Report

Eranda Bregasi, Leanne Lee, Jamie Stankiewiz

12/2/16

1 Abstract

This project focuses on providing our client from the biotech industry with the most qualified schools who fulfill their criterias of recruitment. To determine the best scshools for our client, we need to set a bar on what determines a "top school". To achieve this, we will focus on the use of College ScoreCard dataset. To determine which schools contain the highest achieving biotech students, we have set our dependent variable to be earnings after graduation. We will then look at how our other independent variables influence the earnings.

2 Introduction

The purpose of this research is to find the schools that contain high achieving students in the fields of Science, Technology, Engineering and Mathematics (STEM). In order to determine the top schools in the country, we first have to set restrictions of certain categories that describe the quality of a school. These "qualities" encompass standardized test data such as the school's average SAT math and ACT scores as well as the school's admission rate. Higher average test scores imply that the students at that school are smarter than students of schools with lower average test scores. Schools with lower admission rate imply that the school is much harder to get into, opposed to a school with a higher admission rate. To achieve this, we will look at the data of the schools and partition the schools such that we only have the "top" schools in the country. After our data cleaning process, we focus particularly on these top schools that remain to figure out the percentage of students who received a 4 year degree in STEM fields. After this step, we can narrow our search to look for specific criterion of interest. For example, we can look to see if students who received pell-grant with high SAT math/ACT scores attend more selective schools, which have a low admission rate and higher graduation rate. This directly affects our dependent variable of earnings, since it leads to higher earnings. For the completion of this project and to have an effective and accurate comparison of our data, we chose to use three different kinds of regression models that we learned throughout the semester. We used multiple linear regression, ridge regression, and random forest to fit our data against different tiers of earnings. The reason

we chose these regressions are because of the collinearities. Our predictors include the SAT math and ACT scores, the admission, graduation rate after the 4 years of attendance in a higher institution and pell grant recipients. Finally our dependent variable consists of earnings 10 years after graduation in different tiers.

3 Dataset

We use the dataset *CollegeScore Card* from `https://collegescorecard.ed.gov/data/`. In *data-subsetting.R*, we subset 34 columns that we need for the project. We carefully examined each predictors and see which predictors fit our needs to increase the performance in analysis. In *data-cleaning.R*, there's a lot of null data and privacy suppressed data so we have to delete those rows and replace some with N/A. After cleaning the data we put them in different csv files. In the *pre-processing.R* we select schools with top SAT math and ACT scores that are at least one standard deviation above average, and then we turn those columns that we need into numerics from factors. We decided on 5 predictors and 1 dependent variable out of the 34 options. Finally, we export the clean dataset into a *topschool.csv*.

```
 [1] "X"                    "INSTNM"               "CITY"
 [4] "STABBR"               "ZIP"                  "ADM_RATE"
 [7] "SATMT75"              "ACTMT75"              "COMP_ORIG_YR4_RT"
[10] "PCIP11"               "PCIP14"               "PCIP15"
[13] "PCIP26"               "PCIP27"               "PCIP29"
[16] "PCIP40"               "PCIP41"               "C150_4"
[19] "MN_EARN_WNE_INC1_P10" "MN_EARN_WNE_INC2_P10" "MN_EARN_WNE_INC3_P10"
[22] "PCT90_EARN_WNE_P6"    "ICLEVEL"
```

We have decided it was useful to keep descriptive categories such as

```
[1] "X"      "INSTNM" "CITY"   "STABBR"
```

which give the institution name, city, state, and zip code respectively. In deciding which categories will give a college certain desirable characteristics for our client, we broke the factors into 3 parts:

1. Pre-Admission: Data that describe how hard it is to get into a certain college
2. In-school: Data to tell us how many students graduate from each school and department of interest
3. Post-graduation: Data that will describe how much students make after graduating

Among the quantitative variables, pre-admission data includes the school's 75th percentile of the SAT math and ACT scores:

```
[1] "SATMT75" "ACTMT75"
```

as well as the school's average admission rate:

```
[1] "ADM_RATE"
```

In-school data includes the school's graduation rate

```
[1] "COMP_ORIG_YR4_RT"
```

as well as the "percentage of degrees awarded in" 8 specific programs named,

```
[1] "PCIP11" "PCIP14" "PCIP15" "PCIP26" "PCIP27" "PCIP29" "PCIP40" "PCIP41"
```

which respectively stand for:

