Evaluation of Affordability of Postsecondary Education and Compensation Outcomes

CSC869 Final Project Report, Spring 2017

Paul Previde

I. Problem Definition

Every year, tens of millions of Americans must decide on whether, and where, to invest in postsecondary education [1, at 2]. In today's economy, postsecondary education can be essential to provide a student with economic opportunities later in life [2]. For example, of the thirty fastest-growing occupations, over half require postsecondary education [2]. However, not all universities and colleges serve students equally well: there is great variation among schools in terms of how well they provide cost-effective preparation of students for future careers. Therefore, the decisions that students make regarding their postsecondary education is both important and difficult.

The main problem that I want to address in this project is to explore how graduating from a cost-effective educational institution affects a person's earning power relative to graduating from a more expensive university. I have long been curious whether students achieve the same or better outcomes by attending a cost-conscious choice like SFSU, as contrasted with incurring more debt to attend a more expensive school. My main focus is on the institutions rather than on student-specific criteria, such as whether students with certain attributes have more lucrative careers.

II. Dataset

A. Overview

The College Scorecard dataset is the product of a collaboration between the United States Department of Education and the Department of the Treasury [1, at 23]. Under the Higher Education Act of 1965, all institutions that participate in federal student aid programs provide the Department of Education with a broad spectrum of academic performance information, including tuition and cost of attendance, enrollment of Pell grant recipients, demographic and financial information about aid-receiving students, and many other categories [1, at 19]. The aforementioned data on federally aided students was then linked to earnings data from tax records [1, at 23] and used to produce aggregated, anonymized estimates of institution-level statistics describing the earnings of the university graduates.

The College Scorecard dataset is a comma-separated values file available for download from the Kaggle web site [3]. The entire download is approximately 605 MB, and in decompressed form, the data file is over 1.3 GB in size. Each row in the dataset represents a particular institution in a particular calendar year, ranging from 1996 to 2013 (Table I).

Number of rows	124,699
Number of variables per row	1,731
Years covered	1996-2013
Number of missing values	92,185,358 (42.7% of dataset)
Mean and standard deviation of missing values per row of dataset	739 ± 421 out of 1,731

Table 1: High-level description of the College Scorecard data set.

Each column in the dataset represents a variable describing an aspect of the school or its class, as a whole, for a particular year. Earnings figures refer to the *aggregate* of the students (for whom information is known) from that university in that year. As such, the table does not contain student-specific information.

It is important to understand that for student privacy reasons, the dataset does not present student-level information; rather, student-level information has been used to generate aggregated statistics for each university in a given year. Thus, the dataset can describe institutions, not individuals.

B. Limitations of the College Scorecard Dataset

The dataset has a large number of missing values. Variables concerning incomes are especially prone to missing values, and there is data only for relatively recent years (2005 or later). The main reasons for the missing values are the fact that some of the data has been suppressed for student privacy reasons, and also that many institutions were exempt from providing information about their students or provided it in a form that prevented the information from being included on this dataset [1, at 51].

Other limitations exist on this dataset. The dataset only includes data for first-time, full-time students who participate in federal aid programs. As such, part-time students, transfer students, adults going back to school, and students who do not use federal aid are not captured in the data set [1, at 24]. The federal aid requirement eliminates approximately 30% of all students from the dataset [1, at 24]. Earnings data is provided only for non-enrolled workers [1, at 31].

Most significantly, the data represent aggregations of all students for which data is available at each university at each year. The dataset does not provide student-level information, so it is not possible to categorize individual students by their academic qualifications, performance in college, or their earnings after college.

III. Main Strategies

The first strategy used in this project was to perform a review of the extensive documentation about this dataset [1, 2] and literature research regarding how it has been cited. These steps revealed useful insights about the limitations of the dataset, as described in the previous section. The documentation review also provided guidance on which variables of the dataset were initial candidates for analysis by regression and classification. The candidate variables related to topics including: post-graduation median earnings from a university's alumni; the university's size and proportions of majors from a sample of areas (*i.e.*, business, computer science, life sciences, mathematics, and literature); the costs associated with attendance of the university; and the quality of the students, as measured by the average entering test scores and the admission rate of the university (with the premise that high average scores and low admission rates are the signs of a more competitive university).

