The Real Value of a College Degree

Jake Marsh, Jeff Bertucci, Michael Znidarsic, Tim Zalk

# Specification

## Problem

With most jobs college graduates are able to get with their degrees, the average time to repay student loans is 21 years. Knowing what factors contribute to a student’s ability to repay their loans can better inform the college and degree decision process. We intend to answer the question: *What is the value of an undergraduate college degree based on the rising costs of tuition compared to students’ expected income?*

## Hypothesis

There exist correlations between attributes of colleges and the income and employment of students after obtaining a degree.

## Data

The US Department of Education provides the College Scorecard for evaluating colleges based on various metrics and demographics. The raw data of the College Scorecard contains aggregated information on school demographics, tuition, expenses, programs, and many other metrics of schools combined with student and alumni information from the financial aid system and federal tax returns.

# Observation

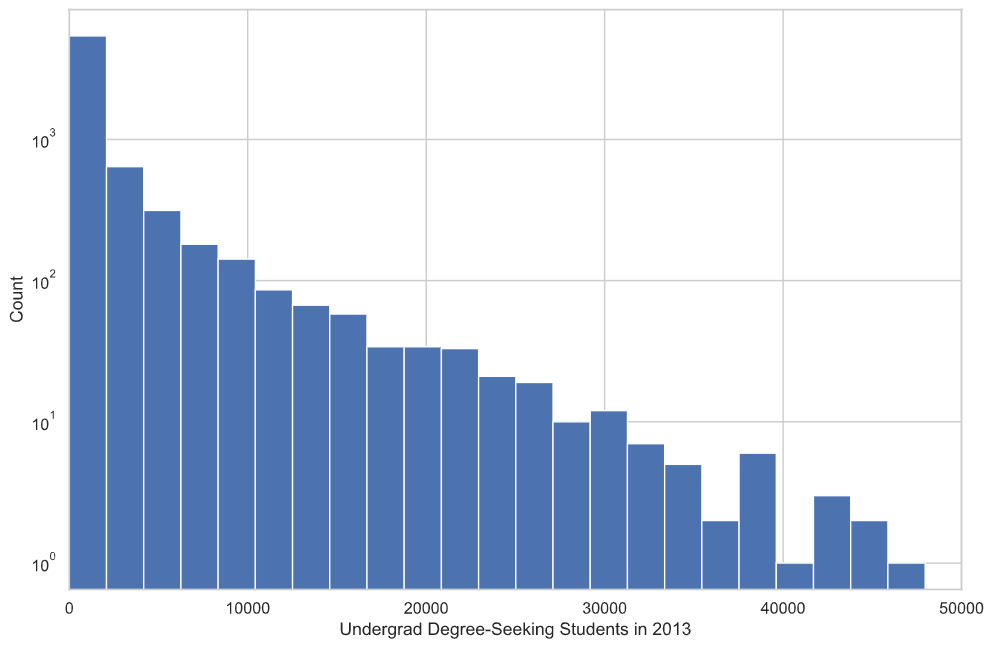
The College Scorecard dataset is highly dimensional, with 1,741 columns. Adding to the complexity of so many columns is the abbreviations and acronyms used for most columns are not self explanatory. A data dictionary must be used to interpret each column’s meaning as well as many columns encoded values. Data provided is from school years between 2003 and 2019, with data not available for every feature from every school each year. The dataset covers almost every organization of higher education in the United States, including very small schools with a student population in the dozens, and schools that do not offer traditional bachelors degrees. Data is also missing from some features for some schools based on various privacy restrictions.

