that one of the bins has a p value less than .05 and thus can be considered inaccurate. The rest of the bins can be considered accurate based on their respective p values.

7.2 Factors during freshman year

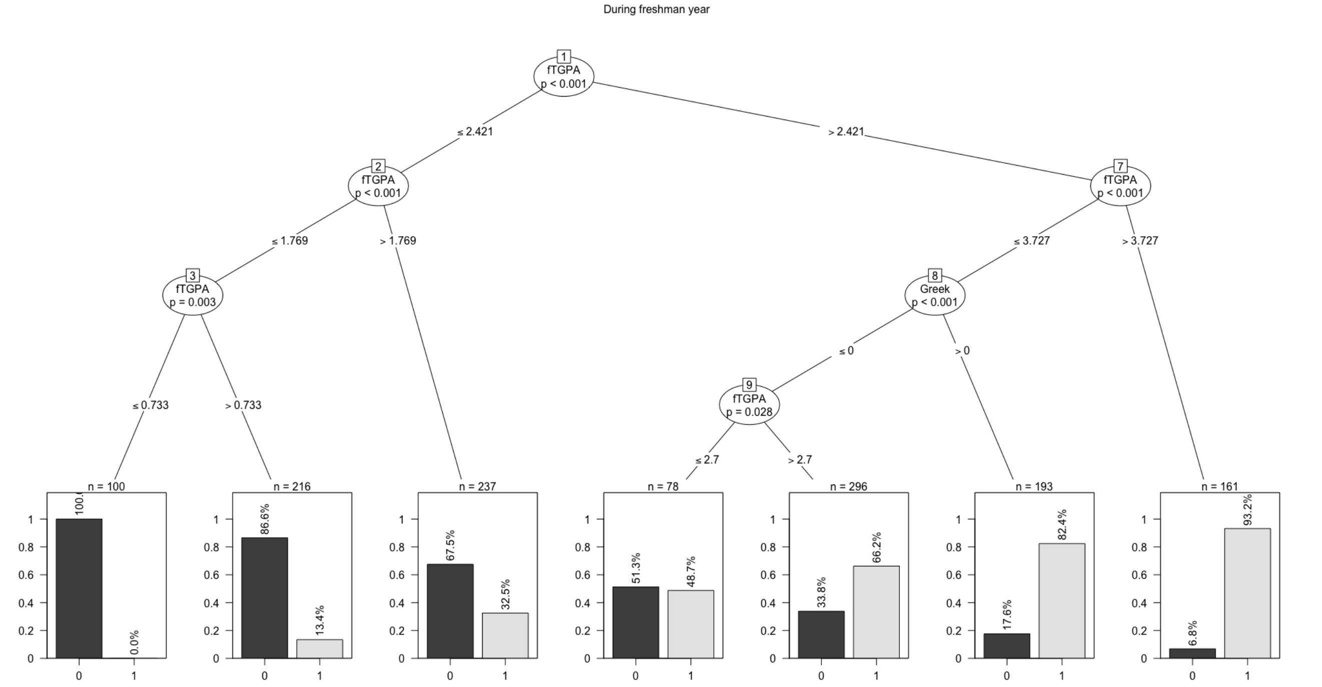
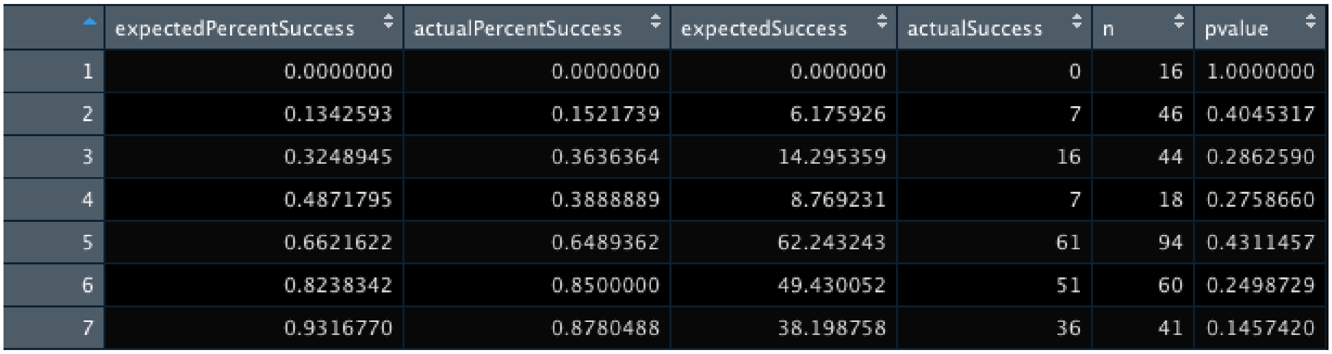


