College Scorecard Analysis of the Institutions of America

CSCI4502 project

Kyle Hartland Brown, Victoria Velasquez, and Rami Alqunaibit

April 30, 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 completed students tend to have lower levels of debt than not completed students?

The results all indicate that high debts should be avoided. High debts have negative impact on all student even if did not graduate. The degree does not always pay of the amount of debt as it is the case with high debts. In summary, borrow less than your projected annual salary.

Introduction

Our questions are focused mostly on the graduating debt. With graduating debt becoming more concerning than ever in the United States, more and more students graduate with debts. Attending college is an investment. Some students think it’s worthwhile to take out loans as their hope is that a bachelor’s degree will lead to a fruitful career. So, we wanted to ask questions that can help and guide the people thinking about a degree. We looked at the short-term investment, income after 6 years after graduating. This is not considerd very short, as students would like to pay of before their second year of working. The importance of our questions is relative to the important of life after college. Understanding the amount of loans and income in terms of number will make better decisions. In fact, the questions give an obvious picture of the finicial risk.

Related Work

- "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 used:

- $\text{GRAD}_{\text{DEBTMDN}}$

- count$_{\mathrm{wneinc3p6}}$
- ADM$_{\mathrm{RATE}}$
- PREDDEG
- DEBT$_{\mathrm{MDN}}$
- AVGFACSAL
- C150$_{4}$
- LO$_{\mathrm{INCDEBTMDN}}$
- MD$_{\mathrm{INCDEBTMDN}}$
- HI$_{\mathrm{INCDEBTMDN}}$
- RELAFFIL
- COSTT4$_{\mathrm{A}}$
- mn$_{\mathrm{earnwneinc1p6}}$
- CONTROL
- mn$_{\mathrm{earnwneinc3p6}}$

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.

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.

Clustering

Key Results

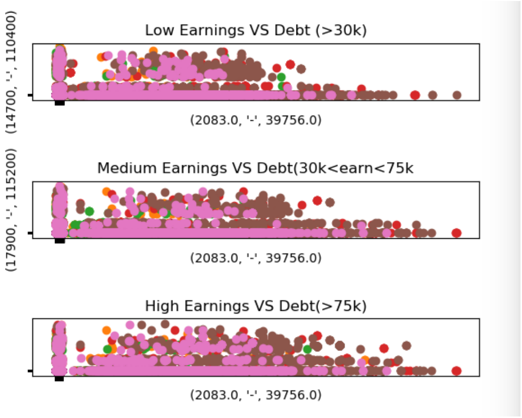

Figure 1: first image

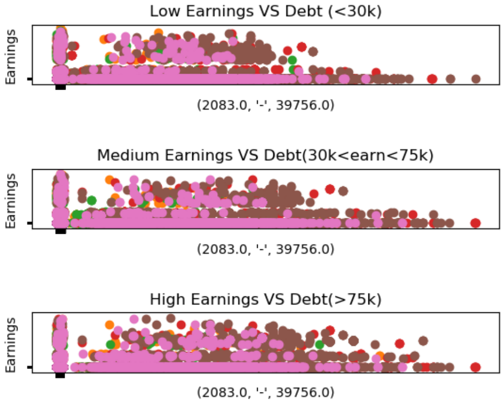

Figure 2: second image

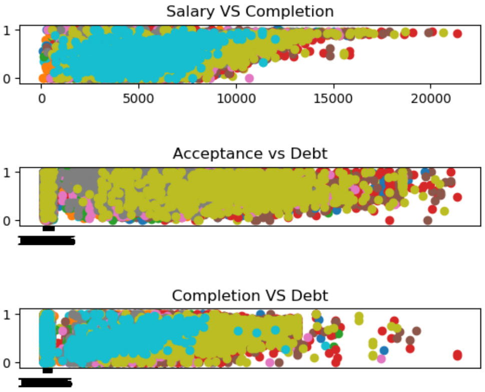

Figure 4: first image

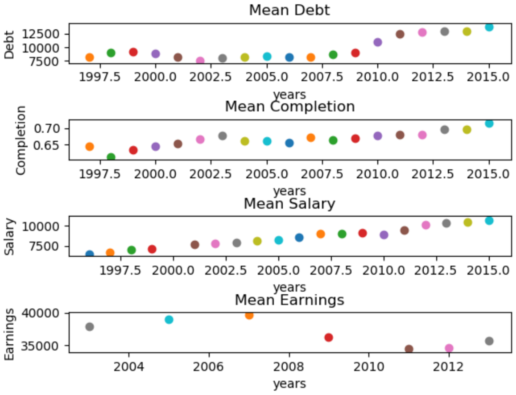

Figure 3: first image

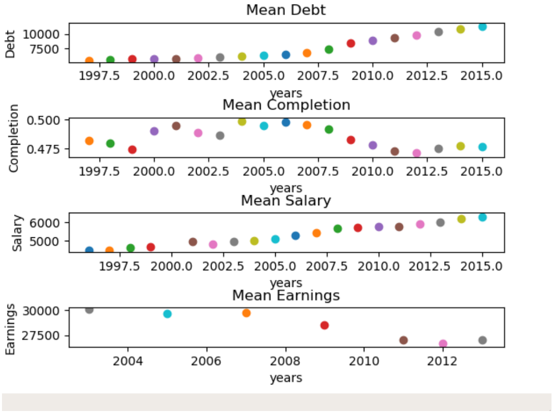

Figure 5: first image

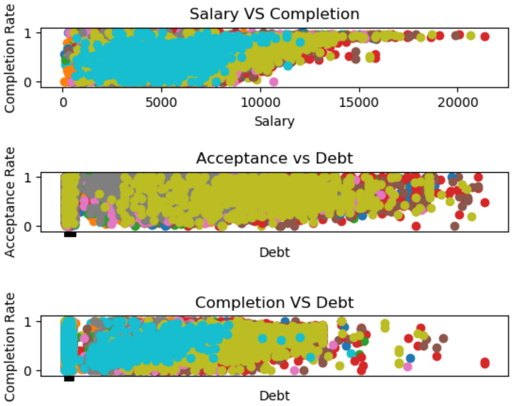

Figure 6: first image

Applications