Some of the more pertinent features in the dataset include:

* School
* Neighborhood Type (city, suburb, rural, etc)
* Degrees Offered
* Admission Details
* Student Demographics
* Tuition Cost
* Student Socioeconomic Details
* Student Loan and Financial Aid Details
* Loan Repayment and Debt Details
* Student Earnings at Increments Past Graduation

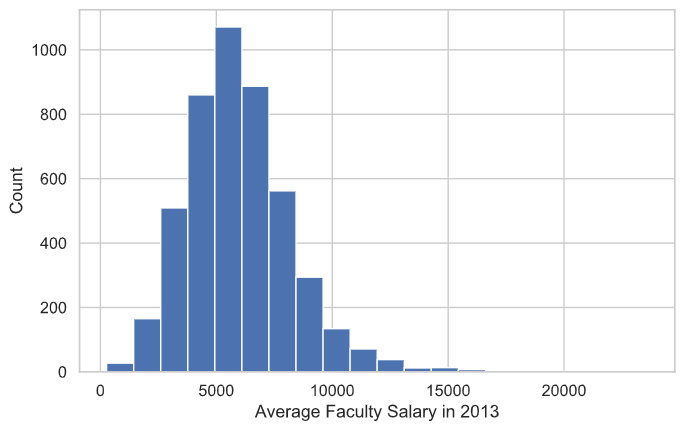
Using data from 2013, a year where most features are available for most schools, we conducted exploratory data analysis.

The histogram below shows the size of each school based on undergraduate degree-seeking students, defined as the number of students enrolled in the school working toward a program completion. This includes students enrolled in certificate and other certification programs and excludes students auditing courses or otherwise not seeking completion of a program.

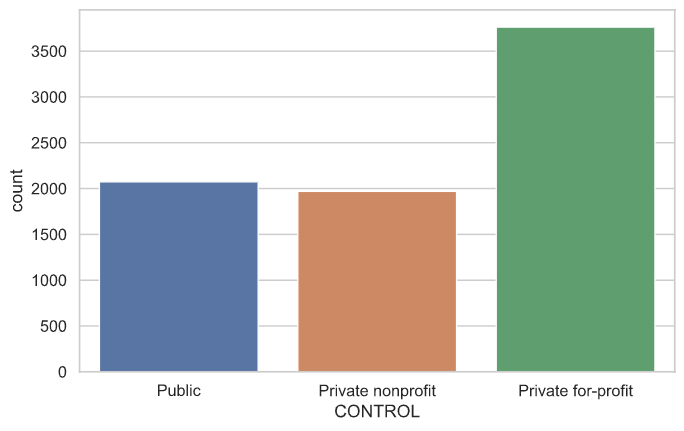
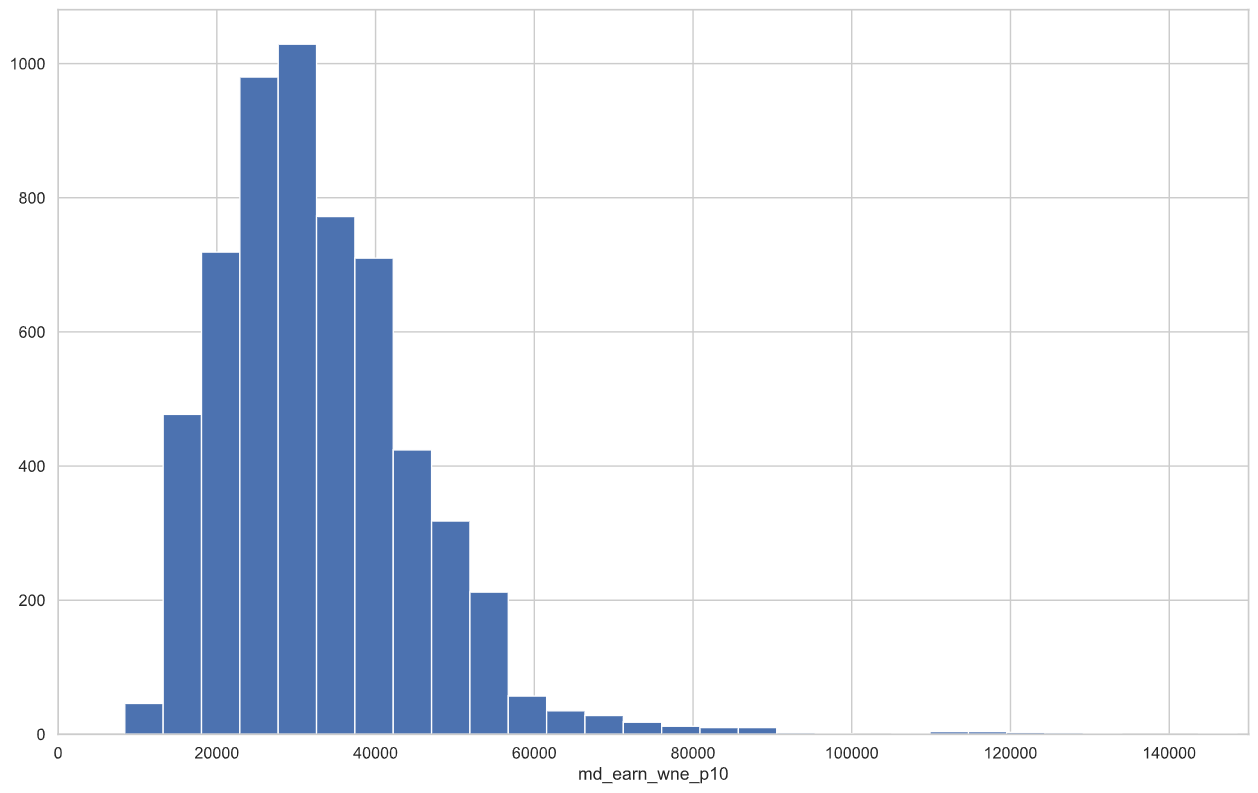