PCIP11 = computer and information sciences and support services
PCIP14 = engineering
PCIP15 = engineering technologies and engineering related fields
PCIP26 = biological and biomedical sciences
PCIP27 = mathematics and statistics
PCIP29 = military technologies and applied sciences
PCIP40 = physical sciences
PCIP41 = science technologies/technicians

These variables will help us to determine the quality of the school, and the size of each program. Post-graduation data includes the "mean earnings of students working and not enrolled 10 years after entry in" 3 different categorical earnings denoted:

```
[1] "MN_EARN_WNE_INC1_P10" "MN_EARN_WNE_INC2_P10" "MN_EARN_WNE_INC3_P10"
```

MN_EARN_WNE_INC1_P10 = "lowest income tercile $0-$30,000" - low tier
MN_EARN_WNE_INC2_P10 = "middle income tercile $30,001-$75,000" - middle tier
MN_EARN_WNE_INC3_P10 = "highest income tercile $75,001 " - top tier

These categories will be used in our regression-models to help us determine the top schools as well as the visualizations created.

Once we have properly subsetted our dataset, it was necessary to clean it in order to be properly used. In particular, all of the PCIP columns that contained the value 0 was replaced by NA. We then needed to take out only the schools that were a 4 year institution,

```
[1] "ICLEVEL"
```

in this dataset, this column is categorical:

1 = 4-year institution
2 = 2-year institution
3 = less than 2-year institution

For our analysis, the only schools our client should be interested in are top 4-year institutions. So we subsetted our data set appropriately.

4 Methodology

We used regression analyses in this project because we wanted to determine if average earnings after graduation is dependent on the school's level of difficulty. We picked 3 different regression analyses to find out which model fits our data the best. The 3 regressions of choice are: multiple linear regression, ridge regression and random forest. In our regressions, the independent variables are: SAT math & ACT scores (SATMT75 & ACTMT75), admission rate (ADM_RATE) and graduation rate (COMP_ORIG_YR4_RT). These factors will help us to predict our response variable, highest earnings 10 after graduation (MN_EARN_WNE_INC3_P10). We have also gone through to analyze the the same regressions with the response variable being the other 2 categories of middle and lowest earnings (MN_EARN_WNE_INC2_P10, MN_EARN_WNE_INC1_P10).

We used multiple linear regression to provide a basic analysis to look at how the predictor variables independently affect the response variable. While multiple linear regression is a good start, it is most likely that the predictor variables are not independent. For that case, ridge regression is a better model. This is because our predictor variables are collinear. After comparing the results between multiple linear regression and ridge regression, we use random forest to predict the future earnings of high achieving students. The random forest method is a good fit for our case because it is more robust to multi-collinearities.

4.1 Overview of Regression Models

Multiple Linear Regression
The goal of multiple linear regression is to see how our response variable changes when it is dependent on other variables. These variables are assumed to be invariant from each other so we can look at how the response variables changes when only one predictor variables changes at a time, keeping the other variables constant. The multiple linear regression model is given by:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p + \hat{\epsilon}$$

Where the betas are the effective increase in Y with a unit increase in X_i, holding the other variables fixed. The goals is to find the estimates of β_i by using $\hat{\beta}_i$ and use the ANOVA analysis to determine which variables are significant to predicting the response variable.

Ridge Regression
Ridge regression is similar to OLS with the coefficients estimated by minimizing a slightly different quantity. By minimizing RSS, we can find the coefficient estimates that fit the data well.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

However, the shrinkage penalty is that $\lambda\sum_{j=1}^{p}\beta_j^2$ is small when β are close to zero. Ridge regression will produce a different set of coefficient estimates for each value of λ .

Random Forest
Random forest is another regression method that constructs many "decision trees" on the training data, more of a machine learning bootstrap. Decision trees partition the data one variable at a time to decrease the residual sum of squares.

Random forest also outputs 2 additional measurements:

- Variable importance: this is essence, ranks the variables of greatest importance (done by looking at how much the error value changes when you only change the independent variable, keeping the other variables constant)
- Neighbor proximity: this is used to determine the "structure" of the data

$$\hat{y} = \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}(W_j(x_i, x')\right)y_i$$

Here, W is a weight function that takes the new point x' and weights it relative to a neighbor point, x_i. Altogether, these predict the new $\hat{y}$.