Another strategy was to "get to know the data" using Tableau visualizations and R scripts to explore the the dataset and quantify missing values within each variable. A pruned version of the entire dataset was then prepared in R that eliminated a large number of variables that were not of interest to this study.

Based on the literature review and this step, the dependent variable of interest was determined to be the median earnings of graduates for a specific university, and the independent variables included: the student body size; the tuition in-state and out-of-state; the net price for students to attend, after financial aid; admission rates and average SAT math and verbal scores; and the proportion of students in various majors.

For the independent variables, the following strategies were used to handle missing values, and the results were compared to one another:

1. Do not impute missing values, simply omit any row with missing values.

2. Impute missing values using the mean of that variable across the data set.
3. Impute missing values using the mean of that variable across all observations within the same class (after discretization of median earnings, as described below).

Another strategy regarding the discretization of the dependent variable, median earnings. In order to be able to handle missing values and to be able to perform classification techniques, the variable for median earnings from each university's graduates was subject to two different discretization approaches:

1. Split the variable into two balanced classes using the median of the variable.
2. Split the variable into four balanced classes using the first and third quartiles and the median as the bin boundaries.

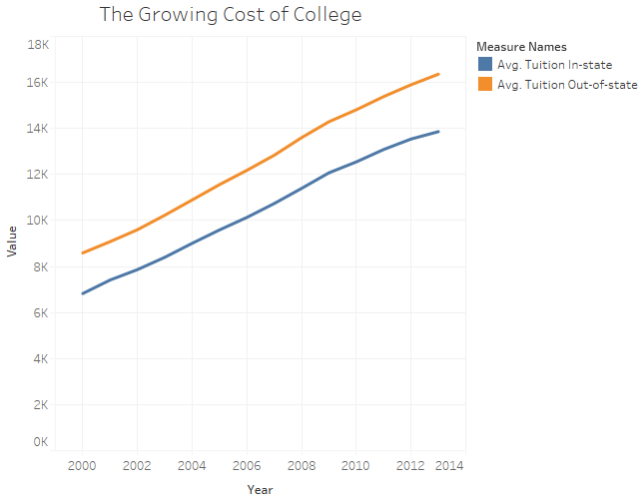
The next strategy involved choosing which, and how many, analytical techniques to use on the preprocessed data. Model interpretability was the key motivation of each choice – ideally, each technique would reveal evidence about the underlying relationship between median earnings and the other variables. To that end, linear, ridge, lasso, elastic net, and polynomial regression were chosen to explore the relationships, if any, between the independent and dependent variables, and also to explore the fits of models with a variety of complexities using the adjusted R^2 metric. Moreover, the test mean squared error of each technique was estimated using ten-fold cross validation in all models to see if more complicated models produced better results.

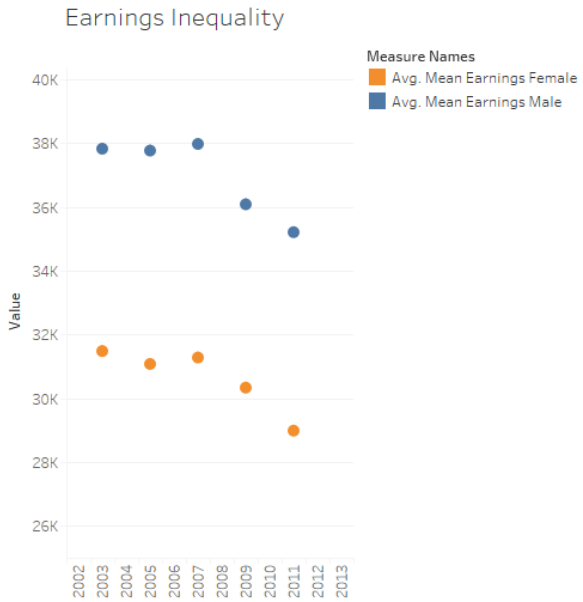
For classification, it was decided to use the tree-based approaches of bagging, random forest, and Adaboost. Tree-based methods offer a look at the importance of each variable in the resulting model. Linear and quadratic discriminant analysis and naive Bayes classification were also used and compared to the tree-based methods.