As can be seen in the histogram, most schools have less than 10,000 students, with few schools approaching 50,000 students.

The next histogram below shows average monthly faculty salary per school. Most schools’ average faculty salaries fall in the $3,500 to $6,500 range, which equates to $42,000 to $78,000 annually if paid for 12 months. This does not factor in the proportion of faculty that are part-time.

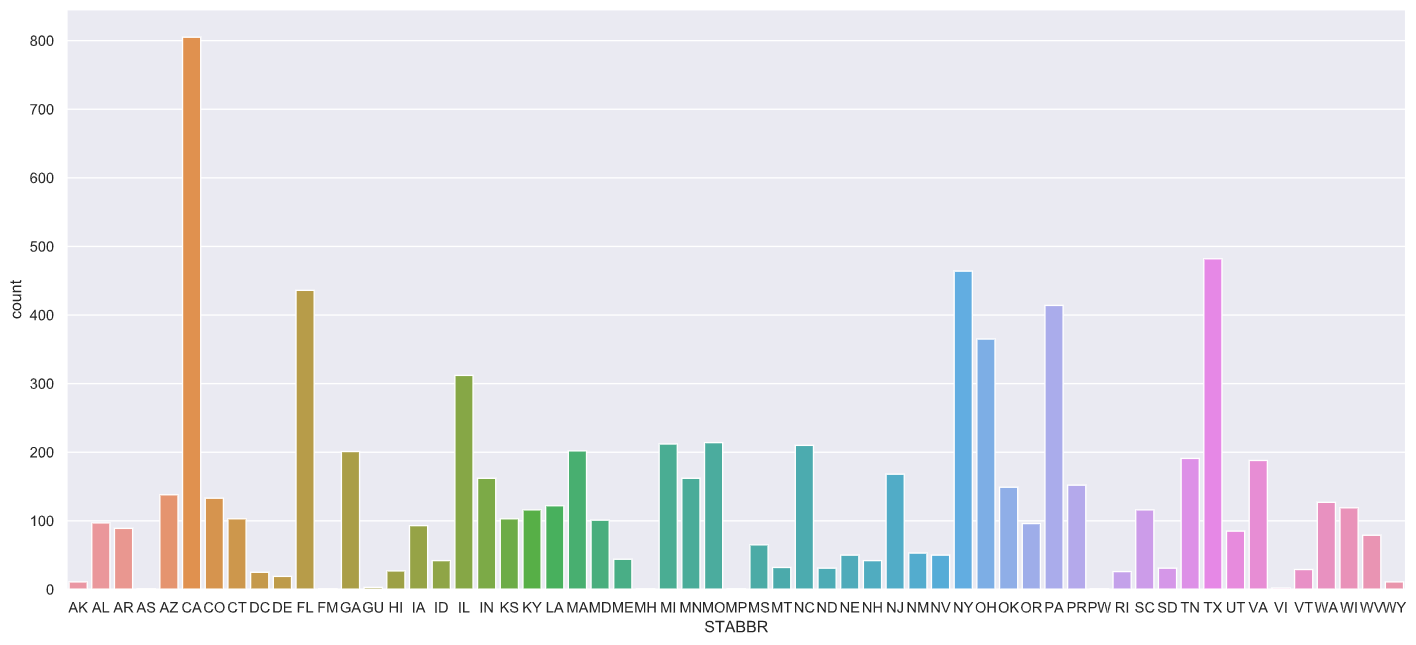


Below is each school’s students’ median income ten years after entering the school, regardless of degree or program completion and length of enrollment. Almost all school’s students’ median income ten years after entering the school is less than $60,000 annually.