5 Analysis

Multiple Linear Regression

R_squared_summary
0.6029138
0.6168404
0.5795355

The chart above shows the r-squared when comparing our predictors against earning after graduation in three tiers. The r-squared in this class is similar for top and middle tiers. However, if we compared the r-squared with just top and bottom tiers, we can see that top tier of

```
> TopLMsum

[1] 0.6029138
```

is performing better than

```
> LowLMsum

[1] 0.5795355
```

MSE_summary
0.3939347
0.3801186
0.4171275

Again, this MSE chart also have values that are similar within different tiers. However, low tier is still performing worse than the top and middle tiers. Since low tier has the highest MSE of

```
> LowMSE

[1] 0.4171275
```

it has more errors from the data than the other tiers.

The reason that top tier and middle tier are very close to each other is because of the large range of the middle tier. The middle tier encorperates the students' earnings after graduation, this range is between $30,001-$75,000. Many STEM and non-STEM students can fall into this bracket because this is the average earning after graduation for most schools in the United States.

The following data are the coefficients of each earning bracket from the multuple linear regression:

```
> TopCoef

         (Intercept)              SATMT75              ACTMT75
       -4.646350e-16         2.517286e-01         3.705993e-01
            ADM_RATE PELL_COMP_ORIG_YR6_RT      COMP_ORIG_YR4_RT
       -1.391154e-01        -3.840951e-01         3.259766e-01

> MidCoef

         (Intercept)              SATMT75              ACTMT75
       -6.646505e-16         2.953424e-01         4.062549e-01
            ADM_RATE PELL_COMP_ORIG_YR6_RT      COMP_ORIG_YR4_RT
       -1.238959e-01        -2.978452e-01         2.041115e-01

> LowCoef
```

```
  (Intercept)              SATMT75              ACTMT75
-6.021958e-16         2.537037e-01         3.882197e-01
     ADM_RATE PELL_COMP_ORIG_YR6_RT      COMP_ORIG_YR4_RT
-1.214981e-01         -2.095447e-01         2.178538e-01
```

First, let look at the coefficient in the top tier. We can see that there is a positive relationship effect for SAT math, ACT and graduation rate against the top tier of earning. For every standard deviation of SAT math score increases, there is an increase in

```
> TopCoef['SATMT75']

  SATMT75
0.2517286
```

standard deviation increase in top tier earning. From the coefficients, we can also see that the admission rate and pell grant recipient have a negative relationship against top tier earnings. For every standard deviation (of admission rate) that increases, there is a decrease in top tier earnings. The negative relationship makes sense because for schools with a higher admission rate, we expect it to result in lower earnings. Since we are only looking at students who perform well in SAT math and ACT scores, these students are more likely be admitted into more selective schools, implying a lower admission rate. When we look at the negative relationship for the pell grant recipient, we also see a negative relationship. For every standard deviation increase in pell grant recipient, there is a decrease in earnings. When we compare the coefficient in the low tier, we can see that SAT math and ACT score still makes a positive effect on earning. However, there is a slight decrease in pell grant and graduation rate. It shows that there is a less increase in earnings for each standard deviation of increase in graduation rate.

Ridge Regression

	MSE
Top Tier Earnings	0.3974380
Middle Tier Earnings	0.3873009
Lower Tier Earnings	0.4223115

Since ridge regression does not have r-squared values, we can compare the MSE values from the ridge regression with the multiple linear regression. We can see that the MSE values from ridge regression are fairly close to the MSE values from multiple linear regression.

The following data are the coefficients of each tier earning generated from the ridge regression. The order of the tables are as follows: Top Tier Earnings,

Middle Tier Earnings, Lowest Tier Earnings

	x
Intercept	0.0000000
SATMT75	0.2530749
ACTMT75	0.3483071
ADM_RATE	-0.1614524
COMP_ORIG_YR4_RT	0.2500293
PELL_COMP_ORIG_YR6_RT	-0.3077931

	x
Intercept	0.00
SATMT75	0.29
ACTMT75	0.36
ADM_RATE	-0.16
COMP_ORIG_YR4_RT	0.14
PELL_COMP_ORIG_YR6_RT	-0.22

	x
Intercept	0.00
SATMT75	0.25
ACTMT75	0.33
ADM_RATE	-0.16
COMP_ORIG_YR4_RT	0.15
PELL_COMP_ORIG_YR6_RT	-0.13

When we compare the ridge coefficients with multiple linear regression coefficients, it shows a larger positive relationship between SAT math/ACT scores and earnings. For every standard deviation increase in SAT math/ACT scores, there is a higher increase in standard deviation of earnings. The ridge regression coefficient also shows a negative relationship between admission rate and pell grant recipient against high earnings.

