Opportunities for Investing in Success for African American and Hispanic College Students

Stacy Chang, JD de Lorimier, Tim McGinley, Philhoon Oh

December 5, 2016

1 Abstract

This report for your organization uses the national College Scorecard data to make recommendations for the best allocation of funds for the benefit of African American and Hispanic students. This report builds a proprietary model to suggest what the graduation rates of these groups *should* be based on other data from College Scorecard, and compared these results to the actual graduation rates for these demographics, identifying schools that are underperforming in these areas. We then take a closer look at selected schools from this category to identify trends and specific institutions whose minority students most need the support of your organization.

2 Introduction

There can be no doubt that, in an increasingly divisive and frightening time for the future of our nation's most vulnerable minorities, ensuring the successful college education of our youth is more important than ever. The mission of your organization to promote the educational achievement of minority students is vital to the continued march towards equality in the United States, and this report is designed to help your organization target its capital investments where they are needed most. There are over 2,000 four-year institutions in this country serving our nation's youth, and we've done the leg work for you in finding schools at risk for low graduation rates for African American and Hispanic students.

3 Data

The data originally comes from the Research Triangle Institute, part of the U.S. Department of Education. This is the data that powers `http://collegescorecard.ed.gov`, a website to help college-bound students find the best schools for their interests and other preferences. While this provides a great resource for students, it also has fostered the collection of an incredible amount of data on the over 7,700

places of higher learning in the country, which we can use to answer deeper questions. Data points are included for types of school, graduation rates, admitance rates, student body demographics, and literally thousands more.

The original data file spans 20 years and has over 1700 variables—over 1700 distinct pieces of information about each school. This results in a file simply too large to be practical to include, so we've extracted just the last three annual iterations of the 27 variables that had possible relevance to the production of the information contained in this report. This resulted in the more manageably sized `data/subset-data.csv` data set, which we have included with this report.

4 Exploratory Data Analysis

Exploring the data set we have been given was an important first step in this analysis. Our EDA started with a simple look through the Data Dictionary to learn about the definitions of the over 1700 variables: what they represent, what kind of data they contain, and how to find them.

After determining which variables we will use in our model-fitting and analysis, the next stage was to obtain a clearer image of the data we will be working with. First, let's take a broad look a the distribution of minority students at four-year institutions nationally. The following figure shows the wide disparity of demographic makeup in four-year schools around the country—while non-white students generally comprise a small percentage of the student body at most schools, there are many extreme outliers.

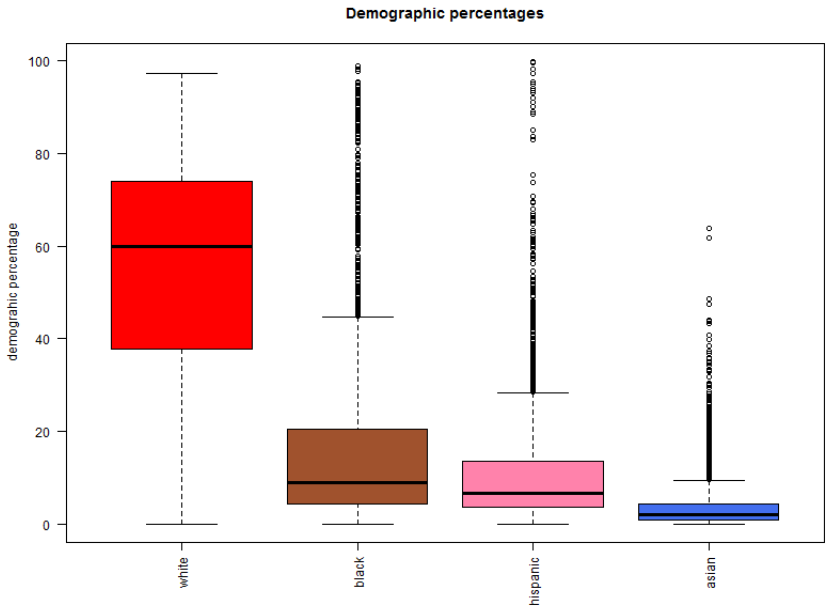


Figure 1: School-by-School Demographic Makeup

Ultimately, the objective of this initiative concerns the graduation rates of minority students, so it seems reasonable to look at the overall rates for these groups in four-year institutions nationwide.

