Final Project Report

December 4th, 2016

1 Introduction

In this project, our objective is to explore the differences between similar schools in terms of demographics and programs offered and how those differences contribute to school competitiveness based on the dataset available on College Scorecard. The client of our consulting project is a group of school administrators, and our data analysis is aimed to provide them with suggestions regarding what factors should they focus on and how to maximize their returns on investment in order to improve their school competitiveness. The requirement for this project can be found here.

2 Data

2.1 Original Dataset Source

The dataset we use is from College Scorecard :

https://collegescorecard.ed.gov/

College Scorecard is developed by the U.S. Department of Education to "key indicators about the cost and value of institutions across the country to help students choose a school that is well-suited to meet their needs, priced affordably, and is consistent with their educational and career goals".

The sources of data behind College Scorecard are available at:

https://collegescorecard.ed.gov/data/

We choose the most recent data, which can be downloaded here.

2.2 Data Cleaning

We first prepare the following datasets besides the original dataset to facilitate data cleaning process. The first is the states shown as abbreivations divided into four regional groups: West, MidWest, Northeast, and South based on location in United States. The second one contains the information of major cities in United States, including state, location, population density, etc. We get the major cities list from wikipedia.

We first calculate the sum of CIP, which includes the precentage of degrees awared in each field of study and whether the insitituion offers the program.

We then determine the way we assess a school's competitiveness. We select the following columns: Number of undergraduate student(UGDS), admission rate(ADM_RATE), average cost of attendence, tuition and fees(COSTT4_A), Mean and median earnings(MD_EARN_WNE_P10), completion rates for first-time, full-time students(C100_4) , percent of undergraduates receiving federal loans(PCTFLOAN).

We look at null values in each column we need and save the result to see what can we do with null values and we decide to remove all the rows which contains at least one null values.

We divide the schools by regions: west, midwest, northeat and south and divide by whehter the schools is in major cities or not.

As for minority, we basically calculate the sum of percentages of different minorty groups. By looking at the result summary, we divide the schools into 4 groups: less than 1st quartile , 1st quartile to median, median to 3rd quartile and larger than 3rd quartile.

We also calculate the number of students applied by dividing number of undergrads by admission rate.

We combine all the columns mentioned above and discard other columns that we do not need to construct the clean dataset for modeling.

One more thing about this dataset is that if some data is protected for privacy purpose, it's shown as PrivacySuppresed and we also eliminate columns which contains it.

2.3 Categorization

We categorize data into 8 groups. Data is first divided by region then further grouped by whether it's in major cities or not. Hence, we have the following 8 groups: West schools in Major city, West schools not in major city, Midwest schools in Major city, Midwest schools not in major city, North schools in Major city, North schools not in major city, South schools in Major city, South schools not in major city. We are going to fit models on each of the 8 categories.

2.4 Train Test Split & Method Determination

We have 1555 valid entries after data cleaning. We use 3:1 train test split ratio for the overall dataset and try to find the best regression method. Also, since we are taking log of STU_APPLIED, MD_EARN_WNE_P10, PCTFLOAN, C100_4, COSTT4_A to perform regression. We also replace the 0 values in those columns to 0.001 so that we could take log. The regression methods we try are introduced in Method section and further discussed in Analysis section.

3 Method

3.1 Least Squares Method

OLS estimators are considered as our base case in this project. OLS estimators are obtained by minimizing the sum of squares where are our coefficient estimates. This is the most common regression method, and it works the best when we have homoscedastic and uncorrelated errors within the data.

Shrinkage Method

Similar to the OLS method, shrinkage method also fits a model containing all independent variables but with a focus on constrains the coefficient estimates toward 0. By shrinking the coefficient estimates, the variance in estimators is significantly reduced compared to that of OLS.

Ridge Regression

Instead of minimizing RSS, Ridge regression minimizes . This regression abandons the requirement of an unbiased estimator in order to obtain a more precise prediction intervals. The addition term is a shrinkage penalty and since it's a sum of coefficient estimators, it will be small when s are all close to 0. Here, is called a tuning parameter and it controls how much RSS and shrinkage penalty affects the regression model. When , ridge regression is the same as the least squares models. Because each corresponds to a different set of coefficient estimates, we usually try a wide range of s and pick the best tuning parameter with the smallest MSE.

One of the main feature of ridge regression is the bias-variance trade-off. When , which is the least square model, the variance is high the estimators are unbiased. As lambda increases, variance decreases significantly but bias only sightly increase. Since along with the relative effect of the change, when is smaller than 10, the variance drops rapidly, with very little increase in bias, so MSE decreases.

