Analysis of the 2011 College Scorecard Data Set

Keith Hultman

October 24, 2015

Introduction

The application and decision process can be daunting for many of the millions of students who enroll in American Colleges and Universities each year. Although the reasons to attend any particular school are numerous and highly individual for each student, there are several common factors that are usually considered when making a decision to apply to a particular school. The College Scorecard provided by the Department of Education provides an unbiased metric for evaluating college quality and costs. The primary website allows potential students a way to search through schools and programs using several parameters and allows students to compare schools based on the average income of previous students. Previous rankings had not included salary or earnings data from graduates, so the release of this data set represents an entirely new opportunity to evaluate schools.

This analysis will explore the College Scorecard data set and attempt to gain an understanding of why some schools are more successful than others in training students for financial achievement.

Description of the Data Set and its Limitations

The data set was compiled by the Department of Education from federal reporting of all accredited post secondary schools in the U.S. Data is available for the years from 1996-2013 at College Scorecard.

I will be using data from 2011, since this is the most recent data set that includes the median salary of students 10 years after graduation. In this case, the 2011 data for salary are based on students who graduated from the institution in 2001. The 2011 Scorecard data set includes 7675 schools and 1729 variables describing various attributes of each school and its student body. The most difficult aspect of exploring this data set was determining which variables I should include in the analysis. I am most interested in the question of how large of an effect a school can have on a student's potential for earning a high salary. I narrowed down the variables to a list of ~50 that I thought might play a role in determining salary, from the types of degrees offered to the fields of study to average SAT and Tuition costs. Several of the initial explorations, such as predicting a student's probability for acceptance to a school given their SAT score, could not be determined without additional data from the school not currently reported on the Scorecard.

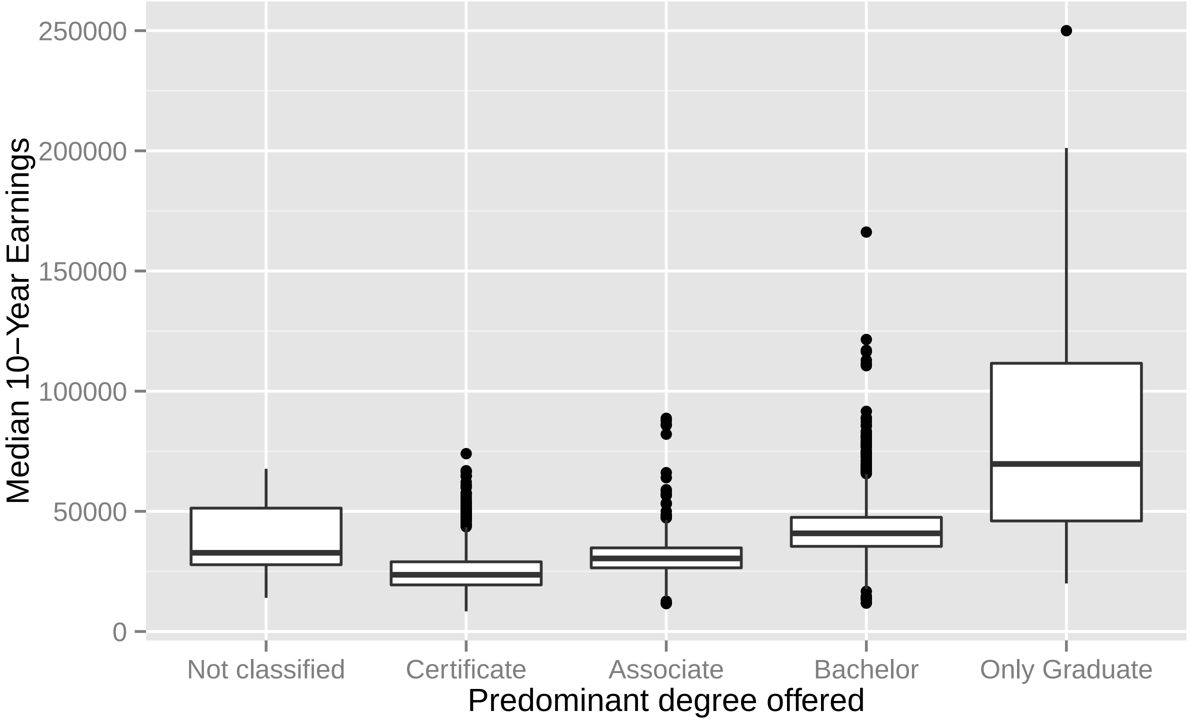
According to the data documentation, many fields in the data were collected from Title IV recipients. These students receive federal grants and loans. This means that the data may not reflect a random sampling of all students and could introduce bias in the accuracy of any given metric that is based off of Title IV recipients. In other words, these data may not be suitable for determining the overall financial situation for all students at the college or university. However, it is appropriate for comparing schools within this subset of students. Analyses from this data set would benefit students who expect to receive federal grants and loans when comparing the parameters of different schools and financial outcomes. Furthermore, schools that deliver beneficial outcomes for Title IV students should also be capable of delivering beneficial outcomes to students who do not qualify or need financial assistance. Several samples within the data are suppressed from publication due to potential identification of individual students. Schools do not report on every variable and there is a lot of missing values for each variable.