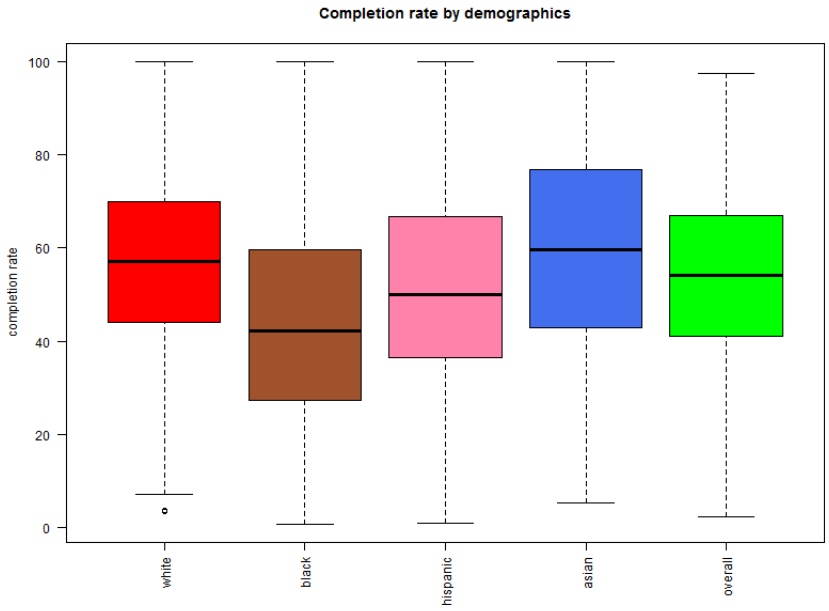


Figure 2: Overall Six-year Completion Percentage By Race

This side-by-side boxplot shows what we already know: white and asian students graduate from college at noticeably higher rates than their black and hispanic counterparts.

Looking more closely at completion rates for minorities, we see that the overall completion rate by school tracks quite closely for white students:

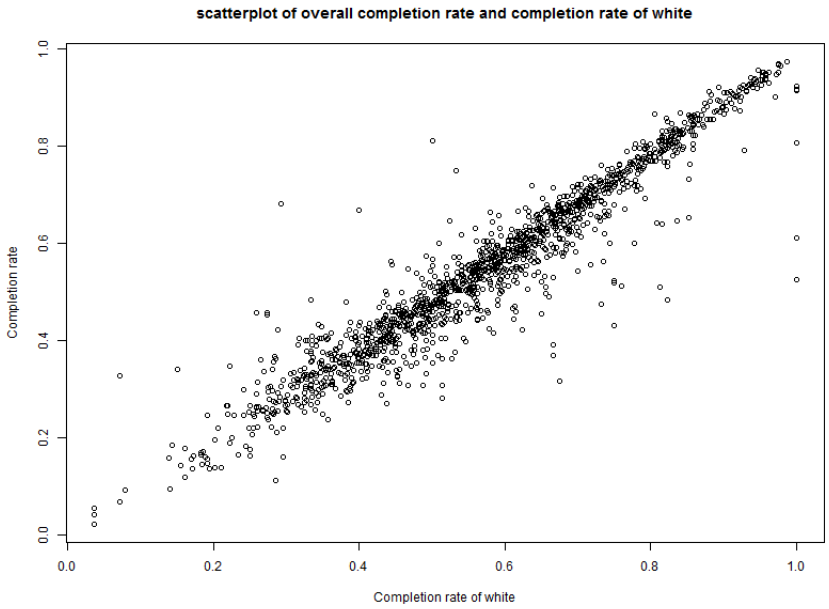
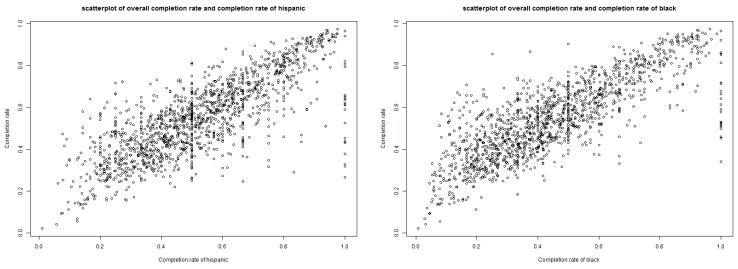