Lasso Regression

The major difference between Lasso and Ridge is that Lasso method aims to minimize + . This is to compensate for the fact that ridge regression with always generate a model with all predictors since it doesn't involve a process of variable reduction. Therefore, lasso improves on this by allowing some of the coefficient estimates to be exactly 0 if the tunning parameter is sufficiently large. Therefore, this variable selection process makes the model much easier to interpret with a reduced amount of predictors.

3.2 Comparison between Ridge and Lasso

In general, lasso is expected to outperform ridge when we have a small amount of predictors having significant coefficients, and the rest close to 0. Ridge regression will perform between when the response is affected by many regressors with equal-sized coefficients.

3.3 Dimension Reduction Methods

Dimension reduction method involves a transformation of variables before we fit in the least squares model. Instead of estimating coefficients when we have p regressors, it transform the data to where represent linear combinations of the original predictors, so that we only need to estimate coefficients .

Principal Components Regression

In principal components regression, we first perform principal components analysis on the original data which constructs M principal components, . Then, we use these components as the predictors and fit the least squares model to obtain coefficient estimates. The key assumption we hold in this regression model is that the directions in which show the most variation are the directions that are associated with Y. If the assumption holds, then fitting a least squares model to will lead to better results than fitting a least squares model to , since most or all of the information in the data that relates to the response is contained in . With fewer predictors, we can also reduce the risk of overfitting. PCR performs the best when the first few principal components are sufficient to capture most of the variation in the predictors and their relationship with the response.

Partial Least Squares

In comparison to PCR regression, which uses unsupervised method to identify the principal components, partial least squares method takes the advantage of a supervised learning process. It uses the response variable Y to identify new vectors that are not only similar to existing regressors, but also select the ones that are related to the response. Therefore, as indicated in the textbook, the PLS approach attempts to find directions that help explain both the response and the predictors.

4 Analysis

4.1 Exploratory Data Analysis (EDA)

The first step of conducting analysis is to understand the data by conducting exploratory data analysis. To conduct the EDA, we obtained descriptive statistics and summaries of all variables. For the quantitative variables, we wrote a function called output_quantitative_stats() to get minimum, maximum, range, median, first and third quartiles, IQR, Mean and Sd of all the quantitative variables including UGDS for all races, UGDS, ADM_RATE, COSTT4_A, MD_EARN, C100_4, PCTFLOAN, CIP_SUM, MINORATIO and STU_APPLIED.

Similarly, we wrote a function called output_qualitative_stats() to generate a table with both the frequency and the relative frequency of the qualitative variables including WEST, MIDWEST, NORTHEAST, SOUTH, MAJOR_CITY, MINOQ1, MINOQ2, MINOQ3, and MINOQ4. To understand the data better, we also want to generate some plots to visualize the data. We wrote the functions histogram_generator() and boxplot_generator() to generate histograms and boxplots of the quantitative variables and condition_boxplot_generator() to generate conditional boxplots between STU_APPLIED and the qualitative variables. To study the association between STU_APPLIED and the rest of predictors, we also obtained the correlation matrix of all quantitative variables using function cor(), the scatterplot matrix using function pairs(), the ANOVA between STU_APPLIED and all the qualitative variables using function aov().

Then, we divide our dataset into 8 separate clusters according to region and its proximity to major cities. In order to develop strategies to improve competitiveness for each cluster, we first tabulate some key statistics for each cluster.

Table 1: Cluster Statistics

The table contains the number of institutions in each cluster, with Northeast region not located in major city cluster having the most institutions. And since we are interested in predicting students applied by variables such as earnings, graduation rate, minority ratio and percentage of students with loans, we look at the mean for this variable in each cluster first.

ANOVA is designed to test whether there are any statistically significant differences between the means of independent groups. Since our measure of school competitiveness is the number of students applied, we will test whether the means of students applied is different among clusters to gauge whether our clustering criteria makes sense.

Table 2: ANOVA Test Result

The test result shows that we have a p-value smaller than 0.01, which means that we can reject the null hypothesis that the means are the same across all 8 clusters.

4.2 Regression Analysis

Then, after the preliminary EDA, we start to run regression and explore the relationship between variables.

To start with, we run an OLS regression for all variables that we believe have an impact on the number of students applied.

The dependent variable is the number of students applied for the fall term. It is calculated by dividing people admitted during fall term by the admission rate.