Analysis

There are several different types of schools in this data set. The first exploration was to examine whether the median 10-year earnings varied significantly based on the type of degrees offered by the school (Figure

1). The relationship was tested using ANOVA, with a reported P value of 2.2e-16. This relationship was somewhat expected, given that income has previously been shown to be related to the type of degree earned, with more advanced degrees receiving higher salaries.

Figure 1:
Earnings by Types of Degrees Offered



After visualizing various columns of the data, I noticed one graph that had an interesting pattern. When looking at the relationship between selectivity and mean SAT score, I noticed two groups of schools with very different relationships (Figure 2a) to these variables. One set of schools seemed to show a direct linear relationship between Selectivity and SAT score, where the more selective the school the higher the average SAT score for the student body. Another set of schools seemed to show no relationship between the variables, and included highly selective schools (rejection rate above 60%) with quite moderate SAT scores (800-1100 Cumulative Average). I further explored this data by using k-means clustering with 3 groups. This resulted in identifying 3 types of schools based on selectivity and SAT scores: Selective High SAT, Selective Low SAT, and Nonselective schools (Figure 2b). Within the Selective High Low SAT group includes art schools, which primarily evaluate students for talent not directly measured by SAT.

Figure 2a:
Selectivity vs. SAT Scores

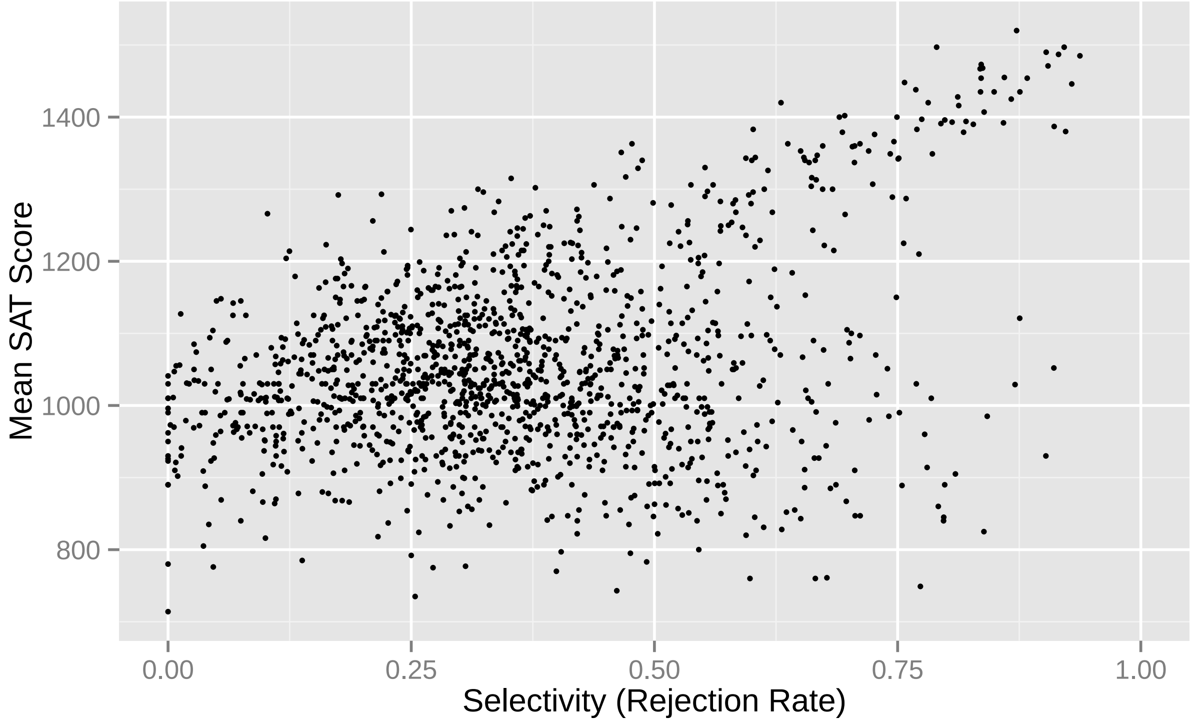
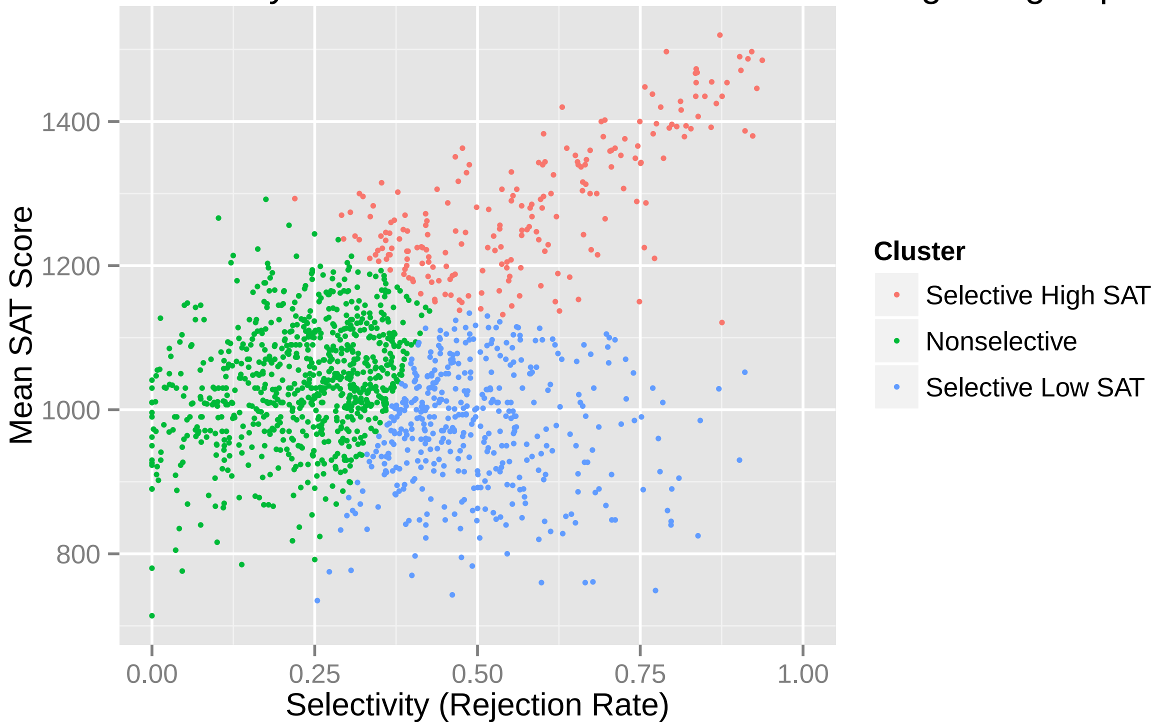
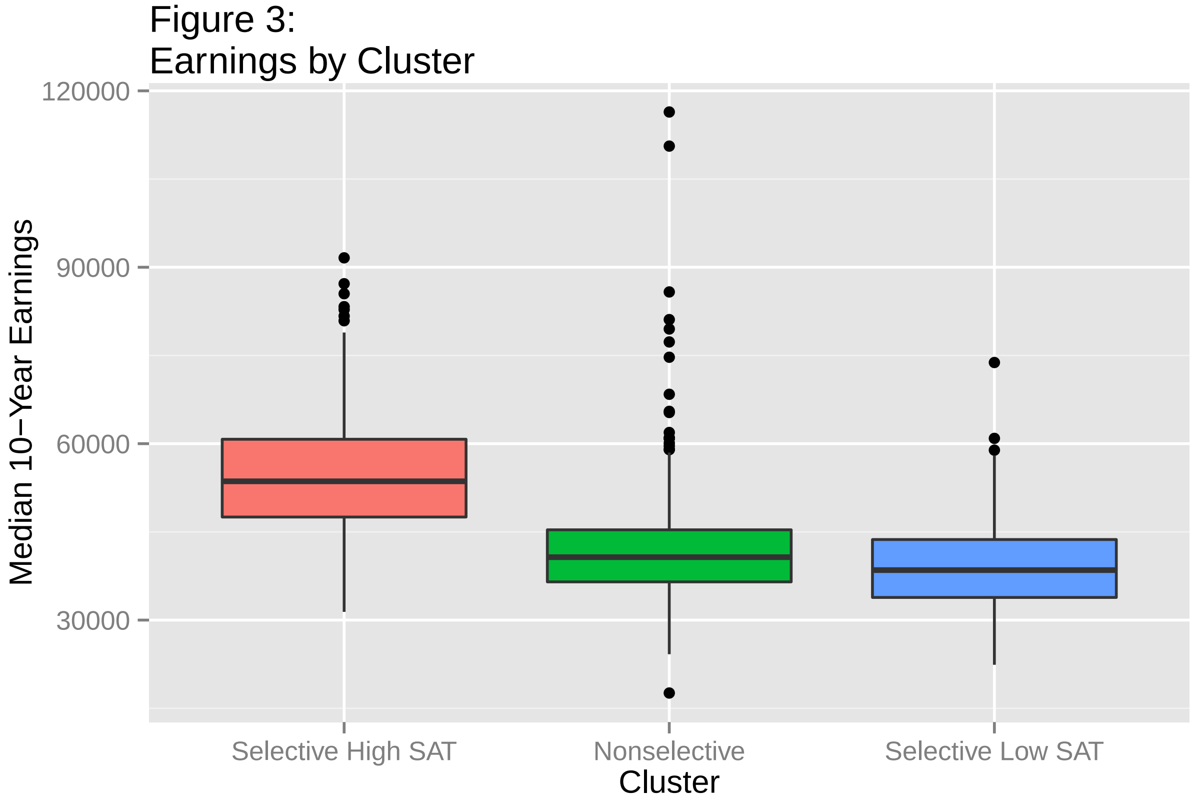


