Project 3 Stat 159

Austin Carango, Youngshin Kim, Alexander Lee, Mandy Zhang

December 5, 2016

1 Abstract

The goal of this project is to create a predictive modeling process for the College Scorecard dataset from the perspective of a consultant whose client is a group of administrators trying to make their school more competetive. The data contains information pertaining to the "cost and value of institutions accross the country" and can be found at `https://collegescorecard.ed.gov/data/`.

2 Introduction

Given the College Scorecard data, which contains a variety of information including academics, demographics, cost, and earnings for thousands of colleges in America, how can a group of administrators increase the competitiveness of their school?

To answer this, we look at 4 different regression methods, including OLS, PLSR, LASSO and Random Tree. Each of these methods will be individually discussed and applied to the College Scorecard data. Then, each model will be compared and recommondations for the administrators will be given based off the results of the regressions.

We will attempt to model earnings of graduates as a function of 33 other variables, which will be discussed in detail in the next section.

3 Data

The dataset used in this project is called Most-Recent-Cohorts-All-Data-Elements.csv, and can be found at `https://collegescorecard.ed.gov/data/`. This particular subset contains only the most recent data, which we are interested in because we want to predict future outcomes. The raw data contains 7703 observations of 1743 variables, each observation being a college.

Before performing analysis of any kind this data was cleaned and scaled. This involved first picking out only private colleges from the data and selecting a subset of variables of interest. Then, rows and columns for which there were an abundance of NA values were deleted. Then, any remaining NA values

were imputed with column means. Finally, the data was mean centered and standardized. This resulted in the dataset we will analyze, final.csv.

This file contains 34 variables. The response variable is called Earning, renamed from MN_EARN_WNE_P10, and indicates the mean earnings of federally aided students 10 years after enrolling. We take this to be a good indicator of competitiveness, as high mean earnings after graduation are desirable. The 33 explanatory variables are as follows:

- HIGHDEG is the highest level of degree awarded by the institution, in descending order of graduate degree/certificate, bachelor's, associate's, and certificate.
- TUITFTE is the net tuition revenue per full-time student
- SATVR25/75, SATMT25/75, ACTEN25/75, and ACTMT25/75 are the 25th and 75th percentile scores for enrolled students in the English and math sections of the SAT and ACT.
- COSTT4_A is the average cost of attendance, tuition and fees
- NPT4_PRIV is the average net price of attending.
- UGDS is the number of undergraduates enrolled in the fall.
- UGDS_WHITE, UGDS_BLACK, UGDS_HISP, UGDS_ASIAN, UGDS_AIAN, UGDS_NHPI, UGDS_2MOR, UGDS_NRA, UGDS_UNKN, respectively indicate the number of white, black, Hispanic, Asian, American Indian/Alaska Native, Native Hawaiian/Pacific Islander, bi or multiracial, non-resident alien, and unknown race students.
- MARRIED, DEPENDENT and FIRST_GEN indicate the share of married, dependent and first generation students.
- PCTFLOAN indicates the share of students who recieved federal loans.
- PCTPELL indicates the share of students who recieved Pell Grants.
- C100_4 and C150_4 are the graduation rates for full-time, first-time students within 100 or 150 percent of expected completion time.
- GT_25K_P10 is the fraction of students earning over 25 thousand dollars per year 10 years after enrolling.