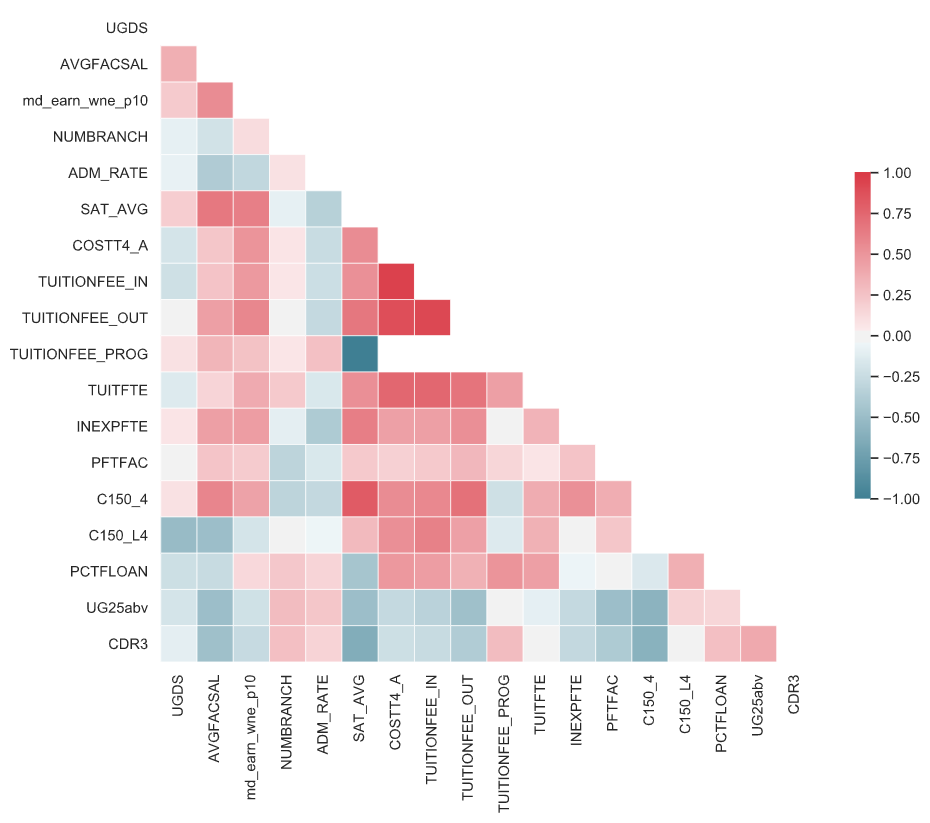


The chart on the right shows the proportion of colleges in the dataset that are public, private nonprofit, and private for profit. Almost three quarters of the colleges are private.

The chart below shows the distribution of colleges across the United States by state. In general, the number of schools in each state are scaled proportionally to the states population, with California, Texas, New York, and Florida having the most colleges, and Alaska, Delaware, Wyoming, and United States Territories having the fewest.



