Regression Analysis Report

Marc Navia, Business Analyst

NC STATE UNIVERSITY

TABLE OF CONTENT

Executive Summary

Introduction

A study was conducted to determine what factors have an effect on colleges and university graduation rates. A sample of 55 colleges and universities from across the United States was used in the regression analysis. The main purpose of this study is to construct a model that can be used to predict the graduation rate of a college or university depending on several factors. The factors that were analyzed were median SAT score, acceptance rate, expenditures per student, the percentage of students in the top 10% of their high school class and whether the institution was university or a liberal arts college.

Recommendations

From this study, education institutions have many ways to increase their graduation rate percentage. These institutions need to be more selective when it comes to choosing their students. One of the best ways is to pick students with higher SAT scores to increase their schools median SAT score and to lower. If this isn't an option, picking students who are the top 10% of their high school class has a positive influence when it comes to increasing graduation rates.

Overall if an educational intuition wants to improve their graduation rate percentage they must focus on Median SAT score, students who are top 10% in their class and acceptance rate based on the research conducted.

Key Findings

Median SAT Score is the best single predictor of Graduation Rate %.

An institution being a University, or a Liberal Arts College has little to almost no effect on graduation.

Schools with a higher acceptance rate will often have a lower graduation rate percentage compared to schools with lower acceptance rate.

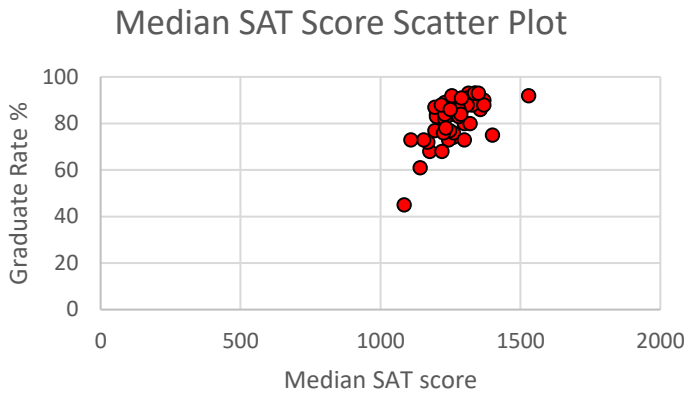
Median SAT score, acceptance rate and the percentage of students in top 10% of their class combined makes the best prediction model.

Student expenditures have a small effect on graduation rate percentage with most colleges having expenditures between $15,000 - $34,000.

Descriptive Statistics

MEDIAN SAT

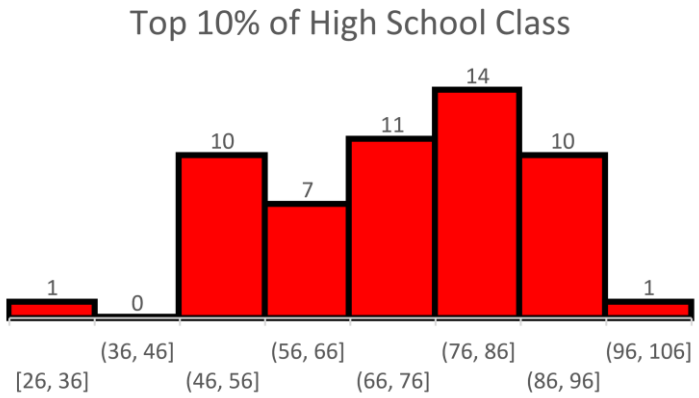
The median sat score has a strong positive association with graduation percentage. The higher the median sat score the higher the school's graduation rate would be. When looking at the scatter plot we can see a curve from the 1200 – 1400 median sat score range. The benefit of having a higher median sat score starts to level off the closer it reaches the 1400 median score.