4 Methodology

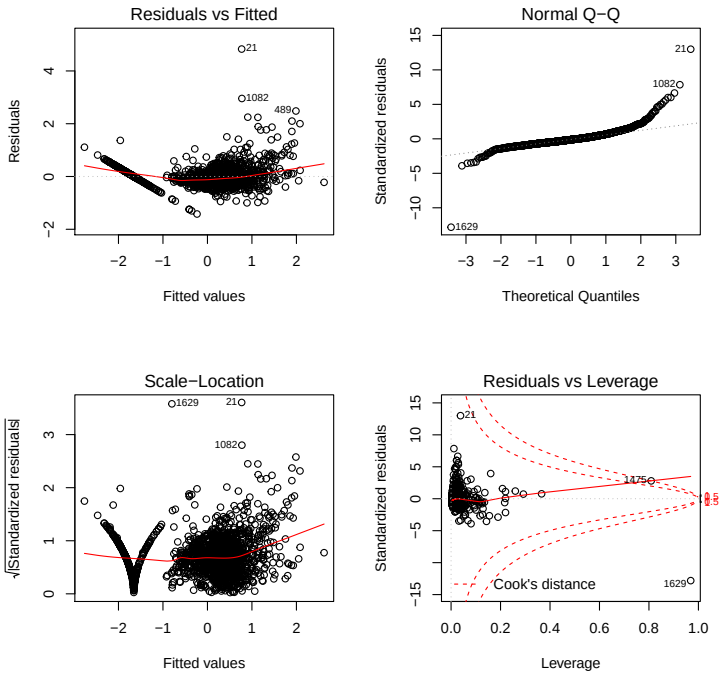
We analyze the data using four regression methods: original least squares (OLS), partial least squares (PLSR), lasso, and random forest. The OLS method simply creates a linear model based on our 34 predictive variables which minimizes the residual sum of squares. PLSR is a dimension reduction method, which is useful when variables are correlated with each other; since it is unknown if our varaibles are correlated, we run a PLSR to generate a model and compare it to the others. Lasso is a shrinkage method, aiming to find a model with the smallest lambda. The random forest method makes predictions using regression trees, which split the data into smaller and smaller subsets such that the residual sum of squares is reduced; the random forest model builds trees based on bootstrapped training samples, which decorrelates the trees, making them more reliable. This is done via the "caret" package, which simplifies regression training.

5 Analysis

In this section we show the processes used in each of the four regression methods by displaying the code used. Diagnostic/validation plots are also displayed for each regression.

5.1 OLS

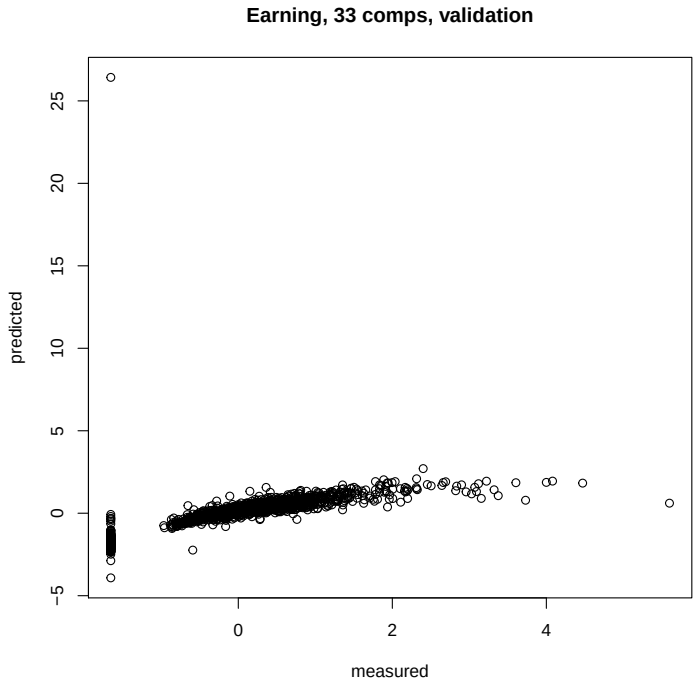
```
> set.seed(159)
> edu = read.csv('../data/final.csv', stringsAsFactors = FALSE)
> edu = edu[, -c(1)]
> ols = lm(Earning~., data = edu)
> ols_summary = summary(ols)
> save(ols, file = '../data/ols.RData')
```



5.2 PLSR

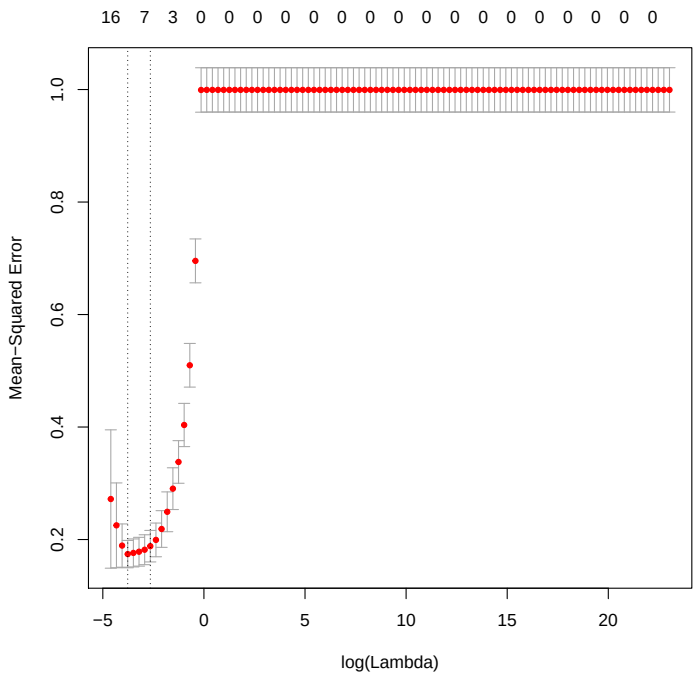
```
> library(pls)
> set.seed(159)
> plsr = plsr(Earning~., data = edu, validation ="CV", scale = FALSE,standardize = FALSE)
> plsr_coef = plsr$coefficients[,1,which.min(plsr$validation$PRESS)]
> plsr_coef
```

```
> save(plsr, file = '../data/plsr.RData')
>
```