Random Forest

After comparing the coefficients, r-squared and MSE values from multiple linear regression and ridge regression, we used random forest to predict earnings.

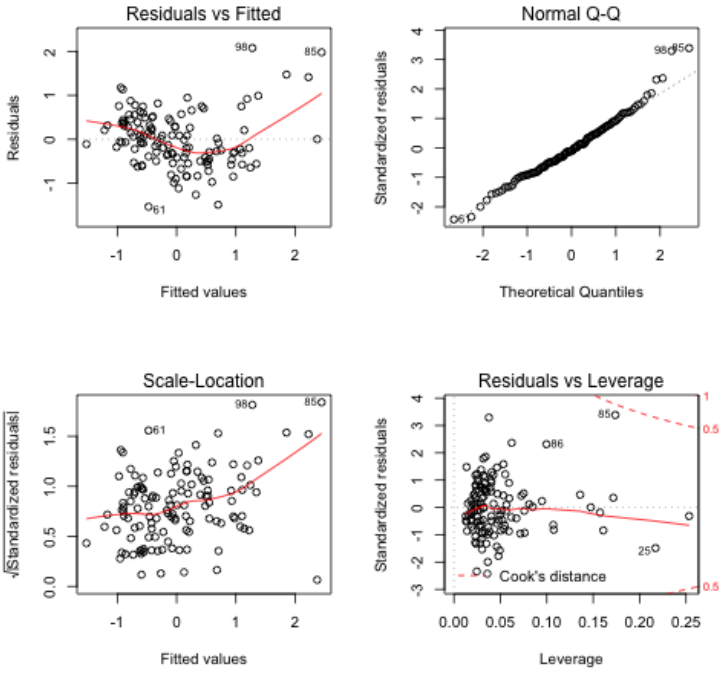
In random forest regression, we can look at the importance of the model. It consists of "IncNodePurity", which is the average cummulative reduction in node impurity due to splits by a variable over the trees. It is also the mean decrease in MSE. From the variance importance, we can see that the main positive relationships are from SAT math and ACT scores, concluding an existing positive correlation with mean earnings. From the random forest model, we can choose

	IncNodePurity
SATMT75	20.1190023
ACTMT75	19.6249217
ADM_RATE	23.9492907
COMP_ORIG_YR4_RT	13.4489085
PELL_COMP_ORIG_YR6_RT	11.8486366

the best features that represent our data. Also, the more selective schools that have a lower admission rate also play an important factor with mean earnings.

6 Results

Multiple Linear Regression

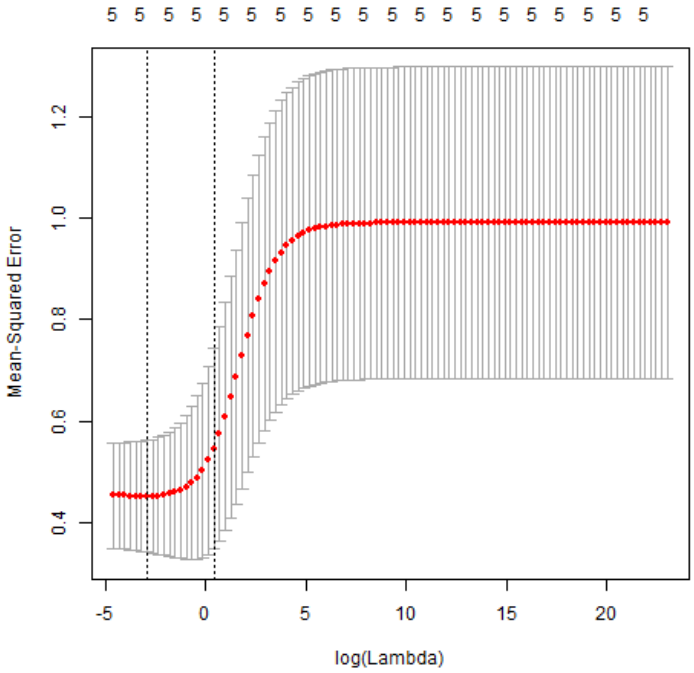


Ridge Regression

Random Forest

The plot shows that the error become more stable after 50 decision trees.

