

# Report

Eranda Bregasi, Leanne Lee, Jamie Stankiewicz

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 scschools 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] "INSTNM" "CITY"  "STABBR" "ZIP"
```

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 subsetting our dataset, it is necessary to clean it in order for it to be properly used for our regressions. For example, all of the PCIP columns that contained the value 0 was replaced by NA. We then need to take out only the schools that were a 4 year institution. This category is represented by:

[1] "ICLEVEL"

in this dataset, this column is categorical, containing 3 possible values:

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 have subsetting the data 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 other tiers of earnings (middle and lowest earnings 10 years after graduation, denoted MN\_EARN\_WNE\_INC2\_P10, MN\_EARN\_WNE\_INC1\_P10) using the same regressions.

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. In 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:

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

The betas are the effective increase in  $Y$  with a unit increase in  $X_i$ , holding the other variables fixed. The goal is to find the estimates of  $\beta_i$  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  $\beta$  are close to zero. Ridge regression will produce a different set of coefficient estimates for each value of  $\lambda$ .

### 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

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

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

The chart above shows the r-squared when comparing our predictors

R_squared_summary	
Top Tier	0.6029138
Middle Tier	0.6168404
Low Tier	0.5795355

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	
Top Tier	0.3939347
Middle Tier	0.3801186
Low Tier	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 encorporates 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 multiple 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 recipients 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 recipients, we also see a negative relationship. For every standard deviation increase in pell grant recipients, 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. This 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	0.0000000
SATMT75	0.2882189
ACTMT75	0.3612080
ADM_RATE	-0.1560997
COMP_ORIG_YR4_RT	0.1362545
PELL_COMP_ORIG_YR6_RT	-0.2166800

	x
Intercept	0.0000000
SATMT75	0.2503716
ACTMT75	0.3328487
ADM_RATE	-0.1557515
COMP_ORIG_YR4_RT	0.1477593
PELL_COMP_ORIG_YR6_RT	-0.1290091

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 recipients against high earnings.

## Random Forest



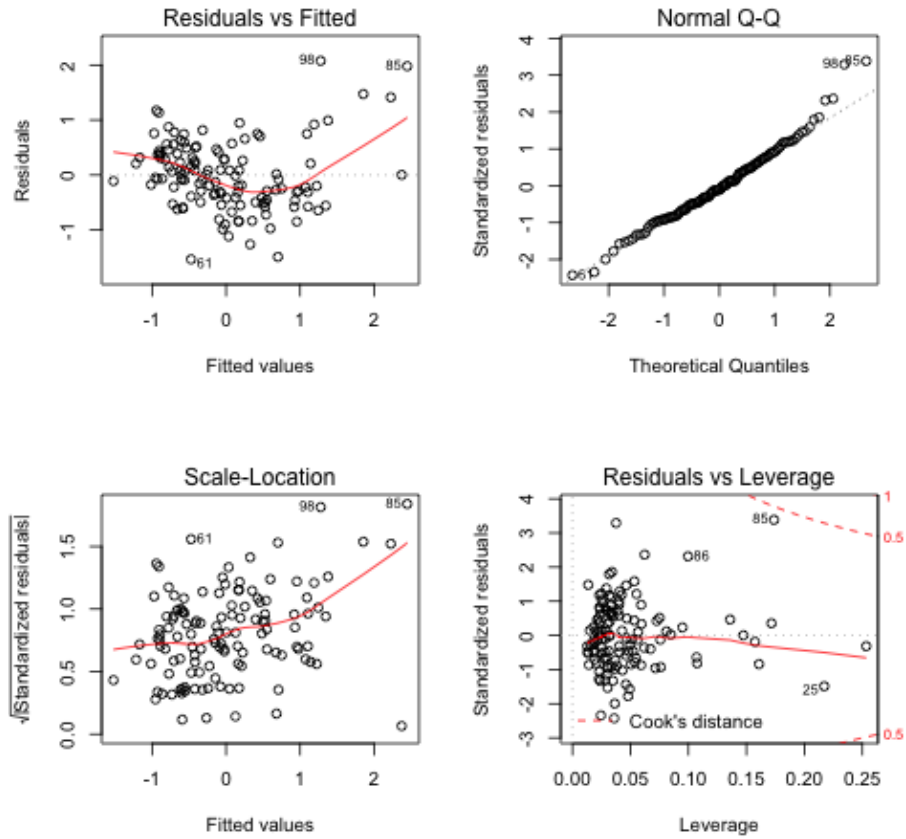
After comparing the coefficients, r-squared and MSE values from multiple linear regression and ridge regression, we used random forest to predict earnings.

