DSC 680 Project 1 – Milestone 2

Draft White Paper

Professor Catherine Wiliams

Presented by:

Sherry Kosmicki

April 1, 2023

Business Problem

When a student receives financial aid at a 4 or 2-year academic institution, does their chances of graduating increase? Can just having financial aid determine whether they will complete their program of study? If we were to evaluate the academic institutions, private vs non-private status or profit vs non-profit, based on the graduation rates to determine what schools a student should choose.

There is a great amount of competition in the marketing of post-secondary education. Would this value impact the marketing needed to increase admissions at any one academic institution?

Background/History

According to an article from greatvaluecolleges.net, those academic institutions which yield high graduation rates tend to be more expensive. If you review colleges with a more affordable tuition rate, you will see them start at a 50% graduation rate. (Great Value Colleges, 2022)

College tuition affordability can be determined by the student. With the use of financial aid, in the form of pell grants or loans, a student can choose to go to any school if they are accepted through the admission process.

Methods

The original concept was to clean data from the National Center for Education Statistics. These statistics aggregate the academic level the student chooses, either 2 or 4-year, and whether they received financial aid. The levels are broken out by public or private and within private it deviates to nonprofit vs for-profit.

The methods used started with regression, then multiple linear regression and logistic regression. Regression and multiple linear regression focused on the graduation rates at various levels. Logistic regression was determining if there are variables that impact whether the student received aid.

Analysis

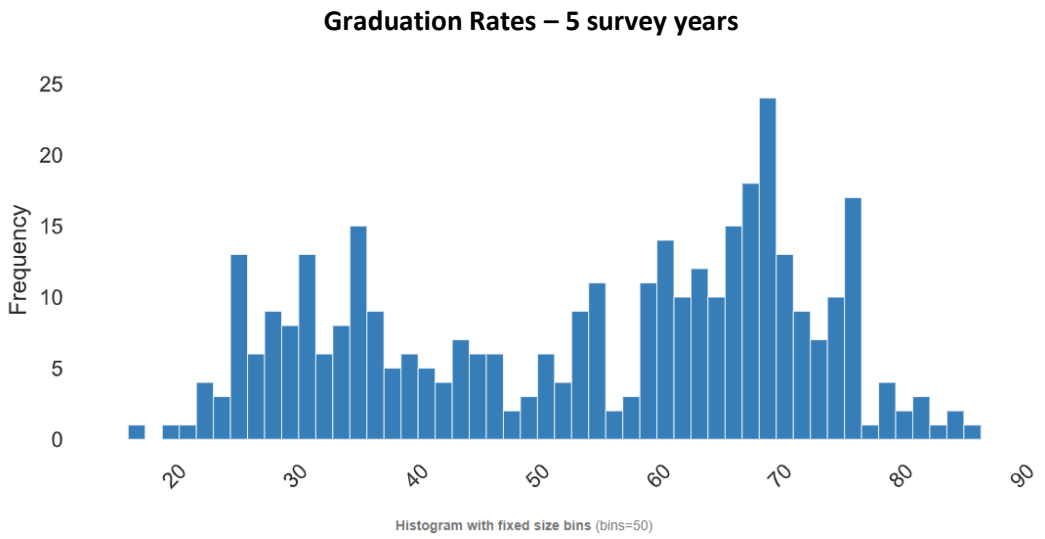
Once the data was extracted and organized, an evaluation of was done for each variable. The final dataset produced 7 variables with 360 observations. The categorical variables are Year, InstType, StudPop, and InstLevel. The numeric variables are NumStudents, Completers, and GradRate. See data sample below:

	Year	InstType	StudPop	InstLevel	NumStudents	Completers	GradRate
0	2010	4-year institutions	All students	Public	1084594	593462	54.7
1	2010	4-year institutions	Received Pell Grant	Public	421649	184115	43.7
2	2010	4-year institutions	Received Direct Subsidized Loan, but not a Pel...	Public	165031	97082	58.8
3	2010	4-year institutions	Received neither a Pell Grant or a Direct Subs...	Public	497914	312265	62.7
4	2010	4-year bachelor's cohort	All students	Public	932667	551604	59.1
...	...	...	...	...	...	...	...
355	2018	2-year institutions	Received neither a Pell Grant or a Direct Subs...	Pri-For-profit	16971	10873	64.1
356	2018	Less-than-2-year institutions	All students	Pri-For-profit	114437	77969	68.1
357	2018	Less-than-2-year institutions	Received Pell Grant	Pri-For-profit	81132	54129	66.7
358	2018	Less-than-2-year institutions	Received Direct Subsidized Loan, but not a Pel...	Pri-For-profit	10313	7789	75.5
359	2018	Less-than-2-year institutions	Received neither a Pell Grant or a Direct Subs...	Pri-For-profit	22992	16051	69.8