Figure 3: Decision tree formed by the attributes of College, First Term Course Load, First Term Pell Status, Honors & Lucky Daye Status, First Term GPA, and Greek Status

In this decision tree a total of six factors were used in the model to determine the success but only two factors are displayed by the tree. First term GPA and Greek status were considered as significant factors to determine the success of graduation.



	expectedPercentSuccess	actualPercentSuccess	expectedSuccess	actualSuccess	n	pvalue
1	0.0000000	0.0000000	0.000000	0	16	1.0000000
2	0.1342593	0.1521739	6.175926	7	46	0.4045317
3	0.3248945	0.3636364	14.295359	16	44	0.2862590
4	0.4871795	0.3888889	8.769231	7	18	0.2758660
5	0.6621622	0.6489362	62.243243	61	94	0.4311457
6	0.8238342	0.8500000	49.430052	51	60	0.2498729
7	0.9316770	0.8780488	38.198758	36	41	0.1457420

Figure 4: Validation table for during freshman year decision tree

The p values determine that all bins in this decision tree are considered accurate.

7.3 Factors before and during freshman year

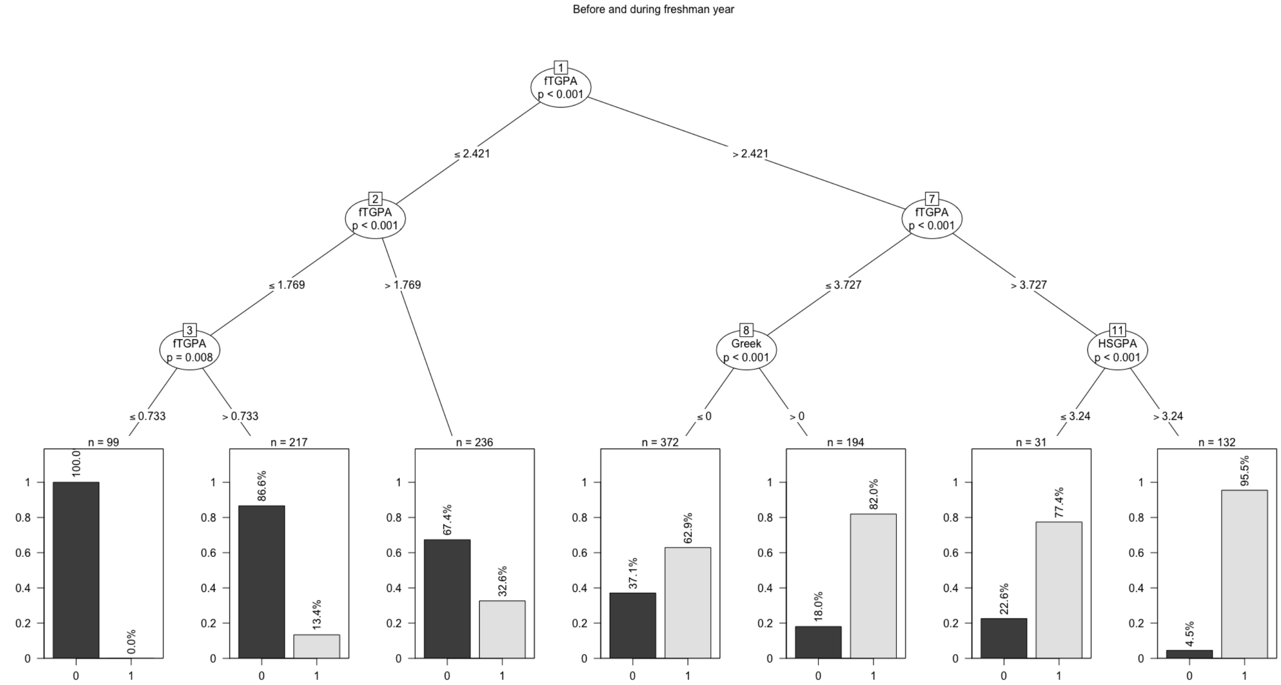
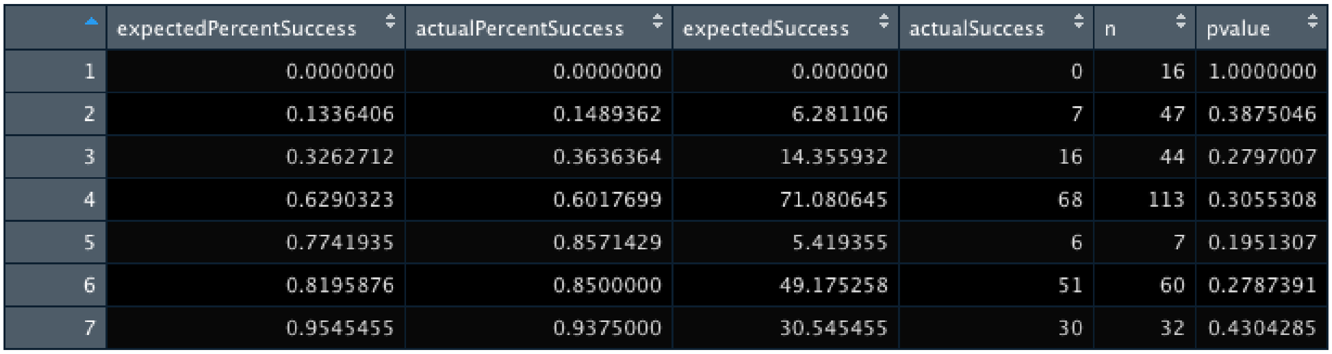


Figure 5: Decision tree formed from the attributes considered in the before freshman year and during freshman year decision trees

Among these factors only first term GPA, high school GPA, and Greek status were found to be significant.



	expectedPercentSuccess	actualPercentSuccess	expectedSuccess	actualSuccess	n	pvalue
1	0.0000000	0.0000000	0.000000	0	16	1.0000000
2	0.1336406	0.1489362	6.281106	7	47	0.3875046
3	0.3262712	0.3636364	14.355932	16	44	0.2797007
4	0.6290323	0.6017699	71.080645	68	113	0.3055308
5	0.7741935	0.8571429	5.419355	6	7	0.1951307
6	0.8195876	0.8500000	49.175258	51	60	0.2787391
7	0.9545455	0.9375000	30.545455	30	32	0.4304285

Figure 6: Validation table for before and during freshman year decision tree

The p values determine that all bins for this decision tree can be considered accurate.