We use 6 main variables as regressors in the regression:

1. Median Earning: Student's median earning 10 years after graduation
2. Completion rate: Percentage of students graduated within 4 years.
3. Percentage with Student Loans: Percentage of students with Student Loans.
4. Major City: Whether the institution is located near a major city or countryside. This dummy variable equals 1 if it is located in a major city, 0 if countryside.
5. Minority Ratio: The ratio of non-white students to the total population.
6. West, Midwest, East: Region dummy variables. We divided all schools into 4 regions and if an institution belongs to a certain region, the corresponding dummy variable will be 1, and others be 0. We drop one dummy variable Northeast in the multiple linear regression to avoid perfect collinearity.

5 Results

5.1 overall regression results

Table 3: OLS Regression Output for the Full Data Set

We noticed that the coefficients are large and each variable comes with different units. In order to make the regression result more interpretable, we take logs of all the quantitative data in the regression. Because we can take log of 0, we replace 0 with 0.0001 in our data set.

In addition, since our data set is large enough, we will first split the set into train set and test set in order to gauge our the performance of our estimated coefficients. We divide the data set according to 3:1 train and test ratio. The train set is used to build the model and test set to calculate the SE.

Table 4: OLS Regression Output After Taking Log

Based on the regression output, some findings match with our expectations:

1. For every 1% increase in median earning 10 years after graduation, we expect the number of students applied to increase by 2.9%, holding other variables constant.
2. Every 1% increase in percentage of students with student loans is associated with a 0.36 decrease in number of students applied, holding other variables constant.
3. If the 4-year completion rate goes up by 1%, the number of students applied is predicted to increase by 0.28%, holding other variables constant.
4. If cost of attendence increase by 1%, on average, we expect the number of students applied decrease by 1.33%, holding other variables constant.
5. If an institution is located near a major city, the number of students applied will be 0.16% higher than schools in countryside, holding other variables constant.
6. If the minority ratio increases by 1%, we predict the number of students applied will increase by -0.98%.

Those 6 coefficients are all very significant at 1% significance level with p-value close to 0.

Since OLS is the most common and versatile method, it is our first choice. However, in order to decide which method fits our data better, we also apply Ridge regression (RR), Lasso regression (LR), Principal Components regression (PCR) and Partial Least Squares regression (PLSR) method.

In order to improve our accuracy on predicting MSE, we use cross validation. We used sample() function to get a 3:1 train test split of our original data and for reproducibility purpose, we set.seed() before running the simulation.

For ridge and lasso regression method, we used cv.glmnet() in R package ”glmnet” to conduct the ten-fold cross-validation on the train set. We then used the best fitted lambda we found from the train set to build a model and calculate MSE from the test set in order to gauge our performance.

Similarly, for pcr and plsr regression method, we used function pcr() and plsr() in ”pls” package to perform the 10-fold cross-validation. We also use the best fitted m from the train set to build the model and obtain the MSE from the test set.

Table 5: Regression Coefficients for 5 Regression Methods

Table 6: MSE of 5 Regression Methods

We notice that all regression methods have similar MSE with lasso resulting the smallest. After interpreting the coefficients, we decide to use ols on the regression that we will run for each cluster for two main reasons:

1. lm() provides us with a p-value associated with each coefficient. This gives us information regarding whether the impact of a certain variable is significant which plays a vital role in determining our advice to schools in a certain cluster. In contrast, due to the way Ridge and Lasso regression are designed, although they give out better prediction, but the SE associated with each coefficient is unreliable, so we will lose this information if we go with those regression methods.
2. We notice that some of our regression variables are correlated. Therefore, ridge and lasso regression, aiming to reduce the dimension by eliminating unnecessary variables, can cause a problem. If both variables can affect the school competitiveness through the same channel, we don’t want the regression randomly drop one since they explain the same portion of change in Y. Instead, we want to make the decision on our own based on the budget required for each change or whether it is practical to execute for a certain institution.

Therefore, weighing all the pros and cons for each regression method, we decide to use OLS to run the regression for each cluster.

5.2 cluster regression results

Here is a summary of all the coefficients and our advice for institutions in different areas.

Table 7: Regression Coefficients WM Cluster

1. **For schools located in the West major cities**:

All the coefficients are significant except minority ratio. This can be largely explained by the fact that this cluster has the highest average minority ratio among all clusters, meaning those institutions already have a very diverse student population. Therefore, further improving this aspect won't have a significant impact on the school's competitiveness.

The most significant regressors is median earnings 10 years after graduation. Most cities on the West Coast are where well-paid technology companies and banking industry clutered. Therefore, there is a possibility that people decide to attend a university on the West Coast in order to get a job with high pay. So during the application process, it is reasonable if they pay extra attention to their future career development and the salary level of previous graduates serve as a plausible measure.