There is a positive relationship for high graduation rate, SAT math and ACT scores, which leads to an increase in mean earnings after graduation. There is



also a negative relationship for lower admission rate and pell grant recipients. Schools that are more selective accept less students; so from our findings, an increase in admission rate would harm the mean earnings. In addition, pell grant recipients usually come from below average family backgrounds, which lead to less earnings after graduation.

In this plot, we can see most top schools come from California.

Then, we determine the biggest programs from these top schools according to the major. Biology

Computer Science

Engineering

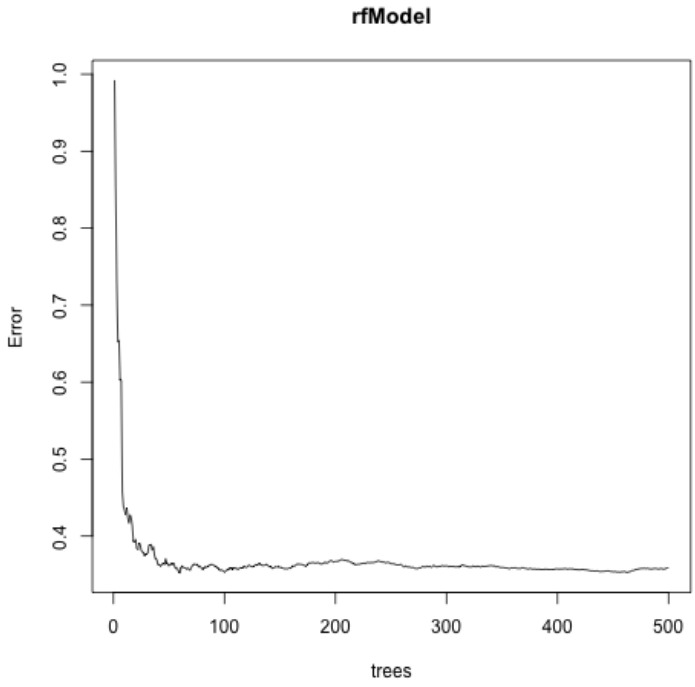
Math/Statistics

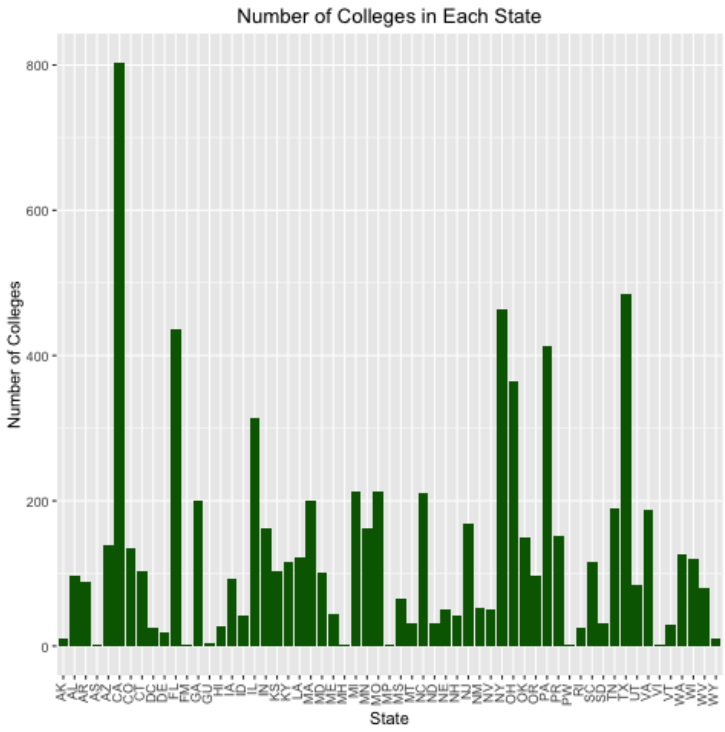
Technicians

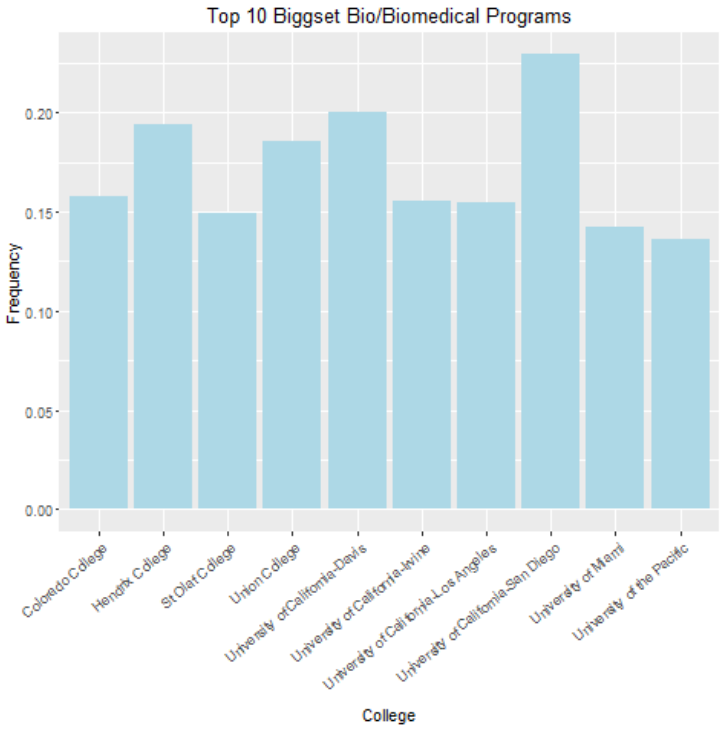
Physical Science

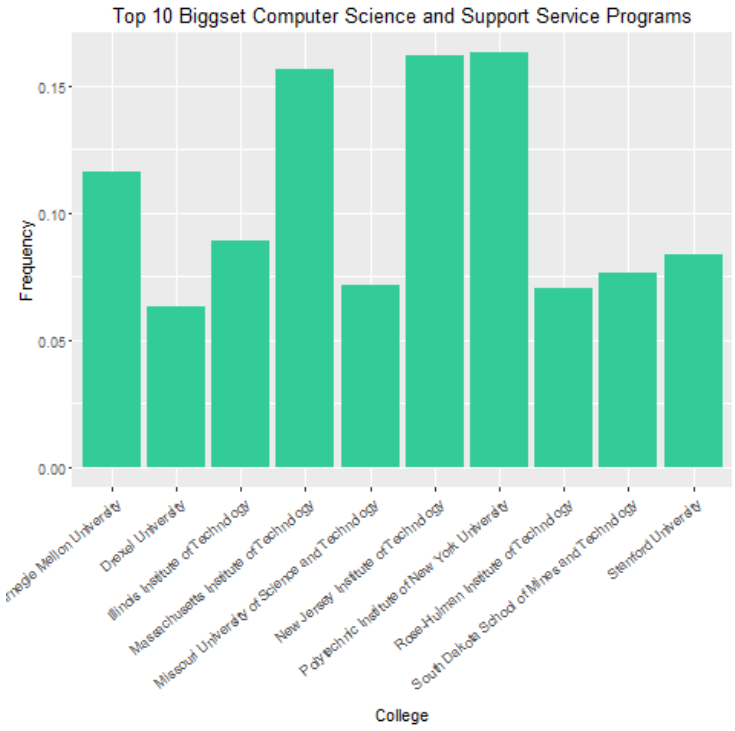
A tech company can easily find which school produces high achieving students with a large program according to the major.