Figure 3: Overall Completion Rate vs White Completion Rate

...but decidedly less closely for African-American and Hispanic Students:



(a) Hispanic (b) Black

Figure 4: Minority Completion Rate vs White Completion Rate

To sum up the relationship between the African-American and Hispanic individual graduation rates and other important selected variables, the following scatterplot matrix was created

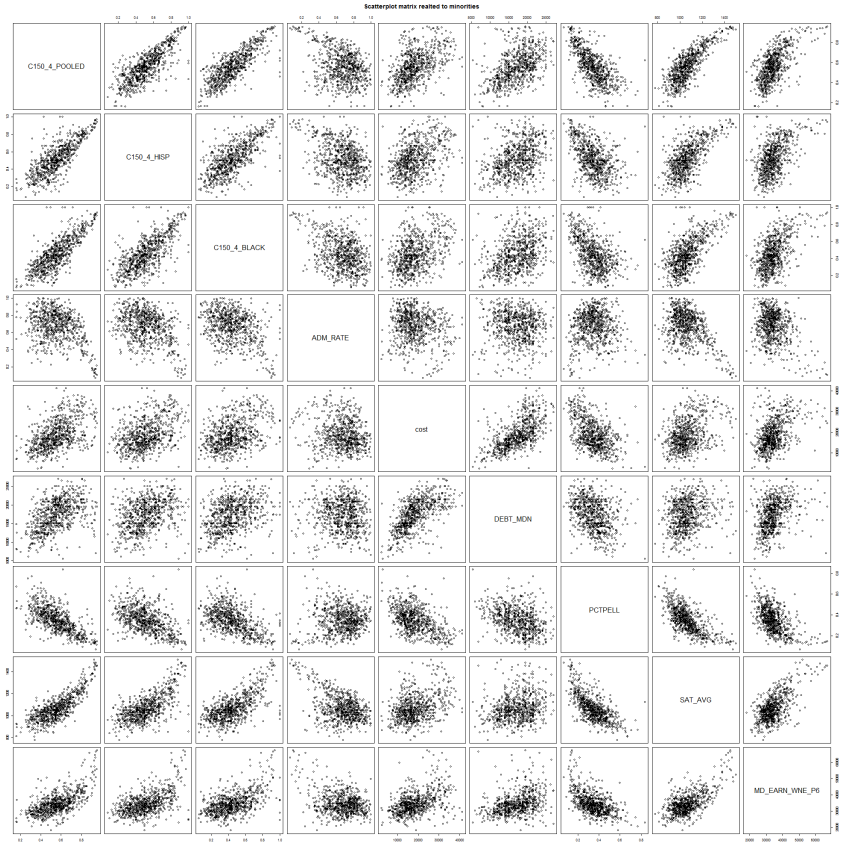


Figure 5: Scatterplot Matrix Comparing Relationships of Variables with Minority Graduation Rates

For more information on the data we used, extensive summary statistics and a correlation matrix of all variables used in this project can be found in the `data/eda-output.txt` file. Many informative histograms and scatterplots that didn't make the cut into this report can also be found in `images/`.

5 Methodology

To conduct the analysis, first the data was loaded from the manageably sized `data/subset-data.csv` and some exploratory data analysis was performed (covered in the next section) on the variables contained.

Our primary goal with our analysis of the data contained was to create a model that relates the data we have to the graduation rate for the minorities we are looking at: African-Americans and Hispanics. This will allow us to identify institutions that should be graduating minorities at a higher rate then they are

- an institution whose students could clearly use more support.

