College Scorecard Analysis of the Institutions of America

CSCI4502 project

Kyle Hartland Brown, Victoria Velasquez, and Rami AlQunaibit

May 1, 2018

Abstract

The goal of our project is to analyze a series of documents created by the Department of Education called "College Scorecards". These consist of various categories concerning the institutions of the United States, including but not limited to: Graduating Debt, Graduating Income, Transfer status during schooling, and acceptance rate(among many others). We analyzed the data in order to answer these questions:

- Do institutions with graduating debt tend to have higher graduating income?
- Do institutions with higher faculty salary tend to have higher levels of completion?
- Do institutions with higher acceptance rates tend to have higher levels of debt on graduation?
- Do institutions with higher graduation rates tend to have higher levels of debt on graduation?

The results all indicate that high levels of debts should be always be avoided. High levels of debt have a negative impact on all students even if they do not graduate. Acquiring a degree does not always pay off the amount of debt these students acquire, even in cases with smaller debts. Another finding was that schools that care for their teachers and give them higher salaries will have higher graduation rates. This shows a correlation between how committed the teacher are, as well as how good of teachers they are for their students.

Introduction

Our questions are focused mostly on graduating debt. With graduating debt becoming more concerning than ever in the United States, more and more students graduate with some amount of debt. Attending college is an investment, and some students think it is worthwhile to take out loans as they hope a bachelor's degree will lead to a fruitful career. We asked questions that can help guide people thinking about attending college and acquiring a degree. The importance of our questions are relative to the important of life after college, and the financial burdens a student could take on when they are starting to begin their life in the real world. Understanding the amount of loans someone can acquire as a student and relating it to income can lead people to make better decisions when deciding on their college career. These questions paint a clear picture of the financial risk students can take on.

Related Work

These are listed as Kernels on Kaggle.com. The most interesting of the kernels at first glance have been: "Admission"(https://www.kaggle.com/stuffypuppy/admission), "For

Whom the Pell Tolls" (https://www.kaggle.com/wrudebusch/forwhom-the-pell-tolls), "Are Affordable Schools a Good Deal?" (https://www.kaggle.com/michaelpawlus/rmarkdowndefault-script), "Escape From Poverty" (https://www.kaggle.com/wrudebusch/escape-from-poverty), and "College Earnings Premium & Value Proposition" (https://www.kaggle.com/apollostar/college-earnings-premium-valueproposition). "Admission" is a paper concerning the admission rates of schools in regards to the overall SAT scores of that school. They also analyze the different institutions with higher earning rates, whether STEM majors make admission more competitive at each institution, and the general price of each institution compared to it's admission rate. "For Whom the Pell Tolls" gives an excellent analysis of where students with pell grants are most likely to attend, here finding that online for profit colleges such as DeVry University get many of the students, as well as that the pell grant students still graduate with high debt. "Are Affordable Schools a Good Deal?" Was able to show a strong correlation between a high cost of attendance and a high repay rate of those students loans, as well as graphs showing a general lower repay rate based with attendance cost below $40,000. "Escape From Poverty" Sorts all colleges by their median earnings six years after graduation and compares that to whether or not they are first generation college students to determine if college was a positive decision for those graduates, their findings were that online for-profit schools keep the best track of their students, are high in first generation student percentage, and generally lead to a rather low salary, with none exceeding $70,000 a year on average. "College Earnings Premium & Value Proposition" was undoubtedly the most interesting concerning our project, as they went through and compared SAT scores, income, and various other attributes to create an Earnings Premium chart in which they predict the expected earnings of a group based on their degree and university.

These are also related sources, but different dataset:

- "Repayment of Student Loans as of 2015 Among 1995–96 and 2003–04 First-Time Beginning Students"

https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018410 The report presents estimates of both cohorts' cumulative borrowing, repayment, and default statuses.

- "Project on Student Debt"

https://ticas.org/posd/state-state-data-2015 State- and college-level data on student debt from federal and private loans, (data visualization).

