Final Project Report

Junyu Wang, Jie Sun, Mingtao Fang, Nichole Rethmeier

Dec 5th, 2016

1 Abstract

In this project we address a group of administrators trying to make their school more competitive, both in diversity and graduation rates. We use the College Scorecard data to apply a predictive model in order to determine factors that may influence greater diversity and graduation rates for schools within California. For our regression analyses, we refered to *An Introduction to Statistical Learning* by Gareth James et al. To learn more about the College Scorecard program which provided the data, refer to the U.S. Department of Education's at the College Scorecard Site.

2 Introduction

In this project we worked with a selection of the College Scorecard data, specifically using schools in California and variables related around diversity, graduation rates, and other important factors of the university such as degrees offer, debt, earnings, etc. We applied PLSR, PCR, Ridge, and Lasso regressions to our data, and ultimately drew conclusions based on the model that minimalized error. From there, we aim to make recommendations to our clients, a group of school administrators on how to improve their schools diversity and graduation rates based on this case study in California.

3 Data

Source The data used in this project was from the College Scorecard. The College Scorecard, introduced during the Obama administration, is maintained by the U.S. Department of Education. This program provides data around all colleges in the United States as a means of helping students compare different institutions to choose the most suitable for themself. The source of all of the data comes through federal reporting from colleges, such as the National Student Loan Data System and the Integrated Postsecondary Education Data System.

Content The content of this data set is vast, covering many important aspects around institutions. Some of the information included in the data set as a whole is:

- Type of degree offered
- Amount of financial aid offered
- Acceptance rates
- Post graduation earnings
- Racial diversity of student body

For our project we included all relevant information about diversity such as race, gender, marital status and completion rate, or the amount of students who successfully achieve their degree within 4 years. We limited our analysis to a case study of California, and included variables that were reported consistently across all institutions as we did not want inconsistent data to tarnish the analysis.

Exploratory Data Analysis

For our analysis, we used 29 variables from the College Scorecard data set and observed 394 colleges. To get an overall understanding of some of our data, refer to the summary statistics below.

	MEN	WOMEN	WHITE	BLACK	HISPANIC	ASIAN	AMERICAN.INDIAN	NATIVE.HAW
Median	0.421	0.579	0.251	0.0580	0.379	0.0661	0.00415	0.00510
First Quartile	0.261	0.510	0.162	0.0309	0.248	0.0347	0.00190	0.00222
Third Quartile	0.490	0.739	0.375	0.1181	0.516	0.1302	0.00830	0.01215
IQR	0.228	0.228	0.213	0.0872	0.268	0.0955	0.00640	0.00993
Mean	0.398	0.602	0.278	0.0912	0.398	0.0987	0.00634	0.00965
Std. Deviation	0.210	0.210	0.159	0.0940	0.196	0.0945	0.00758	0.01203

Table 1: Summary Statistics for Quantitative Variables

Aside from the quantitative variables that were used, there was also a handful of categorical variables within the dataset. The variables `HIGHDEG`,`PREDDEG`, `CONTROL`, and `ICLEVEL` were important categorical variables throughout our analysis. `HIGHDEG` represents the highest level of degree offered at an institution, where 0:4 stood for Non-degree granting, Certificate degree, Associate degree, Bachelor's degree, and Graduate degree, respectively. `PREDDEG` represents the predominant type of degree offered by the institution, following the same scale as `HIGHDEG`. `CONTROL` indicated the control of the institution, where 1:3 stood for Public, Private Nonprofit, and Private For-Profit, respectively. Lastly, we have `ICLEVEL` which represented the level of the institution having 1:3 indicate a school was either a 4-year, 2-year, or less than 2-year institution. To get an overview of the categorical variable used in this analysis, refer to the below frequency table.