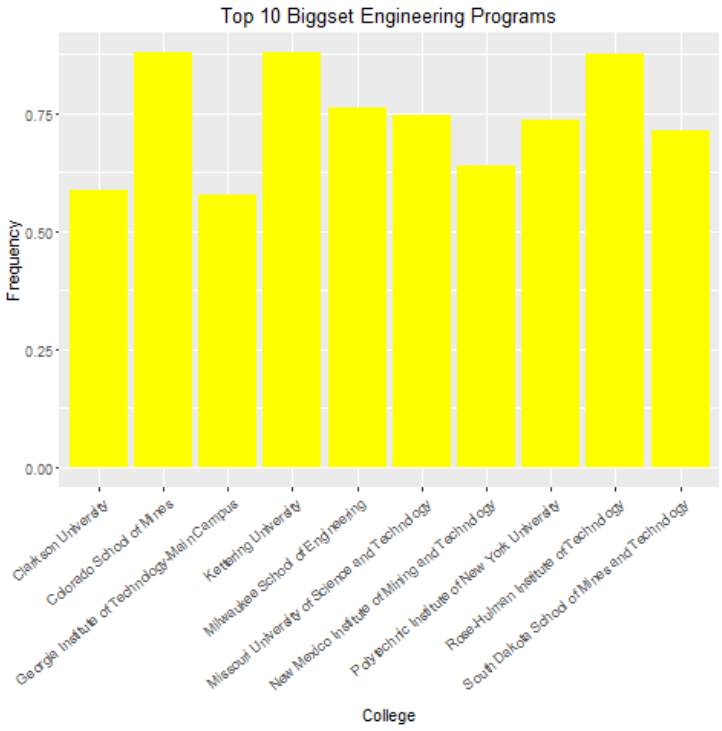
7 Conclusions

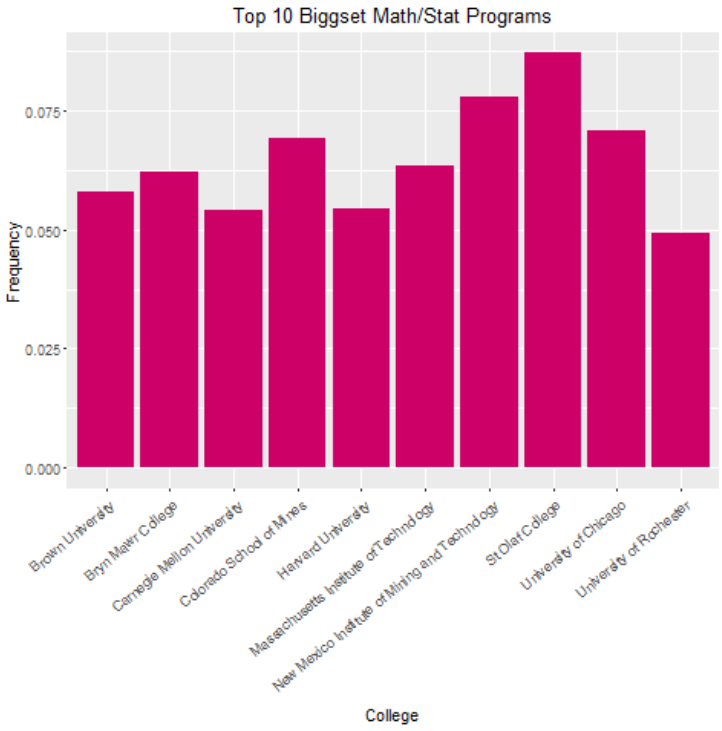

rfModel
Error
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
100
200
300
400
500
trees

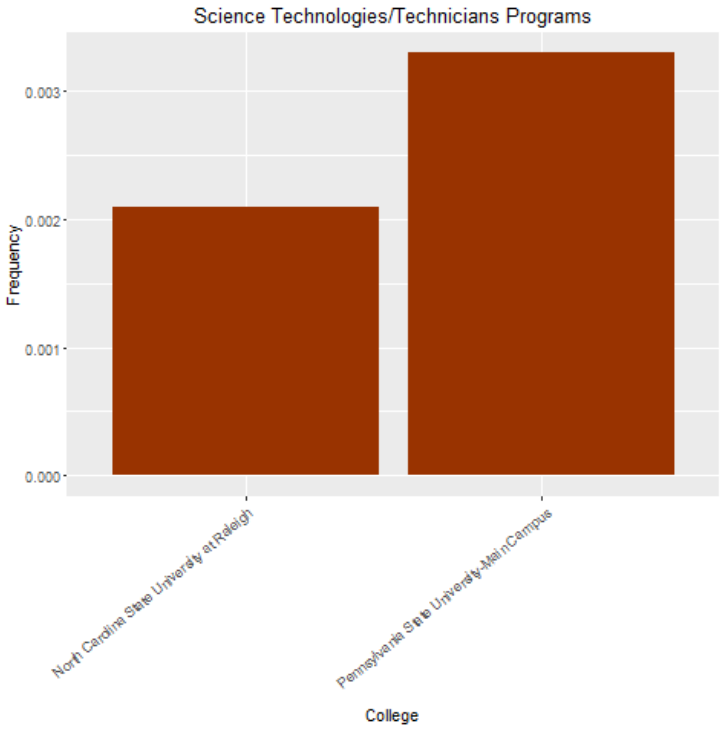

Number of Colleges in Each State
Number of Colleges
800
600
400
200
0
AK
AL
AR
AS
AZ
CA
CO
CT
DC
DE
FL
GA
GU
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MH
MI
MN
MO
MP
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
PR
PW
RI
SC
SD
TN
TX
UT
VA
VI
VT
WA
WI
WV
WY
State