	%IncMSE	IncNodePurity
SATMT75	15.0323194	22.4510535
ACTMT75	15.7735624	20.2170492
ADM_RATE	13.5671110	23.0732246
COMP_ORIG_YR4_RT	7.8260302	12.9750294
PELL_COMP_ORIG_YR6_RT	6.2190014	11.5542793

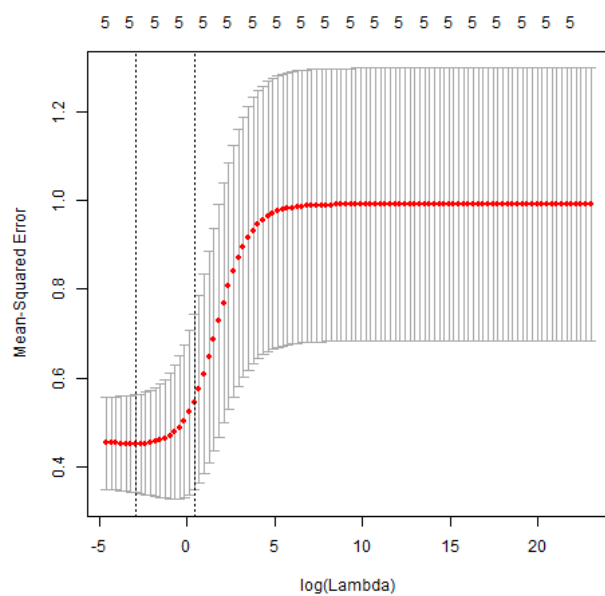
In random forest regression, we can look at the importance of the model. There are two types of importance measures shown above. The “IncMSE” is to show how bad the model perform without each variables. It consists of “IncNodePurity”, which is the average cummulative reduction in node impurity due to splits by a variable over the trees. It is also the mean decrease in MSE. From the variance importance, we can see that SAT math, ACT scores and admission rate are the key components of the model, concluding an existing positive correlation with mean earnings. From the random forest model, we can choose the best features that represent our data.

## 6 Results

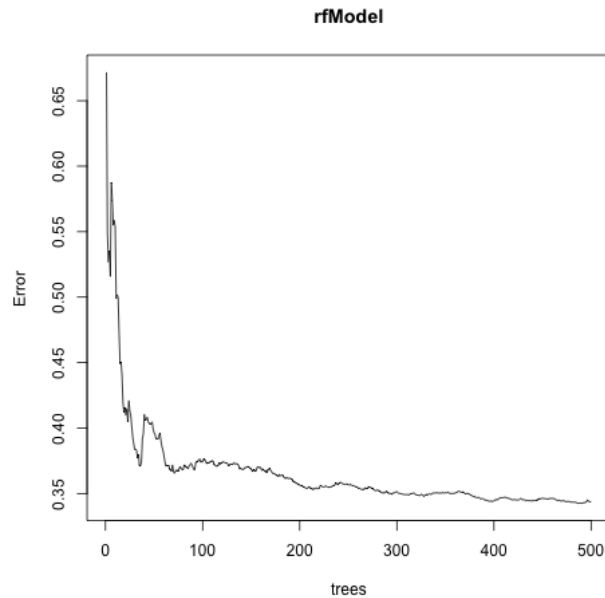
### Multiple Linear Regression



## Ridge Regression



## Random Forest

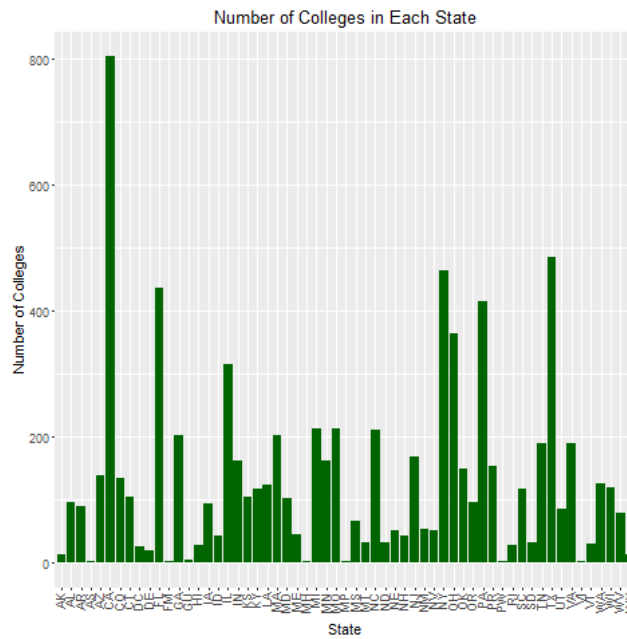


The plot shows that the error become more stable after 200 decision trees.