	Highest Degree	Predominant Degree	Control	Institution Level	Frequency	Frequency Proportion
1	0	0	3	3	1	0.00254
2	1	1	1	2	1	0.00254
3	1	1	1	3	1	0.00254
4	1	1	2	2	1	0.00254
5	1	1	2	3	13	0.03299
6	1	1	3	2	16	0.04061
7	1	1	3	3	65	0.16497
8	2	1	1	2	22	0.05584
9	2	1	2	2	2	0.00508
10	2	1	3	1	1	0.00254
11	2	1	3	2	40	0.10152
12	2	2	1	2	74	0.18782
13	2	2	2	2	1	0.00254
14	2	2	3	1	2	0.00508
15	2	2	3	2	16	0.04061
16	3	1	2	1	1	0.00254
17	3	1	3	1	3	0.00761
18	3	2	3	1	24	0.06091
19	3	2	3	2	1	0.00254
20	3	3	2	1	3	0.00761
21	3	3	3	1	11	0.02792
22	4	1	2	1	1	0.00254
23	4	1	3	1	1	0.00254
24	4	2	2	1	2	0.00508
25	4	2	3	1	3	0.00761
26	4	3	1	1	30	0.07614
27	4	3	2	1	45	0.11421
28	4	3	3	1	13	0.03299

Table 2: Frequency and Frequency Proportion for Categorical Variables

4 Methods

Ordinary Least Square methods
In statistics, OLS is a method for estimating the coefficients in a linear regression model, with the goal of minimizing the sum of the squares of the differences between the observed responses in the given dataset and those predicted by the linear model. (i.e. $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j x_{ij}))^2$, where $\beta_0 ... \beta_p$ are the coefficients estimate)

1. **Ordinary Least Squares (OLS)**
   This is a common linear regression model, and we use it as a benchmark to evaluate the performance of the other 4 regression models.

Shrinkage methods
Shrinkage methods involves fitting a model with all P predictor. However, the estimated coefficients are shrunken towards 0 relative to the least squares estimates. Shrinkage has the effect of reducing variance. Since after the process of shrinkage, some of the coefficients might be exactly 0, it also performs variable selection.

1. **Ridge Regression (Ridge)**
   Ridge regression is very similar to least square, except that the coefficients are estimated by minimizing $RSS + \lambda \sum \beta_j^2$. This equation involves the minimization of two criteria, the first one is the same as OLS, ridge regersssion seeks to fit the data well by minimizing RSS. The second part of this equation, $\lambda \sum \beta_j^2$, is called the **shrinkage penalty**. And the shrinkage penalty gets smaller as coefficients get closer to 0. λ here is called the tuning parameter, and $\lambda \geq 0$. When $\lambda = 0$, Ridge regression is the same as OLS. As λ grows larger, the impact of shrinkage penalty grows bigger as well. The use of tuning parameter λ is also where Ridge regression outperforms OLS, and the mechanism's name is called **bias variance trade off**. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias
2. **Lasso Regression (Lasso)** Lasso regression is an alternative to Ridge and it overcomes oen of Ridge's disadvantage, that even though it uses all P predictors and even though it moves all coefficients close to 0, it won't set any of them to be exactly 0 unless λ is infinity and this disadvantage creates obstacles on model interpretation when P gets really large. Lasso regression minimizes $RSS + \lambda \sum |\beta_j|$. And it is the 2nd term that forces some of the coefficients estimates to be exactly 0.

Dimension Reduction methods
Dimension reduction works by **projecting** the P predictors onto a M-dimensional space, where M < P. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

1. **Principal Components Regression (PCR)**
   The **first principal component** direction of the data is that along which the observation vary the **most**. The PCR approach involves constructing the first M principal components, $Z_1, Z_2, ..., Z_M$, and then using those principle components as predictors in a linear regression model that uses least squares method. This approach reduces dimension because of the fact that usually a smaller amount ($M < P$) of principal components are sufficient to explain most of the variability in the data set, as well as the relationship with the response. PCR is an **unsupervised** approach because in the process of finding the first M principal components that best describes the P predictors, the response Y is not used to help determine the principal component direction.
2. **Partial Least Square Regression (PLSR)**
   PLSR is a **supervised** alternative to PCR. It first identifies a new set of features $Z_1, Z_2, ..., Z_M$ that are linear combinations of the original features, and then fits a linear model using least squares method with those M features. The difference is PLSR identifies $Z_1, Z_2, ...Z_M$ in a supervised way, which means it makes use of the response Y while identifying new features, therefore the new features not only approximate the old feature well enough, they are also related the response.