- "Student Debt Help"

https://www.debt.org/students/ Accurate and accessible information online about financial well-being.

Proposed Work

We propose to sort through the data and determine how much can be learned about the average debt of each institution over the decades of study, and check to see if graduating income increases with level of debt. This will require cleaning the data of the many null values that appear, sorting through schools that have not included this information, and attempting to compensate for any lack of information by using updated college scorecards provided by the department of education. A big challenge will be to overcome the amount of "Privacy Suppressed" data that has been omitted by institutions that do not wish to have certain statistics made public. This is where we can set ourselves apart from previous projects, as no previous projects used the most up to date scorecards from the department of educations website, we will have access to four years more data, as well

as an opportunity to gain insight into what institutions suppress what information and correlate that with other information they may have provided. The privacy suppressed data seems to become more available in more recent years and could have some very interesting stories to tell about what institutions were charging high rates, but not giving students as large of an advantage upon graduation. Data Set https://collegescorecard.ed.gov/data/ Although the data-set is available on Kaggle, we have chosen to set ourselves apart from the other kernels on this website by acquiring the data directly from the department of education, this has a number of benefits, but mostly it has ensured that we are in possession of the most recent data possible. The categories that the data has available are extremely numerous, but are organized into: School, Academics, Admissions, Student, Cost, Aid, Repayment, Completion, and Earnings. All of which have dozens of sub-categories, for example Student can be expanded into Number of Undergrad Students, Race of Undergrads, Undergrad part-time percentage, Age, Income Brackets, First-Generation percentage, and FAFSA submissions. This results in close to a hundred individual categories that we can use over the hundreds of thousands of school entries to find as much meaningful data that can reasonably be acquired by the collegiate institutions of the United States. We also fully intend to start exploring other data sets as we answer our questions with the scorecards. For example, it may be interesting to search for data concerning the average income of residents in cities surrounding colleges to see if there is any affect on the loan rates and tuition costs due to poverty or wealth of certain areas.

Data Set

https://collegescorecard.ed.gov/data/

Although the dataset is available on Kaggle, we have chosen to set ourselves apart from the other kernels on this website by acquiring the data directly from the department of education, this has a number of benefits, but mostly it has ensured that we are in posession of the most recent data possible. The categories that the data has available are extremely numerous, but are organized into: School, Academics, Ad- missions, Student, Cost, Aid, Repayment, Completion, and Earnings. All of which have dozens of sub-categories, for example Student can be expanded into Number of Undergrad Students, Race of Undergrads, Undergrad part-time percentage, Age, Income Brackets, First-Generation percentage, and FAFSA submissions. This results in close to a hundred individual categories that we can use over the hundreds of thousands of school entries to find as much meaningful data that can reasonably be acquired by the collegiate institutions of the United States. We also fully intend to start exploring other data sets as we answer our questions with the scorecards. For example, it may be interesting to search for data concering the average income of residents in cities surrounding colleges to see if there is any affect on the loan rates and tuition costs due to poverty or wealth of certain areas.

The attributes that we mostly used from the 1729 columns are in Table 1.

Evaluation Methods

To evaluate our data the larges challenge will be actually sorting through it to find the percentages that could be correlated. Otherwise the entire data set is composed of percentages which we assume to be taken out of the total population of students for each university. So at that point we can reference how other papers evaluated things like the predicted income or debt, but most of our work in evaluating the data will be in checking the Confidence and Support of the relationships we hope to draw. At that point we will determine a minimum support that would make the data relevant and draw conclusions based on what these metrics tell us.

Tools

Our tools do not exceed the functions that:

- python
- numpy
- pandas
- Bash script

Main Techniques Applied

Note: The dataset is very huge for personal computer to handle with ease. With 1729 attributes and 18 files(millions of data points), a lot of work needed.

1. Cleaning

- Renaming all files year for easier integeration and sorting.
- Sort all files by names.
- Defining poorly organized variables
- Cleaning all variables for privacy supressed schools and nulls.