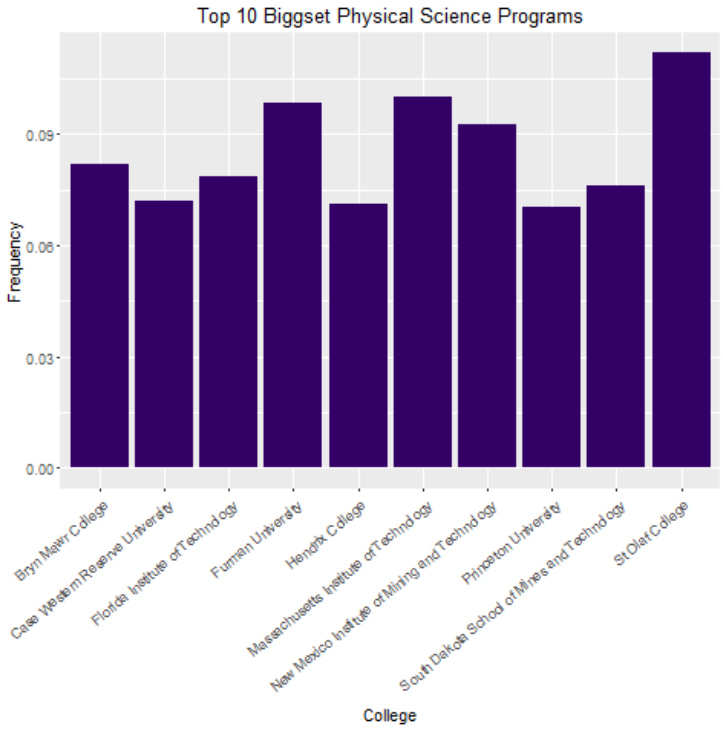

Top 10 Biggset Bio/Biomedical Programs
Frequency
0.20
0.15
0.10
0.05
0.00
Colorado College
Hendrix College
St Olaf College
Union College
University of California-Davis
University of California-Irvine
University of California-Los Angeles
University of California-San Diego
University of Miami
University of the Pacific
College


Top 10 Biggset Computer Science and Support Service Programs
Frequency
0.15
0.10
0.05
0.00
negie Mellon University
Drexel University
Illinois Institute of Technology
Massachusetts Institute of Technology
Missouri University of Science and Technology
New Jersey Institute of Technology
Polytechnic Institute of New York University
Rose-Hulman Institute of Technology
South Dakota School of Mines and Technology
Stanford University
College


Top 10 Biggset Engineering Programs
Frequency
0.00
0.25
0.50
0.75
Clarkson University
Colorado School of Mines
Georgia Institute of Technology-Main Campus
Kettering University
Milwaukee School of Engineering
Missouri University of Science and Technology
New Mexico Institute of Mining and Technology
Polytechnic Institute of New York University
Rose-Hulman Institute of Technology
South Dakota School of Mines and Technology
College


Top 10 Biggset Math/Stat Programs
Frequency
0.075
0.050
0.025
0.000
Brown University
Bryn Mawr College
Carnegie Mellon University
Colorado School of Mines
Harvard University
Massachusetts Institute of Technology
New Mexico Institute of Mining and Technology
St Olaf College
University of Chicago
University of Rochester
College


Science Technologies/Technicians Programs
0.003
0.002
0.001
0.000
Frequency
North Carolina State University at Raleigh
Pennsylvania State University-Main Campus
College


Top 10 Biggset Physical Science Programs
Frequency
0.09
0.06
0.03
0.00
Bryn Mawr College
Case Western Reserve University
Florida Institute of Technology
Furman University
Hendrix College
Massachusetts Institute of Technology
New Mexico Institute of Mining and Technology
Princeton University
South Dakota School of Mines and Technology
St Olaf College
College