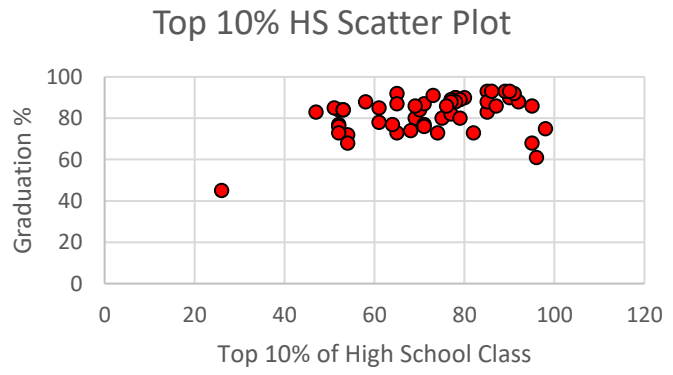


Top 10% of High School Class

The top 10% of High School Class variable had a medium strength positive association with graduation percentage. The higher the top 10% class percentage was the higher graduation rate would be. Looking at the scatterplot we can see a curve from 50% to 90%. Like median sat we see the increase in graduation % start to slow down and then decrease at the tail end.

The average value for the percentage of students in top 10% of their high school class among all schools in the data set was 72.4%. When looking at the histogram we see there is a large spread amongst the data. Sixty-eight percent of the data has a percentage between 57% and 88%. This is the range where we start to see increase and decrease of the influence of top 10% have on graduation rate. Lastly, we can see there is an outlier in the histogram in the 26%-36% bucket which is Middle Tennessee State.

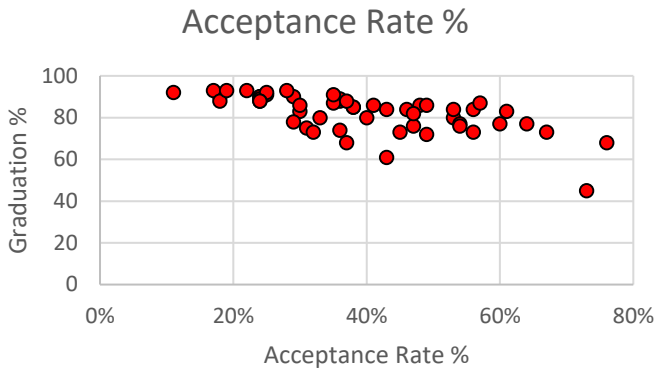




Descriptive Statistics Cont.

Acceptance Rate Percentage

The acceptance rate percentage has a strong negative association with graduation percentage. As the acceptance rate increase the graduation percentage will then decrease. Unlike the median sat and top 10% variables this data follows a more linear path. Schools can have the same acceptance rate but a different graduation rate. This is most likely due to the other consideration schools look for when picking students.



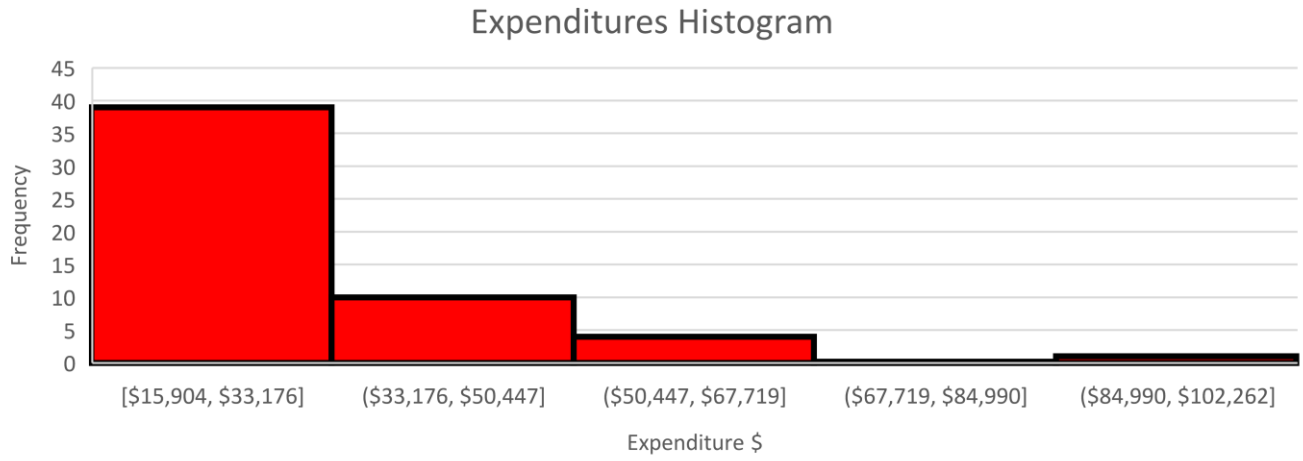
University Type

The university type had a very weak negative association with graduation percentage. If the school is a university then it has a slightly lower graduation rate. From the data set, 29 of the 54 (53.7%) of the schools were liberal arts.