Figure 2b:
Selectivity vs. SAT Score with k−means clustering of 3 groups

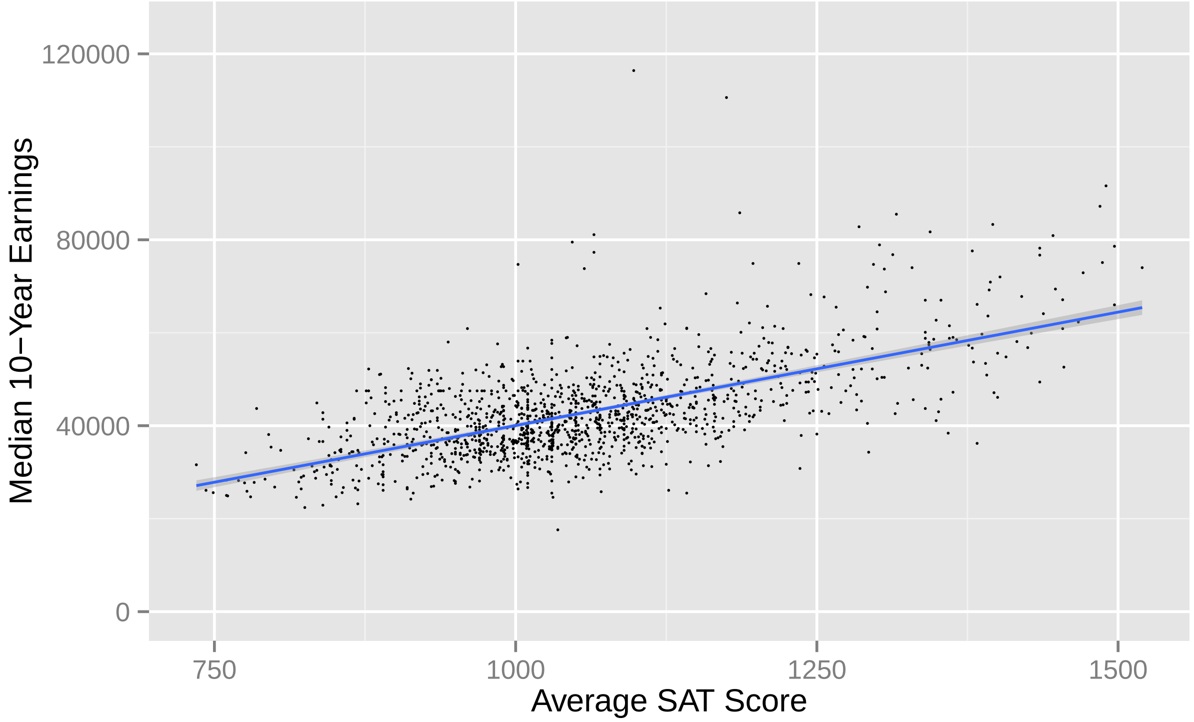


I next looked to see if the different clusters of schools found in the k-means clustering had different outcomes in the median earnings of graduated students. Figure 3 shows the boxplots of median earnings in the three groups of schools. Perhaps not surprisingly, the Selective High SAT schools had an average Earnings outcome that was higher than the nonselective schools. A t-test of the Selective High SAT schools (mean = 55463.37) against all other schools in the data set (mean = 40565.82) showed that the difference in means is statistically significant (P < 2.2e-16). More surprisingly, the Selective Low SAT schools had a relatively low median earnings (mean = 38981.51) that was less than the Nonselective group (mean = 41410.38). This difference was also significant with a P value of 5.0e-7.



It is clear that incoming student ability, as measured by average SAT scores of accepted students, seems to have a bigger effect on earnings than selectivity through non-SAT considerations. In fact, in the data as a whole, SAT scores seems to have a linear relationship with future earnings (Figure 3). The Pearson correlation for this positive relationship is r = 0.624.