5 Analysis

Data Processing
Because of the massive size of our raw data set, a series of data cleaning tasks need to be performed in order to make the analysis feasible. And our data cleaning includes the following steps:

1. **Removing unrelated columns**
   The original data set has more than 1700 columns but we certainly don't need to use all of them. Since we are measuring competitiveness in terms of diversity and completion rate, we picked related columns like **UGDS_MEN, UGDS_WOMEN, UGDS_WHITE, C100_4** etc.
2. **Converting column types to numeric**
   The columns of data we read in are mostly of type character and factor, so we need to convert them to numeric values for further computation and regression.
3. **Dealing with NAs and non-computable values**
   There are a lot of missing values in columns like **C100_4** (completion rate) which are essential to our regression, so what we did is we first removed columns that has more than 75% NAs in it, and then we delete rows with NAs or with non-numeric data ("PrivacySuppressed" in our case). This left us with 394 rows in our data set and 24 columns. And since there are a lot of NAs in completion rate column, we chose to produce a separate data set for regression on that column, which ended up having about 105 rows in total.
4. **Calculating diversity score**
   We need a way to numerically measure diversity of each school, so we came up with our own diversity score formula. In order to define diversity in terms of **Gender, Race, Marital Status, and ratio of First Generation Student**, we calculated the **chi-square** ($x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$) of columns belonged those those 4 categories separately as their own diversity scores, then take the average as the final diversity score.
5. **Save to data files**
   In order to access the cleaned data more efficiently, we saved them to separate files from the raw data.

Training and Testing Data Sets
In order to benefit from running cross validation on several regression models, another step we took before building our model was to take a set of data for model building and another for testing model performance. The set of data used to build the model, our training set is 300 (out of 394) rows randomly selected from our cleaned data set for diversity regressions, and 80 rows (out of 105) from our data set for completion rate regressions, using the *sample*() function and the rest of rows are the test set. We also used the *set.seed*() function for reproducibility.

Model Building Process
For each of our regressions we fit the model to our training set after performing 10-fold-cross-validation and resampling the data. For the lasso and ridge regressions, we use the function *cv.glmnet*() to apply the model and perform the cross validation whereas with the PCR and PLSR regression models, we simply use the *pcr*() and *plsr*() functions and set the validation argument to "CV". After we have fit the functions to our data, we selected the best fit model looking for the minimum values of $lambda.min for the lasso and ridge regressions and of $validation$PRESS for the PLSR and PCR regressions. Once we selected our best models, best_comp_num in our PCR and PLSR regressions and min_lambda in our ridge and lasso regressions, we use our test set to compute the Mean Square Error to compare all of our best models. Finally to finish off our regressions we used the full set of data to get our actual coefficients for the models. The full data set was used in combination with our regression functions to compute summary statistics such as R^2, F-Statistic, and more.

6 Results

To understand the data, we computed descriptive statistics and summaries of all variables and generated corresponding plots in EDA phase, which can be found in `data/eda-output.txt` and `images`. Since we

are interested in studying the factors that affect diversity or graduation rates, we also obtained matrix of correlations. Please refer to eda-output.txt in the data section of this report, as the matrix is large.

We regressed diversity scores and graduation rates on different predictors separately. For diversity model, after fitting all models to the full data sets, we summarized coefficients and test MSE values. While the coefficients vary across models, we can spot some trends shared in common. The predictor that had the largest effect on diversity score was level of institution `ICLEVEL`, which conveys the highest level of award offered at the institution: 4-year, 2-year, or less-than-2-year. There are several other less influential elements that identify the degree profile of the institution including highest degree `HIGHDEG` and predominant undergraduate degree `PREDDEG`. Percent of undergraduates receiving federal Loans `PCTFLOAN` and governance structure `CONTROL` (public/private nonprofit/private for-profit) are also identified as related predictors by all models except Lasso. In contrast, the predictors `DEBT_MDN` and `MN_EARN_WNE_P10`, which stand for cumulative median debt and mean earnings, barely affect diversity of colleges.

