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

5 Analysis

6 Results

7 Conclusions