7.4 Factors before and during freshman year excluding GPA

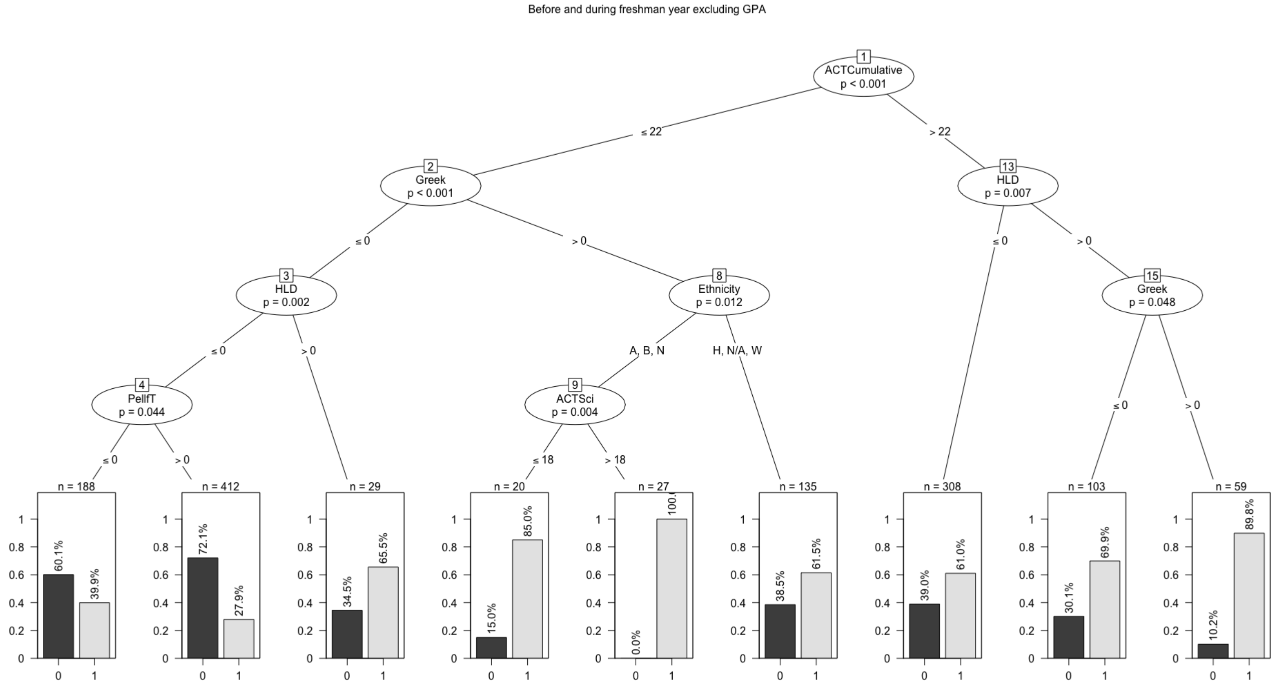
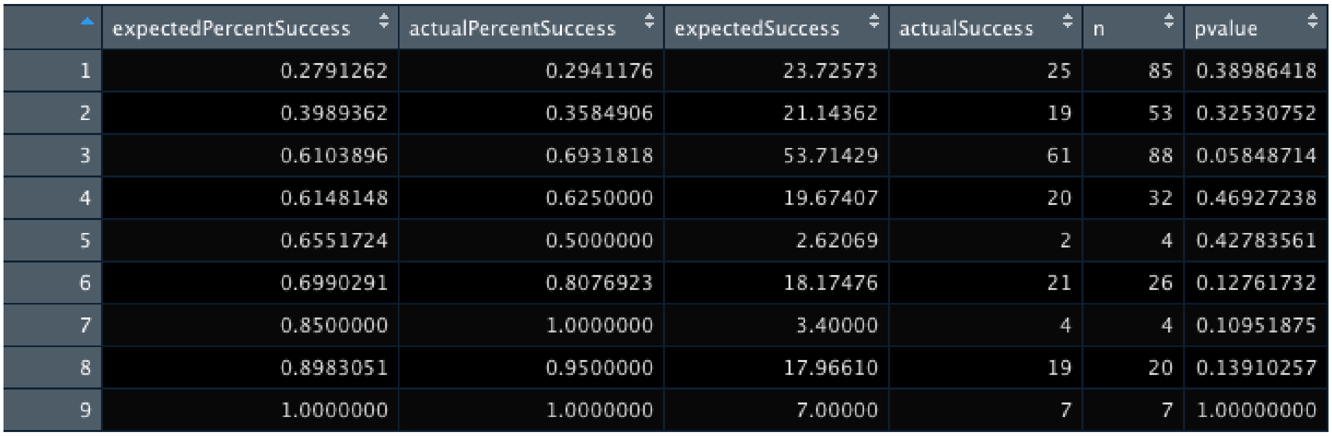