	OLS	Ridge	Lasso	PCR	PLSR
CONTROL	-0.042	0.042	0.013	-0.028	-0.037
PCTFLOAN	0.039	-0.025	0.000	-0.013	0.018
HIGHDEG	-0.012	0.112	0.140	-0.010	-0.014
PREDDEG	-0.004	0.056	0.013	-0.000	-0.001
ICLEVEL	-0.079	0.174	0.198	-0.076	-0.081
DEBT_MDN	-0.000	-0.000	-0.000	-0.000	-0.000
MN_EARN_WNE_P10	-0.000	0.000	0.000	-0.000	-0.000

Table 3: Regression Coefficients for All Diversity Models

In order to see which regression is the best fit for our data set, we can take a closer look at mean square errors on test data sets to examine their quality. While all the models had relatively close MSEs, the two that were the lowest were the PCR and PLSR regressions. Therefore, it would appear these were the most suitable for predicting the diversity score.

	Ridge	Lasso	PCR	PLSR
Test MSE	0.0250	0.0245	0.0163	0.0166

Table 4: Test MSE Values for All Diversity Models

To analyze factors that affect graduation rates, we generated another model. Surprisingly, Lasso eliminated all variables except `MN_EARN_WNE_P10` though the coefficient estimate for `MN_EARN_WNE_P10` is pretty close to 0. And OLS generated an oddly large estimate for `UGDS_AIAN`, which might result from outliers. Among the rest predictors, `UGDS_MEN`, which reflects male-to-female ratio, had a relatively large coefficient, which makes it a prominent predictor. `FIRST_GEN`, which calculates the share of students in which neither parent completed college, also affects the graduation rates. Other influential predicts include those related to ethnic ratio (`UGDS_*`) or degree profile such as `PREDDEG` as explained in the previous model.

	OLS	Ridge	Lasso	PCR	PLSR
UGDS_MEN	-0.575	-0.095	0.000000	-0.431	-0.492
UGDS_WHITE	0.218	0.111	0.000000	0.207	0.260
UGDS_BLACK	-0.442	-0.067	0.000000	-0.044	-0.207
UGDS_HISP	-0.199	-0.046	0.000000	-0.255	-0.206
UGDS_ASIAN	0.132	0.021	0.000000	0.027	0.039
UGDS_AIAN	-10.762	-0.003	0.000000	0.000	-0.015
UGDS_NHPI	-2.598	-0.003	0.000000	0.003	-0.011
UGDS_2MOR	-0.729	0.007	0.000000	0.021	0.005
FIRST_GEN	-0.570	-0.067	0.000000	-0.218	-0.310
MARRIED	0.426	-0.023	0.000000	-0.127	-0.079
CONTROL	-0.077	0.017	0.000000	-0.018	-0.058
PCTFLOAN	0.387	0.065	0.000000	-0.115	0.172
HIGHDEG	-0.131	0.012	0.000000	-0.149	-0.179
PREDDEG	-0.227	-0.089	0.000000	-0.173	-0.173
DEBT_MDN	0.000	0.000	0.000000	0.000	0.000
MN_EARN_WNE_P10	0.000	0.000	0.000007	0.000	0.000

Table 5: Regression Coefficients for All Graduation Rate Models

In terms of model comparison, PCR and PLSR fit better on the data sets, which means dimention reduction methods generally outperform shrinkage methods on our data sets.