To decide on the best model to use to predict the expected graduation rates of these minorities, we built both a normal least-squares model and a ridge regression model using a training set of the previous few years of data, before testing the performance of the models using a testing set comprised of the latest data in the table. By comparing the Mean-Squared error of the predicted values compared to the test set, we determined which type of regression was better to use. Following is a brief discussion of the two techniques.

In an ordinary linear squares regression (OLS), we assume that the relationship between the predictor variables and the minority graduation rates can be described roughly by the equation:

$$GR = \beta_0 + \beta_1(A) + \beta_2(B) + \beta_3(C) + ... \tag{1}$$

Where β_p are random variables determined from the data by a least-squares fit called "coefficients" that scale the values of predictor variables $A, B, C...$ so that the graduation rate GR is optimally close to the true value. The function `lm()` determines these variables from the data for us, and this information as well as a summary are located in `data/ols-black-model.Rdata` and `data/ols-hisp-model.Rdata`.

A slight variation on regular least squares regression (which seeks to minimize the residual sum of squares, $RSS = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{p}\beta_j(x_{ij}))^2$), ridge regression (RR) adds a term to the end of the expression called a shrinkage penalty: $\lambda\sum_{j=1}^{p}\beta_j^2$. λ in this case is called a "tuning parameter" and is ultimately chosen by the model-fitter, ideally as the value that minimizes the variance of the estimates as much as possible while maintaining a threshold of bias. This leads to the final expression that is minimized in ridge regressions,

$$\sum_{i=1}^{n}(y_i - \sum_{j=0}^{p}\beta_j(x_{ij}))^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{2}$$

In practice, computing power is used to calculate the estimates for many different values of λ at once to allow for easy discovery of the 'ideal' value for each, and this is what we have done. This shrinkage penalty is a necessarily positive term that increases as λ does. The minimum value the above expression takes therefore must have smaller coefficients β_j than simple linear regression, hence "shrinkage".

After fitting our models, we compared the predicted values for minority graduation rates with the actual values in the table and found the schools which fell short.

6 Analysis

Using the data from the previous two years of this report, '12 and '13, we built both an OLS model and a Ridge Regression model to predict the African-American and Hispanic graduation rates for each school based on 12 other re-

lated variables in the dataset. These variables are listed below, as well as a brief description.

Table 1: Variables Used in Model

Variable Name	Description
UGDS_BLACK	Black Student Body Percentage
UGDS_HISP	Hispanic Student Body Percentage
COMPL_RPY_5YR_RT	5-year Overall Completion Rate
NPT4_PUB	Net Price (Public and Private Schools Combined)
ADM_RATE	Admission Rate
RET_FT4	4-Year Retention Rate
DEBT_MDN	Median Post-Graduate Debt
PCTPELL	Percentage of Student Body on Pell Grants
MD_EARN_WNE_P10	Median Post-Graduate Earnings
grad_total	Total Graduation Rate

Note: Only the corresponding demographic percentage was included for each model—i.e. our Hispanic model wasn't based on UGDS_Black at all

After we have built and trained our two models using the data from previous years, we tested them with the current year's data to determine which model produced the lowest error, measured via the Mean-Squared Error.

Group	OLS MSE	Ridge MSE
Black	1.129	0.453
Hispanic	1.038	0.432

Table 2: MSEs by Race and Method

For both minority groups we looked at, Ridge regression far outperformed OLS, so we used our ridge model throughout.

Our goal is to uncover the schools that this model predicts would have a higher graduation rate for minority students than it does, but first, we needed to make some logical restrictions on the data. First, a school with a vanishingly small percentage of undergraduates that fall into one or the other category not only leads to small sample size issues, but also doesn't seem like the type of institution worth investing capital in anyway. So we set a logical cutoff, requiring at least of 3% of the student body to fall into the racial category of interest. Secondly, we only included schools with at least 500 students in the entire student body. Finally, we restricted this search to schools classified as either "Highly Selective" or "Selective" by the US Department of Education—an easy catchall for schools holding good academic credentials.

