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 applicant who fulfill their criterias of recruitment. Though at first we'd have to go through all individuals, the talent would only be restricted to women who are very high-achieving in STEM fields, while excluding the male gender. To determine the bar for the applicants' ranking, we would have to focus on the use of College ScoreCard dataset. Since our dependent variable would be earning, the rest would be independent variables. To achieve the final process of data cleaning, we need to filter out the top schools using SAT math and ACT scores, since they are standard elements of admission process. Then we would have to consider the relationship between admission rate and graduation rate, while taking into account the SAT math and ACT scores and observe how they affect the future earnings.

2 Introduction

The purpose of this research is to find high achieving students in the fields of Science, Technology, Engineering and Mathematics. To achieve this, we first look at the SAT math and ACT scores and consider only the ones with at least one standard deviation above, since we have to set a minimum score to categorize the high achieving students that the companies are looking for. After this step, it is easy to filter out schools in terms of their admission rate criteria. Based on our data analysis, we can easily see that 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. After our data cleaning process, we focus particularly on these top schools that remain after our filter and we have to figure out the percentage of students who received a 4 year degree in STEM fields, such as BioEngineering, Engineering, Mathematics, Statistics and others. 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 ridge regression, random forest and multiple linear regression to fit our data against

different tiers of earnings. The reason we chose these regressions because of the co-linearities and to effectively compare. 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, and finally our dependent variable which consist of earnings 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 math SAT and math 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: math SAT & 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

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

R_squared_summary
0.6029138
0.6168404
0.5795355

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

```
> 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 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 earning. For every standard deviation of admission rate increases, there is a decrease in top tier earning. The negative relationship make sense because schools higher admission rate will not result in higher earning. Since we are only looking at students who perform well in SAT math and ACT scores, they are more likely to admit into more selective schools, resulting in 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 earning. 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 less increase in earning for each standard deviation of increase in graduation rate.

Ridge Regression

Since there is no r-squared in ridge regression, we can compare the MSE from

MSE
0.3974380
0.3873009
0.4223115

ridge regression with multiple linear regression. We can see that the MSE in ridge regression are fairly close to the multiple linear regression.

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

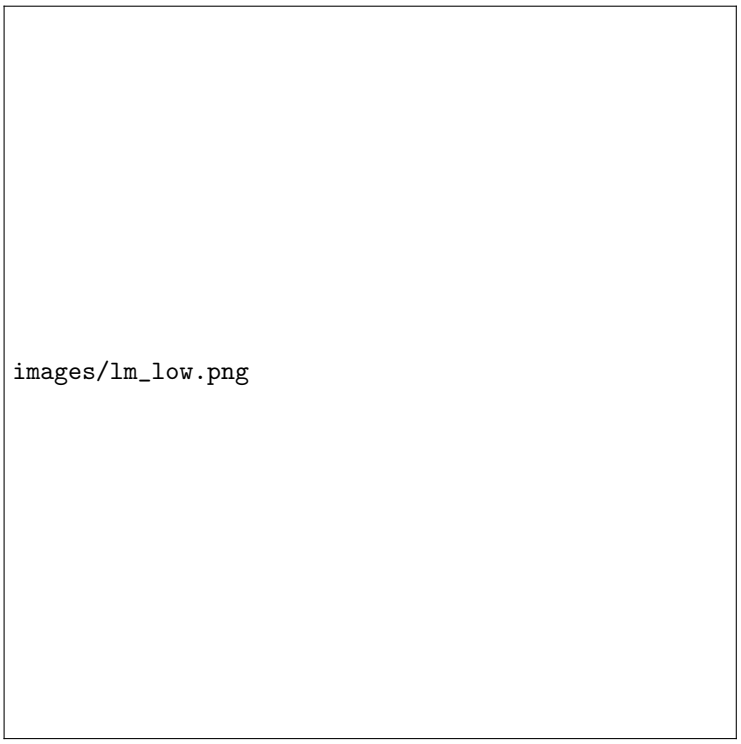
Random Forest

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

	IncNodePurity
SATMT75	20.1190023
ACTMT75	19.6249217
ADM_RATE	23.9492907
COMP_ORIG_YR4_RT	13.4489085
PELL_COMP_ORIG_YR6_RT	11.8486366

6 Results

Multiple Linear Regression



From the plots above, we can see that we could fit the multiple linear regression model in our case.

7 Conclusions