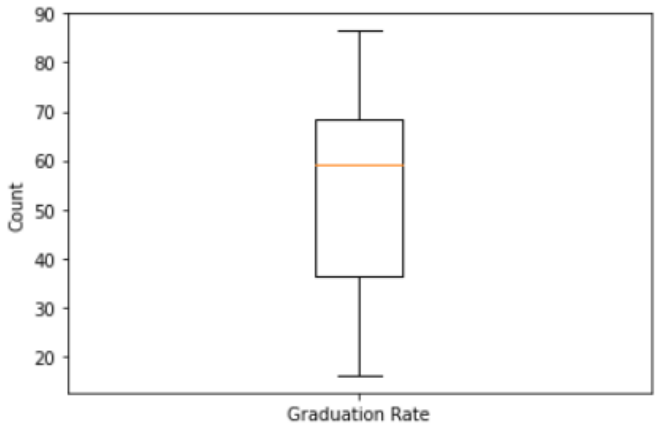
I then created two aggregated variables that identified values as either 2-year or 4-year institutions called agg_Itype and creating 2 categories of the StudPop data as Yes Aid or No Aid. All Students were left blank to be removed later when evaluating.

agg_Itype	agg_Stpop
4-year	
4-year	Yes Aid
4-year	Yes Aid
4-year	No Aid
4-year	
...	...
2-year	No Aid
2-year	
2-year	Yes Aid
2-year	Yes Aid
2-year	No Aid

A histogram of graduation rates shows a wide spectrum of values.

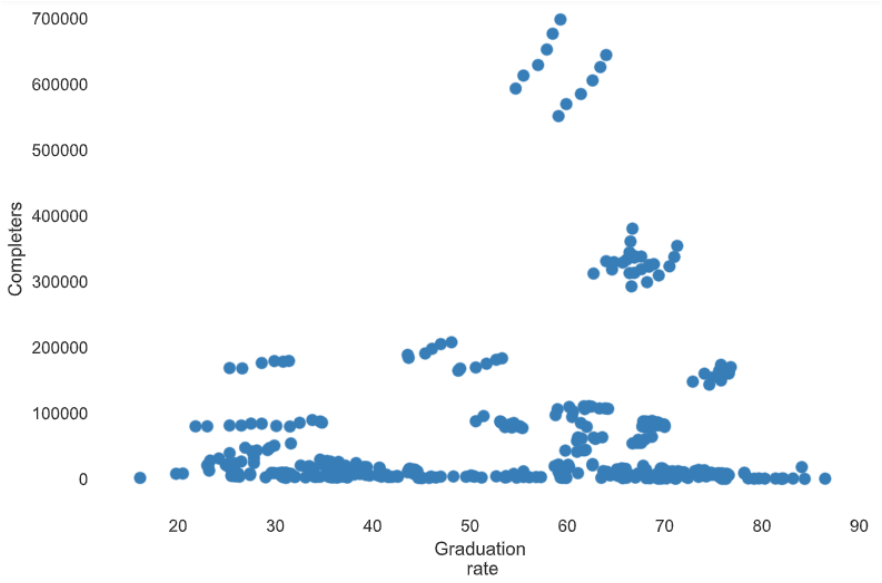


This box and whisker plot gives an idea of the range of the graduation rate like the histogram, but this provides a mean value of around 60 percent.



As I looked for correlations between the data, it is known that the StudPop and Completers are correlated with the GradRate due to the determination of GradRate us StudPop divided by Completers. In this visualization, there is an increase in graduation rate when more students complete a 2-year or 4-year academic institution. This graph is similar when comparing student population and graduation rates.

Graduation Rates vs. Completers



I created dummy variables to represent each of the categorical variables and put them in the same dataset with the numeric variables. Then they were split into a train and test environment.