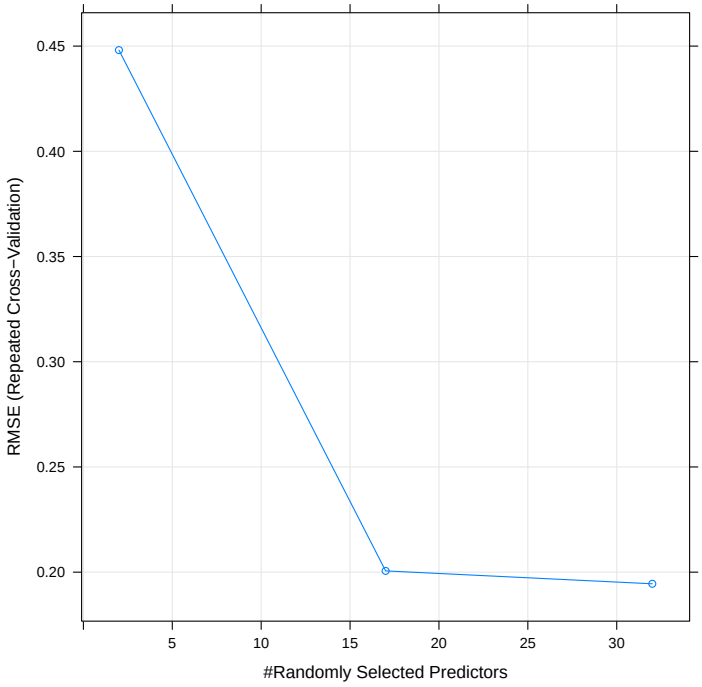
5.3 Lasso

```
> library(glmnet)
> edu = read.csv('../data/final.csv', stringsAsFactors = FALSE)
> edu = edu[, -c(1)]
> grid = 10^seq(10, -2, length=100)
> set.seed(159)
> lasso = cv.glmnet(as.matrix(edu[,c(1:33)]), as.matrix(edu[,34]), intercept = FALSE,
+                   standardize = FALSE, lambda = grid, alpha = 1)
> save(lasso, file = '../data/lasso.RData')
>
```



5.4 Random Forest

```
> library(caret)
> edu = read.csv('../data/final.csv', stringsAsFactors = FALSE)
> edu = edu[, -c(1)]
> train_index = sample(1:1636, 1336)
> train = edu[train_index,]
> test = edu[-train_index,]
> fitControl = trainControl(method = "repeatedcv", number = 10)
> model.rf = train(Earning ~ ., data = train, method = 'rf', trControl = fitControl)
> importance(model.rf$finalModel)
> save(model.rf, file = '../data/randomforest.RData')
>
```



6 Results

6.1 OLS

From these results we can see that the primary predictors of earning are SAT and ACT scores. Other significant variables are tuition, annual cost of attendance, the average net price of attendance, the number of undergraduates, dependence, the borrowing rate, and reception of a Pell grant.

6.2 PLSR

The coefficients of the PLSR model with the lowest lambda value. Of particular note are the number of Asian undergraduates (the highest value) and the number of students receiving a Pell grant (the lowest value).

6.3 lasso

The coefficients of the lasso regression model. The highest is the number of Asian undergraduates and the lowest is the number of students receiving a Pell grant.