Expenditures

Like the university type variable, expenditures have a very weak negative association with graduation percentage. The more a student spends on expenditures the lower the school graduation percentage. The average amount student spent on expenditures is $30,106.

The histogram below shows that at most schools in our dataset, students spend between $15,000 – $33,000. The obvious outlier which can be seen on the right tail end of the chart is Cal Tech where students spend over $102,262 on average.



Simple Regression Model

The Best Single Predictor Variable

From the five variables that were analyzed, median SAT score was the best single predictor of graduation rate percentage. This makes sense as it had one of the strongest association to the graduation rate. Using this variable in the simple regression model it was able to explain 52.12% of the variability occurring with the graduation rate.

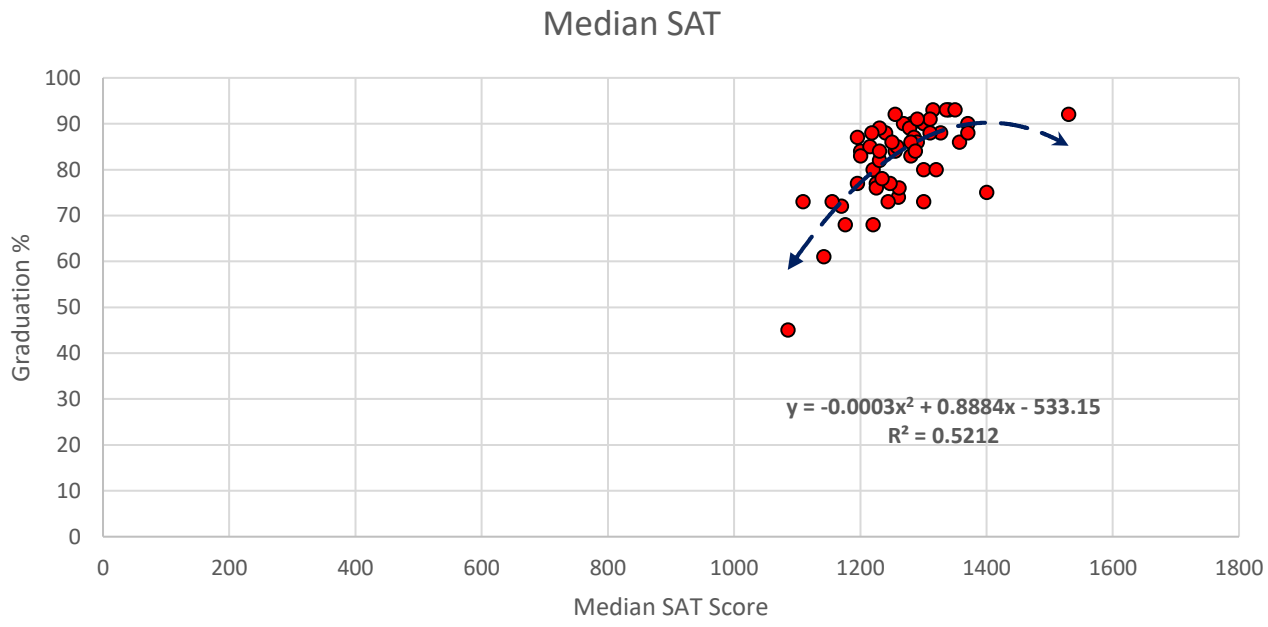
What Does This Model Tells us?