Table 8: Regression Coefficients WN Cluster

2. **For schools located on the West coast countryside:**

   This cluster is our smallest one, so the coefficient might not be as accurate as clusters with larger population. However, from the EDA stage, we notice that WN cluster has the lowest average percentage of students with loans. This perfectly explains the fact that a 1% increase in this percentage won't have a significant effect on school competitiveness because their current level of students with debt is very low.

   The most significant factor here is similar to that of group WM, which is median earnings. And this finding shows that there exists similarities between schools within the same region, possibly due to local culture and regional economic development.

Table 9: Regression Coefficients MM Cluster

Table 10: Regression Coefficients MN Cluster

3. **For schools located in the Midwest major cities and countryside:**

   Those two clusters generate very similar results with all the coefficients being significant, especially MN cluster. According to the EDA, MN is the region with the lowest average number of students applied. Therefore, it makes sense for all the varaibles to show statistical significance since they have a lot of room for improvement. The most significant coefficient in both clusters, median earning, can be interpreted as if we can increase the earnings after graduation by 1%, we predict the students applied will increase by 2.13% and 3.51% in schools located in the midwest cities and countryside, holding all other variables constant. In addition, they have relative low minority ratio among all clusters, therefore boosting their minority ratio by 1% is expected to bring an additional 1.82% and 1% in MM and MN cluster respectively, holding all other variables constant.

Table 11: Regression Coefficients NM Cluster

4. **For schools located in the northeast major cities:**

   All the coefficients are significant besides that of graduation rate. This cluster has the highest 4-year graduation rate, which explains why an additional boost in this rate won't bring as much increase in school competitiveness as other variables. Similarly to the West region, Northeast has prosperous economy and harbour millions of high-tech and finance related-companies. Therefore, it is not surprising to see that the effect of pay after graduation is has a t-value that is significantly higher than other factors, and we expect a 1% increase in graduation pay is associated with a 2.14% increase in number of students applied, holding all other variables constant.

Table 12: Regression Coefficients NN Cluster

5. **For schools located in the northeast countryside:**

   With a relatively low minority ratio to start with, improving minority ratio will bring the largest increase in school competitiveness. For each 1% increase in minority ratio, we expect the number of students applied will increase by 1.79%. Similarly, due to the regional effect we mentioned for the northeast area, we also have a large value of coefficient for median earning. Each 1% increase in median earning is expected to add 1.8% in number of students applied, holding all other variables constant.

Table 13: Regression Coefficients SM Cluster

Table 14: Regression Coefficients SN Cluster

6. **For schools located in the south major cities and countryside:**

   SM and SN have similar statistics in terms of graduation rate and the number of students with student loans.

   This cluster has the lowest average of number of students applied and median earnings among all clusters with the second highest coast of attendence. Therefore, all the coefficients are significant with coefficients of median earnings and cost of attendence. If the school can implement programs to improve the earnings of their graduates by 1%, we expect there will be 2.45% more students applied, holding all other variables constant. If the school will be able to reduce the cost of living by 1%, they can expect to receive 1.61% more applications each year.

6 Conclusion

Those regression results and intepretations give us great insights in developing strategies to boost the number of students applied, therefore improving school competitivenss.

Combining all the findings we had from analysis section, we conclude 3 main sugguestions:

6.1 Overall regression results

Based on our overall sample of 1555 institutions, median earnings, completion rate in 4 years, its proximity to major city and minority ratio all have a positive impact on the number of students applied, while percentage of students with loans and cost of attendence negatively correlated to number of applications received.

6.2 West and Northeast Region

West coast and northeast region are where well-paid technology and finance firms concentrated and historically have diverse population. Although their average earnings are already higher than other regions, our regression shows that students still place higher weight on median earnings when deciding which school to apply. If those schools want to improve their competitiveness, we believe investing in career development programs is the most effective measure. Concrete steps include: establishing long-term relationships with companies, building an extensive network with alumni and expanding programs that cater to market demand such as admitting more students to computer science, busines and engineering majors.

6.3 Midwest and South Region

Schools located at midwest and south region have lower number of students applied, so they have greater space for improvement. In addition, midwest has the lowest minority ratio and south region has the lowest completion rate. So besides implementing better career development programs, which is still a significant factor in the regression, midwest schools should also focus on improving the diversity and institutions located in the south should work on improving completion rate in order to attract more students. Concrete steps to promote diversity includes: improving their student loan system to make sure that minority students have access to the financial resources they need, set certain criteria in the admisssion process with a mindset of promoting diversity and encouraging minority student groups on campus. Concrete steps to improve completion rate includes: personalizing students' experience by reducing class size, introducing freshman interest groups so people can find support from their peers within a small group setting and expanding their academic advising program to address people's academic concerns.