Stat159 Project 3: College Recommendation

Liang Hao, Bret Hart, Andrew Shibata, Gary Nguyen

December 5, 2016

1 Method

In this section, we continue the discussion on how we decided upon the composite "quality" metric for high-quality, low-cost education for at-risk students. We later move on to their weighting and actual method implementation in the shiny app.

One of the most common reasons students cite in choosing to go to college is the expansion of employment opportunities. This is of even greater intuitive importance to poorer students, who absolutely cannot afford to go to an expensive school that may leave them in permanent debt. To that end, data on the earnings and employment prospects of former students can provide key information. To measure the labor market outcome of individuals attending institutions of higher education, data on cohorts of federally aided students were linked with earnings data from de-identified tax records and reported back at the aggregate, institutional level. This dataset, however, is separate from the main College Scorecard dataset, and is included in a different .csv, *Most-Recent-Cohorts-Treasury-Elements.csv* As mentioned earlier, although these datasets were separate at the start, we merge the two to allow for easy access to pertinent information that may only be contained in one of the datasets. Obvious data of interest only within this dataset are median earnings and median debt after graduation.

So - we choose a set of columns that we feel, in aggregate, can represent in some small, objective measure, a school's quality - now, how do we actually create this score in practice? First, we standardize all of the included variables within their respective columns, then take z-scores, and use these z-scores in place of the actual values within each column. Thus, the average of any statistic will be given a score close to 0, with scores above given a more positive score, and scores below a more negative score. Then, we weight each metric with deliberation, experimentation, and consultation of literature. Importantly, some data points are actually desirable when substantially *below* the average, such as net price or median debt, while others are obviously better if higher, such as median earnings or completion rate. Thus, we actually weight some

pieces positively and other negatively in regards to the sign of their z-scores. Finally, these scores will alter and adjust both weighting and actual information as students enter and change their inputted answers.

Eventually, we decided upon the following for our data included in the response and their weights:

1 - Net Price, stratified by income. There are 5 income brackets within the dataset, so dependent upon which income bracket a student inputs, the score is recalculated to consider the relative z-score of their projected respective net cost of attendance. For students below a certain income level, we actually want to more heavily reward schools with very negative z-scores, as these schools, in totality, are anomalously cheap when compared to their counterparts. We decide that for lower income students, this is of more significant importance, as this is the net price of attending AFTER financial aid - and these are students who need to know exactly how much attending could cost. Thus, the columns the information is drawn from and the weights themselves are dynamically adjusted based on user input.

	Variable	Weight	Condition
1	NetPrice	-0.05	Default
2	NetPrice	-0.3	Income in $0-$75,000
3	NetPrice	0.4	Income in $75,000+

Table 1: Table of Net Price by conditional weights

2 - Repayment rate, split into 3 year repayment rate and 5 year repayment rate. To clarify, this column represents the proportion of students who have completely paid off their debt to the school after 3 and 5 years. Specifically, we choose the columns of repayment rate by completers, as we do not wish to muddle the recommendation algorithm with students who are struggling to pay back their loans because they did not actually finish their education. While this could be interpreted as an oversight, we believe we are discussing college admissions with students who understand the hardships that getting a college education entails, and are willing to take the risk of attempting to complete. It would be unfair to their passion to attend to include the repayment rate for students who weren't able to finish - we presume that they are willing to take the risk and want to know the information pertaining to the goal they're trying to achieve, not what could happen. To eliminate variance, we include both the 3 year and the 5 year repayment rate at equal weighting in the final response quality metric.

3 - Completion rate, in 150% time (4+2 years) for all students at that particular institution. Of course, this is a loaded statistic, and although we briefly touched upon the discussion around which students are most likely to complete college and which are likely to not, we did brush over that the students we are working to accomodate are at-risk, and thus, are significantly more likely

	Variable	Weight	Condition
1	3 Year Repayment Rate	0.1	Default
2	5 Year Repayment Rate	0.1	Default

Table 2: Table of Repayment Rate by weights

to struggle or possibly need to drop out. So, admittedly, this metric may be imperfect as it is not stratified to completion rate of students within specific income brackets. However, we believe it can still be appreciated and is important, perhaps indicative of some broader communal feeling at the school, and is worthy of inclusion.

	Variable	Weight	Condition
1	4 Year Completion Rate	0.7	Default

Table 3: Table of Completion rate by weight

4 - Mean and median earnings, 10 years after beginning college. This is another imperfect statistic to include, as to study earnings 6 years (in theory) after graduation, we must use data from students who started college 10 years ago, which is outside of the framework of our data set. However, because it is contemporary information and is often referred to as a significant and crucial piece of measuring a more objective sense of monetary quality of university attendance, we include it despite its flaws. We use both the mean and median earnings 10 years after beginning, again, to reduce variance and to legitimize claims of a potential expectation of salary in the future. While this may not be why every student attends college, it is important to a large number of students and a good indicator of the capability for social mobility provided by an institution - which is exactly what we seek to measure and return.