2. Classification

- Earnings:
  - High-income: \$75,001+
  - Med-income: \$30,001-\$75,000
  - Low-income: \$0-\$30,000
- Completion rate
  - High: $> 50\%$
  - Low: $< 50\%$

3. Limitations

The decision tree was limited by the inconsistency in the data types. Moreover, Pandas functions were not producing any significant or desired effect with the columns which have multiple data types. Hence, the decesion tree was useless and needed more development to a high degree of complexity, or some enterprise level libraries.

Key Results

- Support and Confidence:
  - 68.77% of schools released data about debt, which means that almost third of the data is missing. This makes our task harder to know more about debt. And 67.25% of the schools (that reports) have students with average debt $<$ \$15,000 (support: 0.6725). Of those we have 8% of schools have students with debt $<$ \$15,000 and with earnings greater than \$30,000 (confidence: (Debt $<$ 15,000) $=>$ (earnings $>$ 30,000) = 0.08)
  - Also only 28% of schools released data about completion rate (support: 0.2846). Of those 54% of schools have a completion rate of $>$ 50% (Support: 0.54). And 19% of schools with a completion rate $>$ 50% have earning $>$ \$30,000 (Confidence: Comp $>$ .50 to earnings $>$ 30,000 = 0.19)
- In Figure 1, the top graph represents high-income, middle graph represents med-income, and bottom graph represents low-income. The interesting trend is that high debt have low level income. On the other hand, the 0 debt have higher income range. This trend is seen in all three income levels.
- In Figure 2, the top graph shows a correlation between the faculaty salary and

the students' completion rate. In fact, above 15,000 the completion rate almost goes around 100%. In the middle graph, there is no indication of any realtion between the acceptance rate and the debt. In the bottom graph, there seems to be kind of trend between completion rate and debt. However there is no enough evidence to support that.

- Note: Salary is the faculaty monthly income, and Earnings are the annual graduate income in the graph.
- We used Figure 3 and Figure 4 to compare University of Colorado at Boulder with the Nation's mean. It is clear that CU Boulder is above average in all aspects. CU Boulder has higher debt than the national as well as higher earnings. Additionally, the faculaty salary has always been hire than the Nation's mean. In CU Boulder the faculaty salary in 1997 is higher than the Nation's highest point ever. Moreover, CU Boulder always had a higher completifon rate than Nation's best completion rate.

Applications

This project can be applied to various applications regarding college degree expenses. It can be seen as a good starting point for any person thinking about the debt that comes with the degree. For instance, the project can be expanded into a more accurate debt calculator for students to use. Moreover, the project also looked at the college degree as an investment. Thus, applications concerned with the benefits of degrees can use the project as well. For example, the ease of understanding the amount of debt and profit is very important to any person considering a college degree.

Appendix

Table 1: Attributes

Column	Value	Type
GRAD_DEBT_MDN	The median debt for students who have completed.	float
count_wne_inc3_p6	Number of students working and not enrolled 6 years after entry in the highest income tercile.	integer
ADM_RATE	Admission rate.	float
PREDDEG	Predominant degree awarded 0 = Not classified 1 = Predominantly certificate-degree granting 2 = Predominantly associate's-degree granting 3 = Predominantly bachelor's-degree granting 4 = Entirely graduate-degree granting	integer
DEBT_MDN	The original amount of the loan principal upon entering repayment.	float
AVGFACSAL	Average faculty salary.	integer
C150_4	Completion rate for first-time, full-time students at four-year institutions.	float
LO_INC_DEBT_MDN	The median debt for students with family income between $0-$30,000.	float
MD_INC_DEBT_MDN	The median debt for students with family income between $30,001-$75,000	float
HI_INC_DEBT_MDN	The median debt for students with family income $75,001+	float
RELAFFIL	Religous affiliation of the institution.	integer
COSTT4_A	Average cost of attendance (academic year institutions).	integer
mn_earn_wne_inc1_p6	Mean earnings of students working and not enrolled 6 years after entry in the lowest income tercile.	float
CONTROL	Control of institution 1 Public 2 Private nonprofit 3 Private for-profit	integer
mn_earn_wne_inc3_p6	Mean earnings of students working and not enrolled 6 years after entry in the highest income tercile.	float