Schools that have a higher median sat score will see a much higher graduation rate percentage compare to schools with a lower median sat score. Students who perform well on the SAT are most likely to also perform well in college. With the SAT being a standardize being a test of the major subjects in school, it itself is a useful tool in gauging possible future school performance.

This model also shows how there is also diminishing benefit for schools in increasing their overall median SAT score. From the SAT scores of 1100 to 1300, the increase in graduation rate percentage has the most value. From that endpoint, the benefits start to level off and actually decreased. This shows that there are other variables that come into play then just only median SAT score

1100 – 1300
SAT SCORE

1400+
SAT SCORE



Multiple Regression Model

The Best Predictor Model for Graduation Percentage

On the bottom of this page is the equation that best predicts the graduation rate percentage of a school using three variables. The three variables that this equation uses are median SAT score, acceptance rate percentage, and percentage of students who were top 10% in their high school class. For the percentage of students who were top 10% and median SAT score they appear twice with one of them being squared. What this shows is that as the top 10% percentage and median sat scores increase the effect of increasing the graduation rate percentage starts to slow down and eventual decrease. For the acceptance rate variable, it will decrease the graduation rate percentage the higher it is and increase it the lower it is. This makes sense as it has a strong negative association with graduation rate. Overall this model explains 71.31% of what is affecting the graduation rate percentage of schools.

The Outliers

In the data set, three outliers were found. Outliers are data points that area abnormal compared to the average value in the data set it is in. The three outlier schools are Vanderbilt, Middle Tennessee State, and Cal Tech. Vanderbilt had a significantly abnormal median SAT score, Middle Tennessee State had an abnormally low percentage of students who were top 10% in their class and Cal Tech who had an abnormally high student expenditure. For the university graduation rate prediction model, the outliers were kept.

Removing the outliers from the Model

When the three outliers were removed, it had a negative affect on the prediction model. On the best predictor model, it caused the variables to be insignificant which means they are not applicable for predicting. After multiple regression analysis involving multiple variables, the overall effect of removing the variables caused the prediction accuracy to lower. The negative effect of not having these outliers might be due to the relatively small sample size. These outliers schools only have only 1 to 2 abnormal data points but overall they provide enough data to improve the model.

Testing the model

This test will involve a school with the following data

- Median SAT score = 1210
- Acceptance Rate = 23%
- Expenditure $25,500
- Percentage of students in top 10% of high school class = 79%
- A university

This model would only use median SAT score, acceptance rate and Percentage of students in top 10% of high school class. Using those three variables we can predict that the graduation rate percentage would be 86%.

$$\begin{aligned} &\boldsymbol{Graduation\ Rate\%} \\ &\quad = -0.0099(\boldsymbol{Top\ 10\%\ HS})\text{^}2 + 1.2612(\boldsymbol{Top\ 10\%\ HS}) \\ &\quad - 0.0002(\boldsymbol{Median\ Sat\ Score})\text{^}2 + 0.4503(\boldsymbol{Median\ SAT\ Score}) \\ &\quad - 31.4338(\boldsymbol{Acceptance\ Rate\%}) - 257.3465 \end{aligned}$$

Conclusion

Spending more doesn't mean better……

…….Students who spend more on their tuition will see very a marginal benefit in increasing their chance in graduating. Are intuitions who cost more using the money wisely?

University or Liberal Arts College?

….. Going to either has little to no effect on increasing your graduation chance

The Higher the Median SAT score the more student who graduate…...

……. Students who perform well on the SAT are more likely to be better students

What can be done to improve the model?

Having more data can drastically improve the models. This was obvious during the regression modeling without the outlier data. Having less data decrease the prediction capability of the model. What can be also done is adding extra data by extra variables. While a good portion of graduation rate percentage could be explained by the model, there could possibly more or even better variables that can help explain the graduation percentage. A couple of variables that could be taken into consideration are average teacher pay and the education institution expenditures. The average teacher pay could possibly show how teacher pay relates to the quality of education that students are receiving. Education intuition expenditures could show how much these institutions are investing in their students which could affect the quality of the student's education. From this data set, we focus mostly on the quality of the student and not the quality of the institution. Having this data could give us the necessary information to explain more of what is going on with the graduation rate percentage