This results in the following list of the 35 schools whose African-American graduation rates fall the farthest short of what our model predicted, and that meet our criteria for underperforming schools in regards to African-American graduation rates:

Table 3: Top 35 Most Underperforming Qualified Schools for Black Students

School Name	Black Grad Rate
Asbury University	0.14
Valparaiso University	0.22
Union University	0.23
Austin College	0.33
Wittenberg University	0.22
Rockhurst University	0.35
Kansas State University	0.24
Assumption College	0.36
Embry-Riddle Aeronautical University-Daytona Beach	0.27
Pratt Institute-Main	0.33
The College of Saint Scholastica	0.33
Millsaps College	0.34
University of Missouri-Kansas City	0.25
Albion College	0.43
Drexel University	0.44
Saint Louis University	0.46
Cornell College	0.36
University of Cincinnati-Main Campus	0.34
Eckerd College	0.41
Oklahoma City University	0.37
New Jersey Institute of Technology	0.39
University of San Francisco	0.50
Augustana College	0.53
Seton Hall University	0.46
Texas Christian University	0.58
Florida Institute of Technology	0.39
Rowan University	0.50
University of Delaware	0.59
Ohio Northern University	0.48
Hofstra University	0.45
University of Minnesota-Twin Cities	0.58
Illinois Wesleyan University	0.59
Siena College	0.59
Bradley University	0.54
University of North Carolina School of the Arts	0.40

And the same for Hispanic students:

Table 4: Top 35 Most Underperforming Qualified Schools for Hispanic Students

School Name	Hispanic Grad Rate
Albion College	0.00
Milligan College	0.17
Wabash College	0.36
Elizabethtown College	0.38
Calvin College	0.46
Oklahoma Christian University	0.26
University of Missouri-Kansas City	0.31
Westminster College	0.43
University of Puget Sound	0.54
Illinois Wesleyan University	0.56
Embry-Riddle Aeronautical University-Daytona Beach	0.39
Wagner College	0.45
Embry-Riddle Aeronautical University-Prescott	0.41
Allegheny College	0.57
Florida Institute of Technology	0.39
Wentworth Institute of Technology	0.46
Lipscomb University	0.44
The College of Wooster	0.56
Rochester Institute of Technology	0.48
Clark University	0.59
Harding University	0.42
Oglethorpe University	0.40
University of Maryland-Baltimore County	0.47
Belmont University	0.52
Hofstra University	0.50
University of Rochester	0.69
Christopher Newport University	0.51
Montana State University	0.33
Loyola University Maryland	0.69
State University of New York at New Paltz	0.58
Agnes Scott College	0.57
Rowan University	0.56
Arcadia University	0.48
University of Wyoming	0.38
Whitworth University	0.59

7 Results

The tables in the section above show the schools that our model has shown do not graduate African-American and Hispanic students at the rates they should—that meet our inclusion criteria of school size and demographic proportions. This information alone does not fully answer the ultimate question, though, of which of these schools are the best places to invest scholarship money. Now that we have a manageable selection of schools to choose from, let's dive a little deeper and look at which schools make appealing targets for resources.

The above lists are the top 20 most underperforming schools for each demographic, but far more schools than that do not meet expectations (our model found 216 and 256 for African-Americans and Hispanics, respectively. The full list can be found in textttdata/results-tables.Rdata). Extending this list to the 100 most underpeforming schools, the following five schools from each category have the lowest cost of attendance—in other words, schools where scarce scholarship dollars will go the farthest, making them better investments. Each dollar invested here will go farther towards improving graduation rates by simple result of a lower overall cost.

Table 5: Five Cheapest Underperforming Schools For Black Students

School Name	Net Cost
Arizona State University-Tempe	10858
The University of Texas at Dallas	12050
San Diego State University	12567
University of North Carolina School of the Arts	13625
University of Illinois at Chicago	13811

It's worth noting here that these are all state schools—the financial benefits of offering scholarships at these schools is large, but you must be careful to only offer them to in-state students. Overall, these five schools represent selective or highly selective institutions where African-American students need extra help to graduate, and that help can be most cheaply provided. Following is the same table for Hispanic students:

Table 6: Five Cheapest Underperforming Schools for Hispanic Students

School Name	Net Cost
University of Wyoming	11603
San Diego State University	12567
Stony Brook University	13519
University of North Carolina School of the Arts	13625
University of California-San Diego	14136

As you might imagine, there is some overlap between the two lists.