	Estimate	Pr...t..
TUITFTE	0.02	0.01
SATVR25	0.55	0.00
SATVR75	-0.62	0.00
SATMT25	0.46	0.00
ACTEN25	0.34	0.00
ACTEN75	-0.49	0.00
ACTMT25	0.42	0.00
ACTMT75	-0.30	0.02
COSTT4_A	0.17	0.00
NPT4_PRIV	-0.09	0.00
UGDS	0.07	0.00
DEPENDENT	-0.03	0.02
PCTFLOAN	0.06	0.00
PCTPELL	-0.14	0.00
GT_25K_P10	0.81	0.00

Table 1: OLS Significant Coefficients

6.4 randomForest

The variables with a higher IncNodePurity value were the significant ones, which had more of an effect predicting the earnings of a student. These variables are tuition, race (being white or Asian), and receiving a loan in college.

7 Conclusion

The models above tend to agree that tuition, SAT/ACT scores, total cost of attendance, the racial composition of the undergraduate classes, taking out a loan, and receiving a Pell grant. However, these variables may not be useful in the context of making changes to the campus's administration to increase competitiveness. To begin with, the fact that tuition predicts eventual earnings is clear: schools that require higher tuition levels are typically better schools academically, and thus it would be clear that attending a school with a higher tuition would lead to finding a higher-paying job in the future. It is not the case that increasing tuition at your college would necessarily lead to an increase in competitiveness. There is a similar argument for the cost of attendance, as tuition is included in that cost, and for taking out a loan, as students at universities that have higher tuition would typically have to take out loans in order to pay the amount. Of interest is the negative correlation between earnings later and receiving a Pell grant in college: the models agree that having a Pell grant predicts lower earnings later in life.

	variable	coefficient
1	HIGHDEG	-0.00
2	TUITFTE	0.03
3	SATVR25	0.03
4	SATVR75	-0.01
5	SATMT25	0.06
6	SATMT75	0.01
7	ACTEN25	0.01
8	ACTEN75	-0.05
9	ACTMT25	0.02
10	ACTMT75	-0.03
11	COSTT4_A	0.09
12	NPT4_PRIV	-0.08
13	UGDS	0.09
14	UGDS_MEN	0.04
15	UGDS_WOMEN	-0.03
16	UGDS_WHITE	-0.01
17	UGDS_BLACK	-0.00
18	UGDS_HISP	-0.00
19	UGDS_ASIAN	0.14
20	UGDS_AIAN	-0.01
21	UGDS_NHPI	-0.00
22	UGDS_2MOR	0.00
23	UGDS_NRA	-0.02
24	UGDS_UNKN	-0.02
25	MARRIED	-0.04
26	DEPENDENT	-0.03
27	FIRST_GEN	0.05
28	PCTFLOAN	0.02
29	PCTPELL	-0.16
30	COSTT4_A.1	0.09
31	C100_4	-0.03
32	C150_4	-0.04
33	GT_25K_P10	0.80

Table 2: PLSR Coefficients

	variable	coefficient
1	TUITFTE	0.01
2	SATVR25	0.00
3	SATVR75	0.00
4	SATMT25	0.05
5	SATMT75	0.00
6	ACTEN25	0.00
7	ACTEN75	0.00
8	ACTMT25	0.00
9	ACTMT75	0.00
10	COSTT4_A	0.08
11	NPT4_PRIV	0.00
12	UGDS	0.07
13	UGDS_MEN	0.00
14	UGDS_WOMEN	0.00
15	UGDS_WHITE	0.00
16	UGDS_BLACK	0.00
17	UGDS_HISP	0.00
18	UGDS_ASIAN	0.12
19	UGDS_AIAN	0.00
20	UGDS_NHPI	0.00
21	UGDS_2MOR	0.00
22	UGDS_NRA	0.00
23	UGDS_UNKN	0.00
24	MARRIED	-0.01
25	DEPENDENT	-0.01
26	FIRST_GEN	0.00
27	PCTFLOAN	0.00
28	PCTPELL	-0.13
29	COSTT4_A.1	0.00
30	C100_4	0.00
31	C150_4	0.00
32	GT_25K_P10	0.76

Table 3: Lasso Coefficients

	variable	IncNodePurity
1	TUITFTE	32.97
4	SATMT25	9.60
15	UGDS_WHITE	5.59
18	UGDS_ASIAN	28.33
27	PCTFLOAN	9.84
32	GT_25K_P10	1176.88

Table 4: Relevant Variables from Random Forest