Figure 3
SAT Score vs Earnings



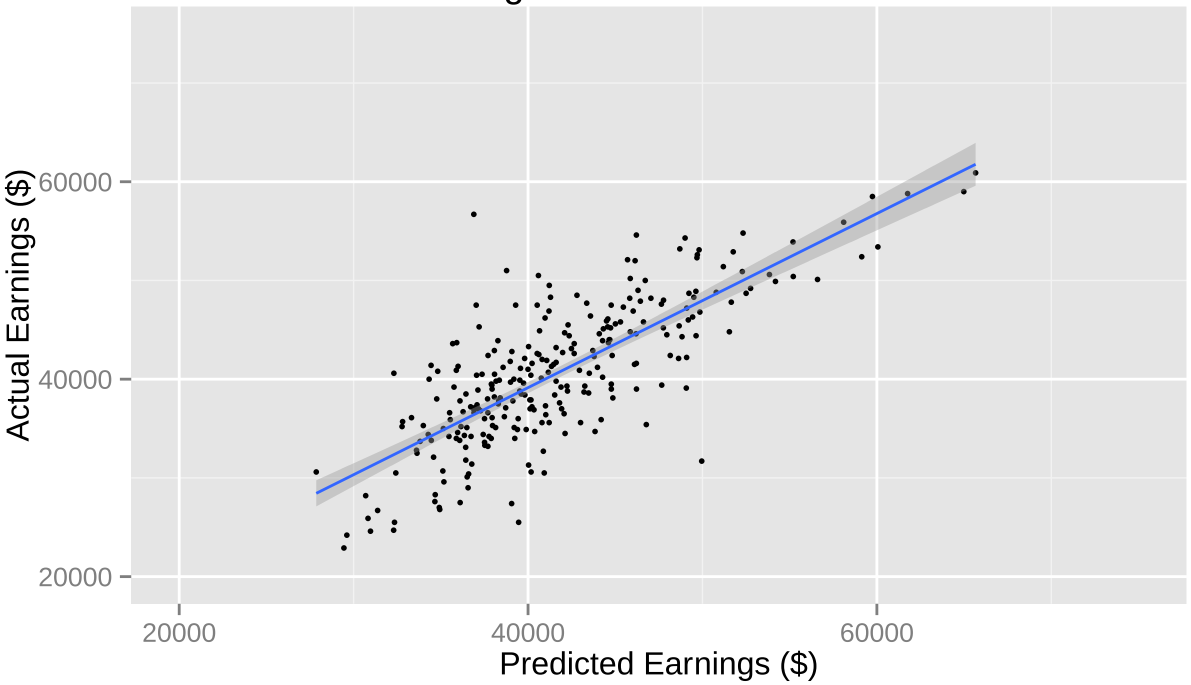
I next wanted to see if median earnings from the schools could be predicted given various parameters from the school. I randomly split the data into a training set with 80% of the observations and a test set with the other 20% of observations using the sample.split function from the caTools package. Using linear regression, I next created several models using combinations of the following parameters:

- Average SAT Score
- Faculty Salary
- Predominant Degree Awarded
- Highest Degree Awarded
- Cost of Tuition
- Total expenditures per student
- Admission Rate

Evaluating the adjusted R squared value for each model and examining each parameter's P value indicated that the simplest model with the most predictive value was likely to be the model Median Earnings ~ Average SAT Score + Faculty Salary + Cost of Tuition with an $R^2 = 0.576$

To validate the linear model, I then made predictions of median earnings using the linear regression model against the test set. Figure 4 shows the plot of Observed Earnings vs. Predicted Earnings for the test data.

Figure 4
SAT Score, Faculty Salary, and Cost of Tuition
Predicts Future Earnings in Test Set



Discussion - Earnings are proportional to student inputs

The variables that I was able to identify as having a predictive role in earnings potential were SAT Score, Faculty Salary, and the Cost of Tuition. It should be noted that the cost of tuition in this case is the overall price of education (tuition and fees) minus any financial aid. As stated earlier, I am most interested in the question of how large an effect a school can have on income. Unfortunately from the point of view of incoming students, two variables are beyond their control and might be based on their inherent intellectual abilities and their economic class. One of the most significant components for predicting earnings was the average SAT of students from that university. This underscores the difficulty in dissociating the inherent quality of the student from the quality of the school in evaluating a school's performance. The ideal experiment would be to randomly assign the incoming class of students to each school, and then measure outcomes. Both the schools and students would probably object to such an experiment, however.

If I were to explore this data set further, I would like to try and find out if there are any other variables that could predict earnings after controlling for SAT score and out of pocket expense. Other variables that might be worth looking into are completion rate of incoming students, and degree awarded. There is a limitation to this dataset, however: there is not enough granularity with regard to the major field of study and outcomes. I would expect any effect of school to be less than the effect of different majors to the relationship with earnings, and it might be easier to see an effect of school within majors.

Perhaps one optimistic conclusion from this analysis is that faculty salary is correlated with student earnings. Schools interested in having a larger pool of alumni funds might be willing to test for a causal relationship by increasing their faculty salaries.