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.

Top 10 states that contain the most amount of colleges are:

CA	TX	NY	FL	PA	OH	IL	MO	MI	NC
804	485	464	437	414	364	314	213	212	211



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

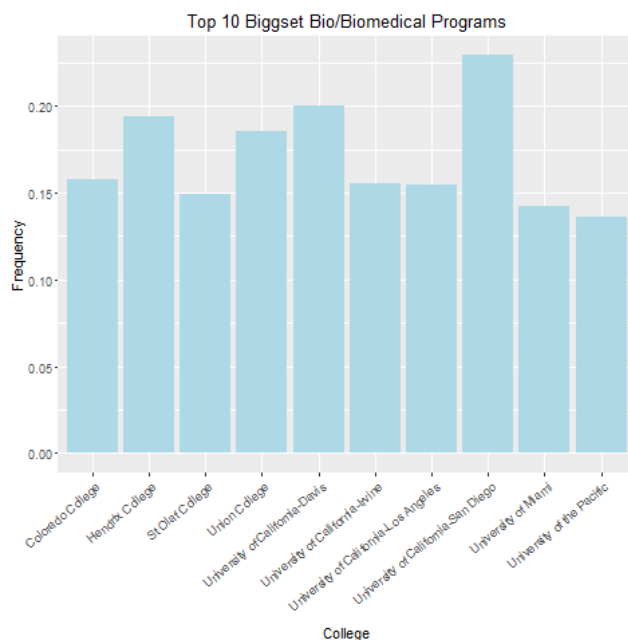
The following graphs are divided according to subject or major category. They depict the schools with the biggest respective programs. For example, the first graph depicts the “Top 10 Biggest Bio/Biomedical Programs”. In this chart, UC San Diego is the largest bio program in the country with a whopping

[1] 0.2294

portion of the student population graduate from this school with a bio/biomedical degree. Then, we determine the biggest programs from these top schools according to the major.

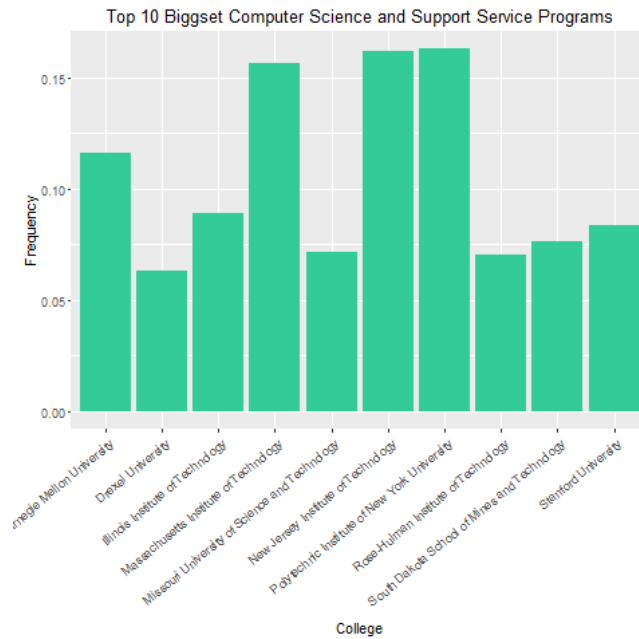
Biology

	INSTNM	PCIP26
1	University of California-San Diego	0.2294
2	University of California-Davis	0.2004
3	Hendrix College	0.1942
4	Union College	0.1855
5	Colorado College	0.1582
6	University of California-Irvine	0.1554
7	University of California-Los Angeles	0.1549
8	St Olaf College	0.1497
9	University of Miami	0.1420
10	University of the Pacific	0.1360



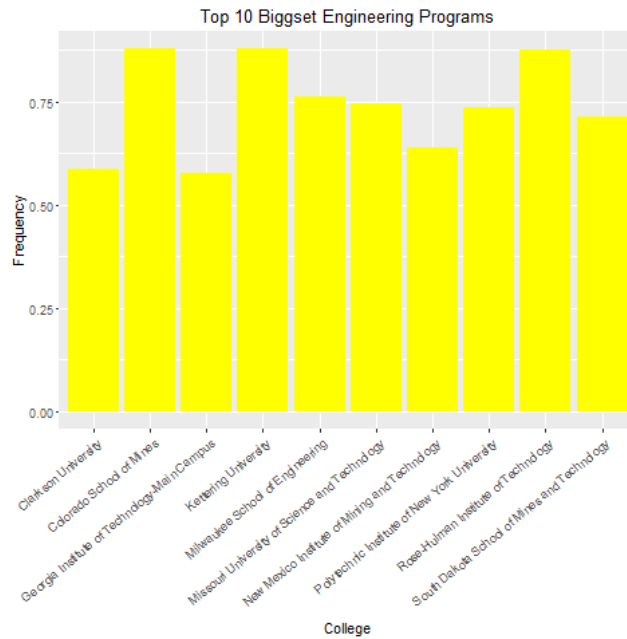
# Computer Science

	INSTNM	PCIP11
1	Polytechnic Institute of New York University	0.1630
2	New Jersey Institute of Technology	0.1619
3	Massachusetts Institute of Technology	0.1563
4	Carnegie Mellon University	0.1159
5	Illinois Institute of Technology	0.0888
6	Stanford University	0.0834
7	South Dakota School of Mines and Technology	0.0760
8	Missouri University of Science and Technology	0.0714
9	Rose-Hulman Institute of Technology	0.0700
10	Drexel University	0.0633



## Engineering

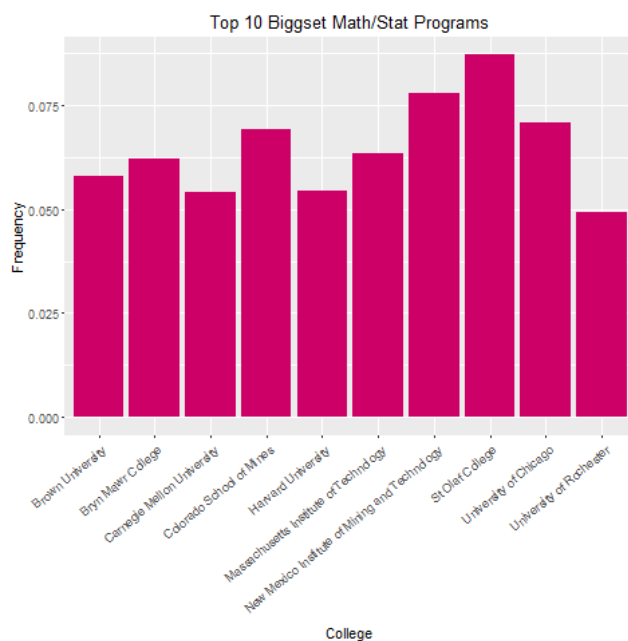
	INSTNM	PCIP14
1	Colorado School of Mines	0.8793
2	Kettering University	0.8783
3	Rose-Hulman Institute of Technology	0.8768
4	Milwaukee School of Engineering	0.7639
5	Missouri University of Science and Technology	0.7470
6	Polytechnic Institute of New York University	0.7355
7	South Dakota School of Mines and Technology	0.7148
8	New Mexico Institute of Mining and Technology	0.6390
9	Clarkson University	0.5886