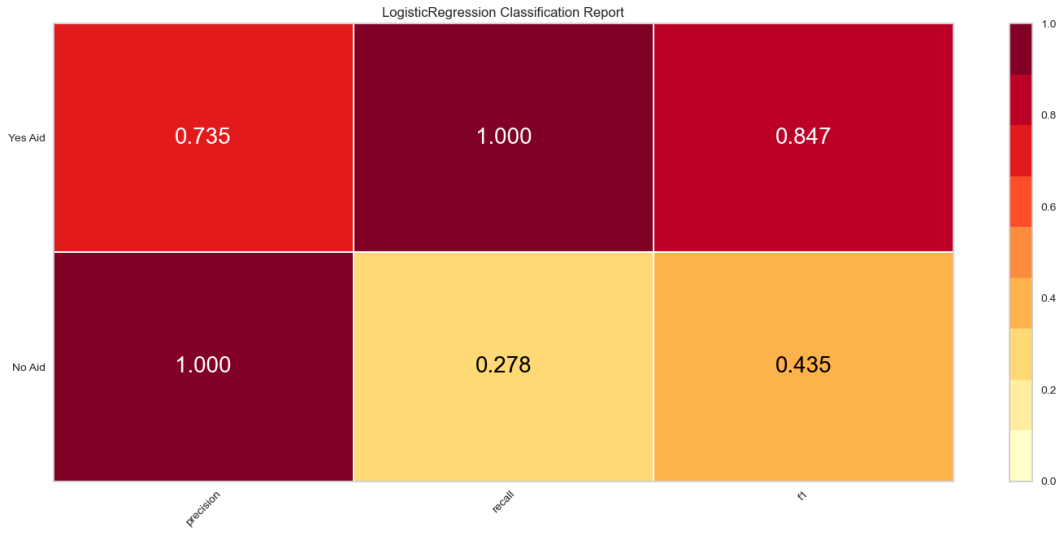
Since a heat map determined a positive relationship between the 2-year variable and graduation rate, a linear regression model was created for those variables. The observation in this model was the coefficient of determination was 15% which means the 2-year school is not a good predictor of graduation rates and it is not a good fit for this model.

The multiple linear regression model took into consideration all the possible variables as determining factors and the coefficient of determination came to 69%. It scored much better than the previous linear regression.

Lastly, I used Logistic regression to take a slightly distinctive look at the data. I wanted to know if the variables involved could determine if the student received aid. This makes the graduation rate, in this scenario, an independent variable.

The confusion matrix shows that there is a positive predictor that the variables can conclude “Yes Aid”. Due to the lack of a large amount of data, in addition to the division to create a train and test, the amount of data for this analysis does not make it definitive.

According to the Precision score, 100% of the “Yes Aid” predictions were correct. While these scores look clean, I would feel more confident with more data points. The ROC curve added little value to the analysis.



Conclusion

There are no variables that stand out as impacting the graduation rate when reviewing the heat map. While the use of multiple linear regression and logistic regression was positive, the lack of data could determine less confidence in predictions.

Assumptions

When I originally started this analysis, I was motivated to see if it could be determined which post-secondary institution category to choose based on a high graduation rate. As I looked at the data closer, this is aggregated data at those levels per academic institution. The law of averages is being used in these models. Based on that and the general lack of volumes of data to model. The data is predictable, but there is not enough of it to feel confident about the outcomes.

Limitations

Once this evaluation was completed, I worked on finding the graduation rate detail information rate at the academic institution level. I was able to find the details used in the reporting and have started evaluating that data. This move could help remove the law of averages that plays into using summary data for the modeling.

Challenges

The challenge right now is finding detailed data in relation to academic institutions. I need to verify the calculation for graduation rate with the total students versus and completers. A replay of the same modeling would be telling and could provide an opportunity to expand the modeling to knn classifiers.

Future Uses/Additional Applications

This data could be expanded to review other variables gathered by the National Center for Education Statistics. These could be identified as determining factors to impact the graduation rate.

Recommendations

My recommendation would be to reevaluate with more data. I am not convinced I am getting a true picture of graduation rates by using the aggregated amount from the summary reporting. This evaluation deserves more detailed data and academic institution level raw numbers.

Implementation Plan

The idea for implementing would have to be a reevaluation of the current data structure and munge the academic institution level. I would like to develop a knn classifier model. Some of the original EDA indicated that there were close in proximity values.

Ethical Assessment

Using averages, unless there are multiple data points, can be unfairly depicting the values at a singular academic institution level. I cannot use the data at an aggregate level if I want to predict a graduation rate on a singular school.

References:

U.S. Department of Education, National Center for Education Statistics. (2019-20). IPEDS Data Explorer. https://nces.ed.gov/ipeds/Search.

Great Value Colleges, 2022. Affordable American Colleges With the Highest Graduation Rates: These Schools Earn Top Marks! https://www.greatvaluecolleges.net/highest-graduation-rates/

10 Audience Questions

1. Do you believe there is a difference in Graduation Rates between 2 and 4-year academic institutions?
2. Is there a graduation rate difference between public and private academic institutions?
3. Do student populations per academic institutions vary in your study? How can you identify?
4. Do the graduation rates look better for any particular year surveyed?
5. When evaluating the data for those students that receive aid or not, what did you determine?
6. Do those students that don't receive aid correlate to graduation rates?
7. Can I evaluate graduation rates of academic institutions with this study?