Next, we looked at which of these schools which boast the highest median income 8 years out of school. If the goal of investing money into minority education is to improve the economic status of oppressed peoples, these schools do the best job of that, of the schools we've identified as underperforming.

Table 7: Five Best Underperforming Schools By Eventual Income For Black Students

School Name	Median 8-Year Income
Georgetown University	77100.00
Harvard University	75200.00
Duke University	72600.00
Yale University	70900.00
Carnegie Mellon University	70500.00

These should come as no surprise: as some of the best schools in the country, they often provide their graduates with high incomes. It is more notable that these schools have been identified as not meeting their expected graduation rates for African-Americans. This suggests that black students at these schools could also use help graduating. The following is the same table for Hispanic students, of which similar things can be said:

Table 8: Five Best Underperforming Schools By Eventual Income For Hispanic Students

School Name	Median 8-Year Income
Carnegie Mellon University	70500.00
Rensselaer Polytechnic Institute	69900.00
Villanova University	65000.00
Cornell University	64300.00
Milwaukee School of Engineering	61100.00

Lastly, we looked at which of these schools' alumni have the lowest median debt after graduation. In a world with exponentially increasing student debt levels, that can burden students and hamper success for decades after graduation. These underperforming schools do the best job at leaving their alumni with manageable amounts of debt, and as such are notable candidates for receiving scholarship money. The top schools in this category are there mostly because of their generous financial aid packages, so we've extended the list to 10 to display some schools that have low median debt due to a combination of price and financial aid factors.

Table 9: Five Best Underperforming Schools By Graduating Debt For Black Students

School Name	Median Post-Grad Debt
Duke University	7000.00
Harvard University	7536.00
Florida Institute of Technology	9375.00
Cornell University	11500.00
Vanderbilt University	12000.00
Embry-Riddle Aeronautical University-Daytona Beach	13000.00
Yale University	13206.00
San Diego State University	13997.00
The University of Texas at Dallas	14000.00
University of Arkansas	14250.00

Table 10: Five Best Underperforming Schools By Graduating Debt For Hispanic Students

School Name	Median Post-Grad Debt
Florida Institute of Technology	9375.00
Cornell University	11500.00
University of Wyoming	12000.00
University of Louisville	12571.00
Warren Wilson College	12750.00
University of Chicago	12955.00
Embry-Riddle Aeronautical University-Daytona Beach	13000.00
Embry-Riddle Aeronautical University-Prescott	13000.00
Lipscomb University	13196.00
Montana State University	13775.00

Investing in the students of these schools will result more often in young professionals less burdened by heavy student loan debt.

8 Conclusion

The lists above represent our best suggestions for schools whose minority students most need—and will be most helped by—financial assistance by your organization. These are schools that our model suggests should be graduating African-American and Hispanic students at higher rates than they are, and whether your organization decides the best course of action is increased scholarships, minority-specific programs, or other forms of intervention, these are the schools whose students could use the most assistance.

One interesting thing to note is that schools that traditionally are well known for high indicators of success for all students—well known schools that have

high expected future incomes— often do not do a good job of graduating their minority students. Schools like Duke University, Georgetown University, and Yale University appear on our lists as schools identified as underperforming our model. Looking more closely at some schools like Carnegie Mellon and Boston College, we find that black students graduate at a rate more than ten percentage points lower than the overall graduation rate. This goes to show that doing research on the institution is often valuable before money is invested.

In this day and age of increased racial divisiveness and worry in our most vulnerable communities, the work your organization does to benefit the next generation of educated minority citizens is more important than ever. We hope this report has helped narrow the scope of your search to find the schools whose students most need your help to achieve their dreams.