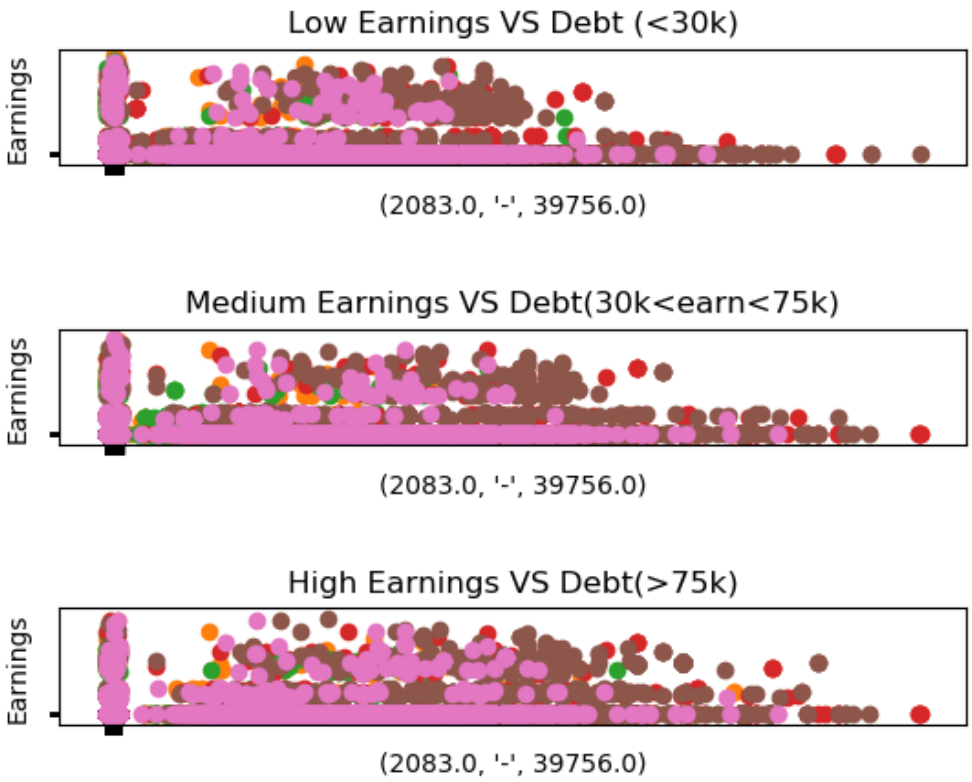

Figure 1: Debt relating to income

Figure 2: Acceptance, Completion, Debt