	Ridge	Lasso	PCR	PLSR
Test MSE	0.0471	0.0607	0.0387	0.0342

Table 6: Test MSE Values for All Graduation Rate Models

7 Conclusion

Exploratory Data Analysis

From our exploratory data analysis, we were able to gain some insights into our intial data. For example, the variables that impacted our diversity score the most were those used to compute our racial diversity score, `UGDS_WHITE`, `UGDS_BLACK`, `UGDS_HISP`, `UGDS_ASIAN`, `UGDS_AIAN`, and `UGDS_NHPI`. This is to be expected as the disparities between racial diversity at the university were larger than those of gender.

Another interesting correlation that was discovered during the exploratory data analysis was between `PCTFLOAN` and minority groups such as `UGDS_BLACK` and `UGDS_NHPI`. Looking at the Descriptive Statistics table in `data/eda-output.txt` we can see that on average the proportion of Black students at universities in California was less than 10% and the proportion of Native Hawaiian and Pacific Islanders at universities in California was less than 1%. Therefore we can infer either that more federal loans were provided to minority students, which can be explained by government programs that want to encourage more minority students to attend college, or that minority students on average accepeted more federal loans, which could be due to minorities not having the ability to fully pay for college themselves. For example, the American Psychological Association states that "unemployment rates for African Americans are typically double those of Caucasian Americans" which could offer one possible explanation for the correlations seen in our data set.

Regression Analyses

As our clients wished to compare themselves to other schools in relation to diversity and graduation rates, we ran two separate regression analyses upon the diversity score and graduation rates.

Diversity Regression For the regression analysis on the diversity scores, as you can see in the Results section, the PCR regression had the lowest MSE at 0.0163. Therefore, we will use this model to make some general conclusions about regression. From this we can see that the `ICLEVEL` variable had the largest impact upon our dependent variable `DIV_SCORE`. Referring back to the data section, `ICLEVEL` represented whether the institution was a 4-year, 2-year, or less that 2-year institution. As these factors are often indicators of other things about a university, it makes sense that they should play an important role in our regression

on diversity. For example, it is typical that a 4-year institution awarding Bachelors degrees would be more likely to attract more students to it than an institution that only was only a less than 2-year level. Especially as students from a young age are encouraged to attend 4-year colleges out of high school. The National Center for Education Statistics reports that in Fall of 2016 an estimated 13.3 million students attended 4-year colleges whereas only 7.2 million students attended 2-year colleges. As these 4-year institutions are attracting significantly more students than 2-year colleges, it makes sense that a larger student body would bring more diversity. Therefore, one recommendation for making your institution more competitive from a diversity stand point would be to offer 4-year programs. This will attract more students to the institution resulting in more diversity. We extend our recommendation to not just offering 4-year programs, but offering higher degrees as well, which goes hand-in-hand, as we saw that `HIGHDEG` was another predictor that had influenced diversity. This is similar in that students will be more attracted to attaining a higher degree, such as a Bachelors, as opposed to an Associates degree.

Graduation Regression

From our regression analysis on the Graduation rates, we can see from the Results that the PLSR regression had the lowest MSE at 0.0342. Therefore, we will use the coefficients from this model to make our recommendations to the client. One of the most interesting conclusions from this analysis is the correlation between `PCTFLOAN` and graduation rates. `PCTFLOAN` represents the share of student who receive federal loans. From this, we can infer that the greater the share of students that accepted federal loans to help pay for college, the more likely the student was to graduate. There are a multitude of reasons that could potentially explain this strong correlation. For example, students who accepted federal loans can more easily afford college. This leads students to pay for college, so that they may not have to drop out due to financial insecurity, and also so that they do not have to work to put themself through school. It seems more likely for a student to finish school if they do not have to worry about working as well. Another possible explanation for this correlation is that students who have accepted loans are highly motivated to finish as they know they'll be paying off loans when they graduate. These are just some potential reasonings. Therefore, to increase the institutions graduation rates, we recommend that the group of administrators offer education programs for students about:

- Types of federal loans offered
- How to determine if you need federal loans or not
- Plans for paying back federal loans after graduating

As a result, these classes could help break the stigmas around student loans as scary everlasting burdens. It would help students understand how they can afford to graduate and to focus on their schooling so they can be successful.