10	Georgia Institute of Technology-Main Campus	0.5788





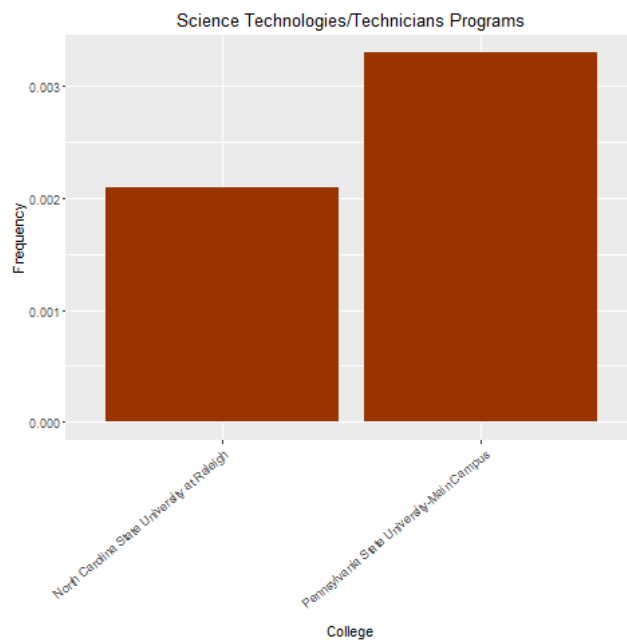
Mathematics/Statistics

	INSTNM	PCIP27
1	St Olaf College	0.0872
2	New Mexico Institute of Mining and Technology	0.0780
3	University of Chicago	0.0709
4	Colorado School of Mines	0.0693
5	Massachusetts Institute of Technology	0.0633
6	Bryn Mawr College	0.0621
7	Brown University	0.0580
8	Harvard University	0.0544
9	Carnegie Mellon University	0.0542
10	University of Rochester	0.0492



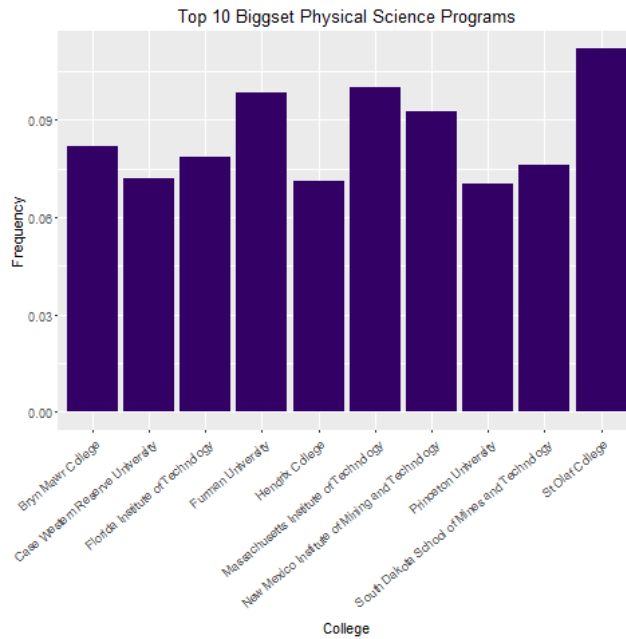
# Science Technologies/Technicians

	INSTNM	PCIP41
1	Pennsylvania State University-Main Campus	0.0033
2	North Carolina State University at Raleigh	0.0021



# Physical Science

	INSTNM	PCIP40
1	St Olaf College	0.1120
2	Massachusetts Institute of Technology	0.0999
3	Furman University	0.0984
4	New Mexico Institute of Mining and Technology	0.0927
5	Bryn Mawr College	0.0817
6	Florida Institute of Technology	0.0787
7	South Dakota School of Mines and Technology	0.0760
8	Case Western Reserve University	0.0718
9	Hendrix College	0.0712
10	Princeton University	0.0705



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

## 7 Conclusions

Our goal is to find high achieving students based on their ACT, SAT math scores, admission rate (how selective is their school), graduation rate and pell grant status. Then suggest schools to our client based on specific skills.

We have 2 regression models, MLR, ridge (better model for our data), random forest is used to find best features to predict earnings. In both regressions, the lowest tier of earnings had the highest MSE value. This concludes that they deviate most from the average.

From the analysis of our project, we found that multiple linear regression produces similar R squares and MSE values for all three tiers. However, the low tier is still appear a disadvantage from other tiers. Ridge regression produces similar R squares and MSE for all tiers. Ridge regression adds a penalty term to reduce overfitting, which produces better results than multiple linear regression.