Figure 7: Decision tree formed by the attributes of before and during freshman year trees excluding any GPA attributes

When we remove GPA from being considered we see a much more diverse decision tree. The factors found to be significant are ACT Cumulative, Greek, Honors and Lucky Daye, Ethnicity, ACT Science, and first term Pell status.



	expectedPercentSuccess	actualPercentSuccess	expectedSuccess	actualSuccess	n	pvalue
1	0.2791262	0.2941176	23.72573	25	85	0.38986418
2	0.3989362	0.3584906	21.14362	19	53	0.32530752
3	0.6103896	0.6931818	53.71429	61	88	0.05848714
4	0.6148148	0.6250000	19.67407	20	32	0.46927238
5	0.6551724	0.5000000	2.62069	2	4	0.42783561
6	0.6990291	0.8076923	18.17476	21	26	0.12761732
7	0.8500000	1.0000000	3.40000	4	4	0.10951875
8	0.8983051	0.9500000	17.96610	19	20	0.13910257
9	1.0000000	1.0000000	7.00000	7	7	1.00000000

Figure 8: Validation table for no GPA decision tree

The p values determine that all bins for this decision tree can be considered accurate.

7.5 Factors determined to be significant from Jesse Robinson's study

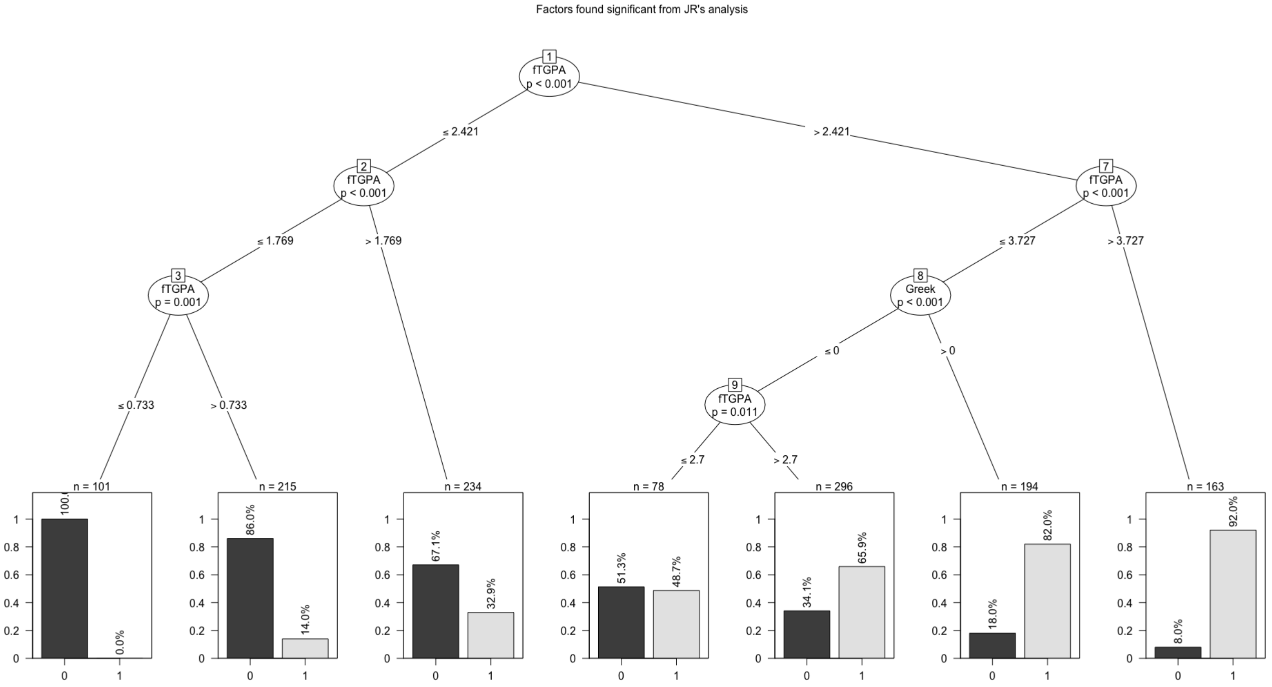
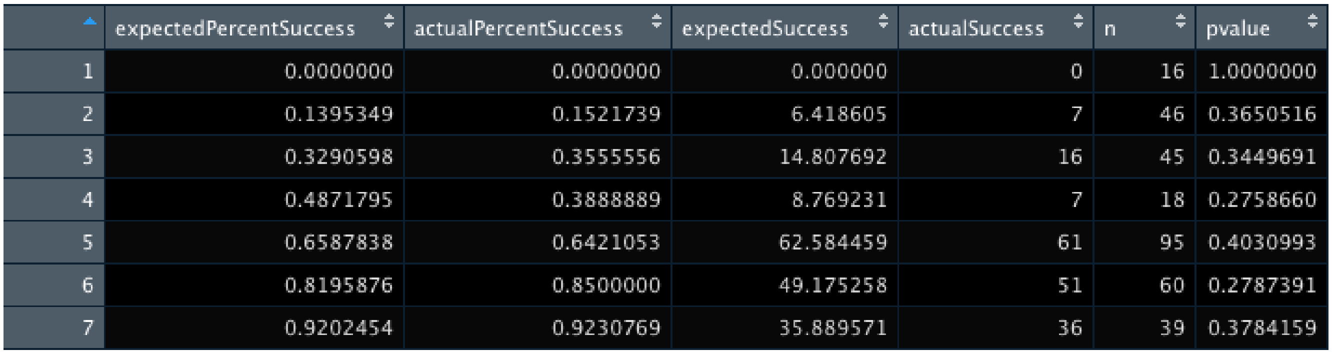


