College Scorecard Analysis of the Institutions of America

KYLE HARTLAND BROWN, VICTORIA VALESQUEZ, and RAMI ALQUNAIBIT

April 10, 2018

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). Our hope is to analyze this data and discover trends concerning average debt of different institutions and other categories listed. We hope to find answers to various questions, foremost among these being "Do institutions with graduating debt tend to have higher graduating income?", "Do predominantly graduate institutions tend to have higher levels of debt?", "Do predominantly graduate institutions' graduates tend to have higher income?", "Do institutions with higher faculty salary tend to have higher levels of completion?", "Do completed students tend to have lower levels of debt than not completed students?", "Do low income families tend to have lower levels of debt?", "Do special mission or religious affiliation institutions tend to have lower levels of debt?", "Do special mission or religious affiliation institutions tend to have lower cost?", "Do special mission or religious affiliation institutions' graduates tend to have lower income?", "Do private for-profit institutions' graduates tend to have higher income?", "Do private nonprofit institutions tend to have lower levels of debt than private for-profit?" and "Do institutions with higher acceptance rates tend to have higher levels of debt on graduation?". We also hope to discover many more interesting correlations in the data sets as we explore the hundreds of columns of information collected about every institution.

Previous Works

The previous work done on this data set is extensive, and should prove incredibly useful for our project. 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-premiu "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 corre-

lation 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.

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.

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 projected needs for tools do not currently exceed the functions that numpy, pandas, and other python libraries are able to provide for us concerning the calculations we need to make and how we want to sort through as well as manage our data.

Milestones [3/5]

- DONE 3/9 - have completely explored the data and determined all questions we would like to answer and whether null values or omitted data will be a problem.

- DONE 3/18 - be close to or finished with cleaning all data of null values, and have decided how to handle privacy omitted data.

- DONE 3/24 - determine if we wish to use other data to draw more information about the colleges.

- TODO ~~4/1~~ 4/11 - be close to if not done with preliminary findings, start creating graphs and graphics, incorporate new methods learned in class if desired.

Questions	Calculations
Do institutions with graduating debt tend to have higher graduating income?	support("GRAD_DEBT_MDN" => "count_wne_inc3_p6") confidence("GRAD_DEBT_MDN" => "count_wne_inc3_p6") lift("GRAD_DEBT_MDN" => "count_wne_inc3_p6")
Do institutions with higher acceptance rates tend to have higher levels of debt on graduation?	support("ADM_RATE=high" => "GRAD_DEBT_MDN=high") confidence("ADM_RATE=high" => "GRAD_DEBT_MDN=high") lift("ADM_RATE=high" => "GRAD_DEBT_MDN=high")

Questions	Calculations
Do predominantly graduate institutions tend to have higher levels of debt?	support("PREDDEG=4" => "DEBT_MDN=high") confidence("PREDDEG=4" => "DEBT_MDN=high") lift("PREDDEG=4" => "DEBT_MDN=high")
Do predominantly graduate institutions' graduates tend to have higher income?	support("PREDDEG=4" => "count_wne_inc3_p6") confidence("PREDDEG=4" => "count_wne_inc3_p6") lift("PREDDEG=4" => "count_wne_inc3_p6")
Do institutions with higher faculty salary tend to have higher levels of completion?	support("AVGFACSAL=high" => "C150_4") confidence("AVGFACSAL=high" => "C150_4") lift("AVGFACSAL=high" => "C150_4")
Do completed students tend to have lower levels of debt than not completed students?	mean("GRAD_DEBT_MDN") mean("WDRAW_DEBT_MDN")
Do low income families tend to have lower levels of debt?	mean("LO_INC_DEBT_MDN") mean("MD_INC_DEBT_MDN") mean("HI_INC_DEBT_MDN")
Do special mission or religious affiliation institutions tend to have lower levels of debt?	support("RELAFFIL" => "DEBT_MDN=low") confidence("RELAFFIL" => "DEBT_MDN=low") lift("RELAFFIL" => "DEBT_MDN=low")
Do special mission or religious affiliation institutions tend to have lower cost?	support("RELAFFIL" => "COSTT4_A=low") confidence("RELAFFIL" => "COSTT4_A=low") lift("RELAFFIL" => "COSTT4_A=low")
Do special mission or religious affiliation institutions' graduates tend to have lower income?	support("RELAFFIL" => "mn_earn_wne_inc1_p6") confidence("RELAFFIL" => "mn_earn_wne_inc1_p6") lift("RELAFFIL" => "mn_earn_wne_inc1_p6")
Do private for-profit institutions' graduates tend to have higher income?	support("CONTROL=3" => "mn_earn_wne_inc3_p6") confidence("CONTROL=3" => "mn_earn_wne_inc3_p6") lift("CONTROL=3" => "mn_earn_wne_inc3_p6")

Questions	Calculations
Do private nonprofit institutions tend to have lower levels of debt than private for-profit?	support("CONTROL=3" => "DEBT_MDN=low") support("CONTROL=2" => "DEBT_MDN=low") confidence("CONTROL=3" => "DEBT_MDN=low") confidence("CONTROL=2" => "DEBT_MDN=low") lift("CONTROL=3" => "DEBT_MDN=low") lift("CONTROL=2" => "DEBT_MDN=low")

- TODO ~~4/10~~ 4/18 - take care of any unforeseen changes and create project final report when ready.

Results so far

At this point of the project, we finished researching the 18 data sets, and grasped the main idea of how we are going to analyze the data. We completed:

- The data set is cleaned and the incorrect or irrelevant attributes were replaced and modified.

- The NULL values for each university were filled using previous years.

- All universities that had too many empty scorecards were removed.

- All the questions for the project were linked directly to the needed calculations for the analysis.

- Working on the data integration of the 18 sets, and combining them all into a single unified view by year.