The correlation matrix below shows some of the more interesting relationships between variables. Tuition for program-based institutions (billed by full program rather than academic year, the variable TUITIONFEE\_PROG) has a strong negative correlation with average SAT score (the variable SAT\_AVG). The completion rate of 4-year or longer programs within 150% of the expected time (i.e. completion within 6 years for a 4-year program, the variable C150\_4) has a strong positive correlation with average SAT score. This can be interpreted as schools with higher SAT scores generally have higher completion rates.

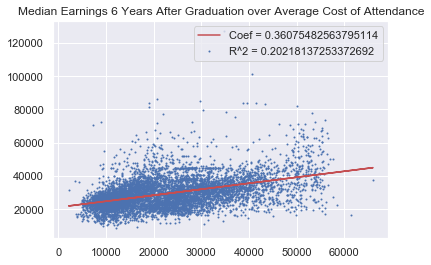


|  |  |
| --- | --- |
| AVGFACSAL | – Average Faculty Salary |
| NUMBRANCH | Number of Inst. Locations |
| md\_earn\_wne\_p10 | Median Annual Income after 10 yrs |
| ADM\_Rate | Admission Rate |
| SAT\_AVG | Average SAT Score |
| COSTT4\_A | Avg Annual Total Cost of Attendance |
| TUITIONFEE\_IN | In-State Tuition |
| TUITIONFEE\_OUT | Out-of-State Tuition |
| TUITIONFEE\_PROG | Program-based Tuition |
| TUITFTE | Net Tuition Revenue per FTE Student |
| INEXPFTE | Instructional Expenditures per FTE Student |
| PFTFAC | Proportion of Full-Time Faculty |
| C150\_4 | Completion Rate (4-year institutions) |
| C150\_L4 | Completion Rate (<4-year institutions) |
| PCTFLOAN | Proportion of Students with Federal Loan |
| UG25abv | Proportion of Students Aged 25+ |
| CDR3 | Proportion of Borrowers in Default after 3 years |

# Analysis

## Model 1: Linear regression model

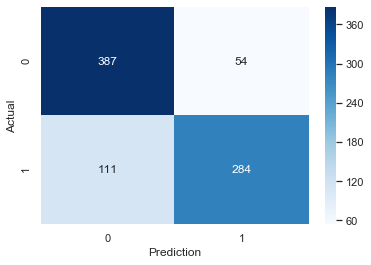
Using a simple linear regression model, we were able to plot with a minimal amount of success a prediction of median earnings based on cost of attendance.



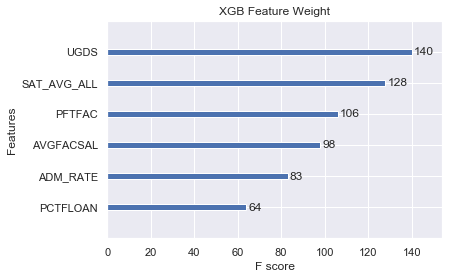
Due to the high number of variables in the dataset, it is clear that there is a larger story than the one variable of cost of attendance, but we do see an upward trend, but with a small R-squared it is difficult to draw any specific conclusions due to only 20% of the variation being accounted for.

## Model 2 - XGBoost Classification

When using XGBoost classification we were able to achieve marginal success when determining wether someone would be able to earn above the mean median income based on the college that they attended.



With a binary classification of above or below median, we were able to predict with 80% accuracy if attending a certain school will result in an above average income.



When looking at the feature weight, we can see that the variables with the most predictive strength include, the number of students, SAT average scores, Faculty salary, and admission rate.

# Recommendation

From our analysis, we can predict with marginal success which colleges will likely produce salaries that are higher than average.

We recommend to refine these models further in order to use this function when selecting schools.

We could market to apps such as Common Application in order to show new prospective students which school will give them more long term value, which is a current issue for students now.

\* This dataset requires significant analysis beyond the scope of this course to account for all 1,700+ columns and 13 years worth of data for thousands of schools.