The results from ridge regression can best be explained through our coefficients. In the coefficients of *both* multiple linear regression and ridge regression, we found that there are positive relationship between future earnings and SAT math, ACT scores and graduation rate. There exists a negative relationship between future earnings with admission rate and pell grant recipients.

Math SAT scores, ACT scores and graduation rate all had positive coefficients with respect to earnings. This is what we expected, that the test scores accurately correlate with earnings because these predict how difficult and competitive a school is. Admission rate is negatively relates to earnings, as we expected because the less schools they admit, the higher the earnings are. This is consistent with what we expected, that lower admission rate (more competitive schools) leads to higher earning students. Pell-grants come from a more disadvantaged family, lack of resources. People who don't receive pell grants have more resources in support, which relates to higher earnings in the future. This is also consistent with what we expected, that more privileged students are more likely to succeed in the future.

From random forest model, we can find out which are the best features to predict the future earnings. In the variance importance, it shows the positive relationships from SAT math and ACT scores with respect to future earnings.

The middle tier range is very large, so it's hard to conclude any significant and distinct results from our analyses. This is probably because the 30,000–75,000 range is very average for all schools across the board, so many STEM and non-STEM people could still fall in that range. Therefore we cannot conclude that the middle tier is associated with high earnings or low earnings. We also found that California has most of the largest top STEM programs, vaguely followed by Texas and New York. This is most likely due to the size of each state.

Our results provide a recommendation of schools to pick from in terms of specific skills. We recommend choosing schools from our top school data set for the smallest admission rate, and highest SAT scores, ACT scores and graduation

rate with respect to each skill.