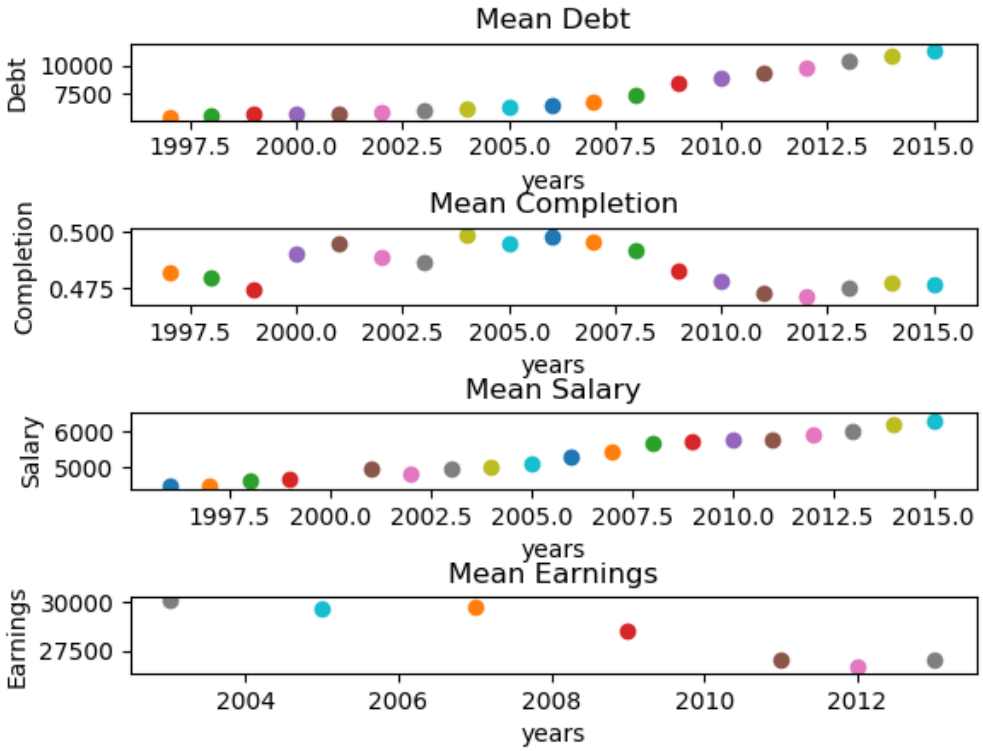

Figure 3: National Data

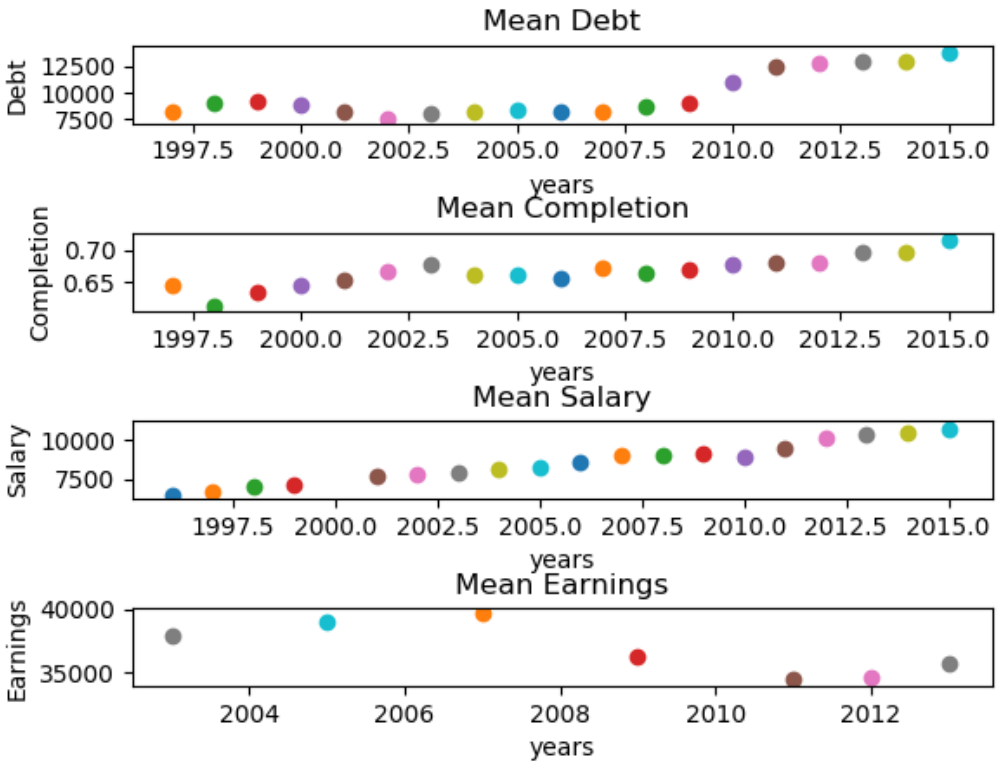

Figure 4: CU Boulder Data