	Variable	Weight	Condition
1	10 Year Mean Earnings	0.7	Default
2	10 Year Median Earnings	0.7	Default

Table 4: Table of Mean/Median Earnings by weights

5 - Percentage of the student body which has/could have a Pell Grant. This is a metric entirely new to us, but is discussed extensively in educational literature as a prescient indicator of built-in, intentional, positive institutional affordances to low-income students. The Pell Grant is a federal loan given to extremely low income students to allow for their attending college with very little to begin with. Thus, scholars reason that schools who integrate into their agenda encouragement and enabling of Pell Grant student attendance may have more amorphous, hard-to-capture support systems in place as well which could allow these students to succeed. In short, they reason that if a school intentionally lets in a bunch of kids with Pell Grants, they'll have programs specifically

tailored to those kids to do well. This is a very positive indicator for at-risk, low-income, minority students, so we include it. To allow for students of all income brackets to use our score generator, though, we only include the Pell Grant percentage for students in lower income brackets, and ignore it for students who do not fit a reasonably similar criterion.

Additionally, the existence of a community of people similar to you at a school is vitally important to success, and if a school has a high number of Pell Grant students, it may reduce the shame or uncomfortability that stems from being from a household of very low income - being around people who have come from similar hardships to you has proven time and time again to help in dramatic ways.

	Variable	Weight	Condition
1	Pell Grant	0.1	Default
2	Pell Grant	0.5	Income in $0-$30,000

Table 5: Table of Pell Grant Percentage by conditional weights

6 - Median debt after attending. This is another statistic that is very difficult to generalize out to all students, but we incldue it anyway as it is of obvious relevance, regardless of equal applicability to all students. This is a very real, concrete number and even if it is unfair to create a median debt over the entire student body, it seems to have some legitimacy when compared to other schools. If a school has a higher median debt, for example, this could either be an indicator of a lack of substantial financial aid afforded to students, or, more interestingly, suggestive of a student body which is generally lower-middle class - able to sort-of afford attendance, not of low enough income to get substantial financial aid, not of high enough income to have no debt at all. Educational scholars also suggest that it is important to include, so we do so, and it plays a minor role in our final composite metric.

	Variable	Weight	Condition
1	Median Debt	-0.1	Default

Table 6: Table of Median debt by weight

7 - First generation student percentage. This is included for very similar reasons to the Pell Grant percentage - if an institution is willing to take on a large number of students, they likely have programs in place to accomodate them. It also suggests that a school will have many other first generation students who would be able to provide a support network. Like the Pell Grant percentage, we only include the variable if a student indicates that they are a first-generation student. We believe that having such a choice for the users would be considerate. They have the freedom to negate it as well.

8 - Intended field of study. During the iterative process of creating this app, we ran into issues with weighting and variable inclusion. At one point, our app

	Variable	Weight	Condition
1	First Generation	0	Default
2	First Generation	0.2	First Generation

Table 7: Table of First Generation percentage by conditional weights

only recommended art schools, partially due to errors in weighting and partialy due to a lack of narrowing the application field by limiting the search to schools which actually had some semblance of a program you were interested in. Thus, we boost a college's score if it offers the program that the student is interested in, without discounting other good schools which may not exactly offer that major.

	Variable	Weight	Condition
1	Major	0.1	Default
2	Major	0.3	Selected major

Table 8: Table of Field of study by conditional weights

9 - Similar Ethnicity percentage. As discussed in previous segments, a school which has real communities of students of many backgrounds and diverse origins is important to both the quality of the school but also to minority students who may want to be near people of the same Ethnicity. This is a simple metric, and is only weighted upon the ethnicity that a student says they are a member of. Thus, if I enter that I am a black student, it will use the z-scores of the percentage of the student body that is black and dynamically alter the weights to favor schools which have higher percentages of black students. Again, having sizable minority percentages may also imply that the school has other affordances to at-risk students who may need the help.

	Variable	Weight	Condition
1	Ethnicity	0	Default
2	Ethnicity	0.1	Selected ethnicity

Table 9: Table of Ethnicity Percentage by conditional weights

10 - Test Scores. While we only begrudgingly include this piece in a minor way, it is admittedly still important to not reccomend students schools which are substially out of their reach via a large test score mismatch. While we're dubious of its status as actually an indicator of quality in any way, pragmatically, schools do have cutoffs on scores and we don't want these at-risk students to waste their time applying to programs they may have no chance of being accepted to. However, as we don't want to rule out any schools completely, we only slightly weight up a school if it fits the entered score threshold.

Thus, we arrive at our 10-section composite metric of school "quality", which is especially geared towards at-risk, low-income, minority students, but can be

	Variable	Weight	Condition
1	Test Score	0	Default
2	Test Score	0.1	Input test score

Table 10: Table of Standardized Test Scores by conditional weights

adjusted to fit any student's needs reasonably well. We chose these criterion based upon deliberation, experimentation, and research.