Figure 9: Decision tree made up of attributes found significant by Jesse Robinson's study. These factors are first term GPA, College, and Greek Status

Jesse Robinson's study found that Greek status, first term GPA, and college were significant factors determining whether a student will successfully graduate. Out of these attributes, our study found that only first term GPA and Greek status were significant.



	expectedPercentSuccess	actualPercentSuccess	expectedSuccess	actualSuccess	n	pvalue
1	0.0000000	0.0000000	0.000000	0	16	1.0000000
2	0.1395349	0.1521739	6.418605	7	46	0.3650516
3	0.3290598	0.3555556	14.807692	16	45	0.3449691
4	0.4871795	0.3888889	8.769231	7	18	0.2758660
5	0.6587838	0.6421053	62.584459	61	95	0.4030993
6	0.8195876	0.8500000	49.175258	51	60	0.2787391
7	0.9202454	0.9230769	35.889571	36	39	0.3784159

Figure 10: Validation table for factors found significant from Jesse Robinson's study decision tree

The p values determine that all bins for this decision tree can be considered accurate.

8 Conclusion

Consistent with other studies, high school GPA and first term GPA were identified as the strongest predictors for graduation. Validation showed that the classification provided by the respective decision trees in all but one instance provided success probabilities that can be considered accurate.

References

[1] Hothorn, T., Hornik, K., & Zeileis, A. (n.d.). Partykit [Scholarly project]. In Ctree: Conditional Inference Trees. Retrieved from https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf

[2] Pew Research Center. (2014, February 02). The Rising Cost of Not Going to College. Retrieved October 03, 2020, from https://www.pewsocialtrends.org/2014/02/11/the-rising-cost-of-not-going-to-college/

[3] Robinson, Jesse Homer, ”A Logistic Regression Analysis of First-Time College Students' Completion Rates at The University of Southern Mississippi” (2018). Honors Theses. 610. https://aquila.usm.edu/honors_theses/610

[4] Sullivan, W. (2017). Machine learning for beginners guide algorithms: Decision tree & random forest introduction. Healthy Pragmatic Solutions.

[5] Wei, C. & Hsu, N. (2008). Derived operating rules for a reservoir operation system: Comparison of decision trees, neural decision trees and fuzzy decision trees. Water Resources Research, 44(2).

[6] Zeileis, A., Hothorn, T. (n.d.). Partykit: A Toolkit for Recursive Partytioning [PDF].