IV. Evaluation Strategy and Results

A. Getting to Know the Data

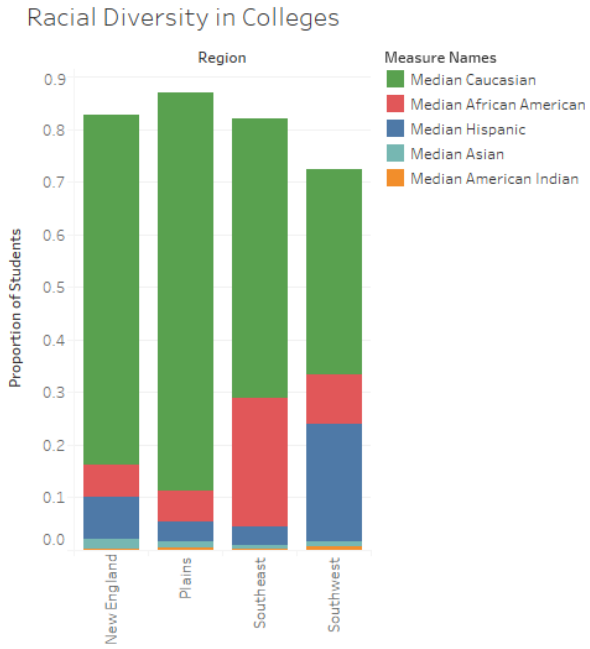
Tableau visualizations were useful for revealing interesting trends in the data on the 20,601 universities contained in the data set. The average tuition for both in-state and out-of-state universities continues to steadily rise, while average mean earnings for a university's graduates is falling and is unbalanced for men and women (Figure 1). Likewise, the southeast and southwest regions of the country includes the highest proportion of African-American and Hispanic-American students respectively (Figure 2(a)), while the average proportion of students studying business-related areas like marketing and economics is decreasing.

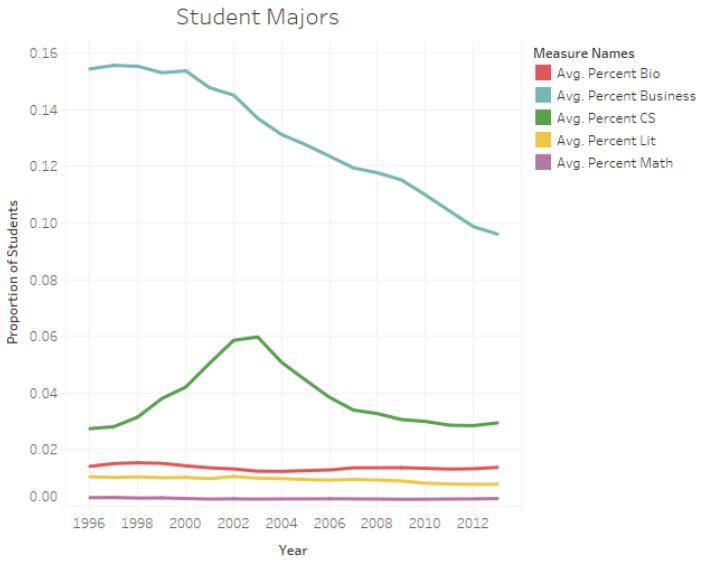




(a) (b)

Figure 1: (a) The average in-state and out-of-state tuition continues to increase. (b) The gender gap in average mean earnings for female and male university graduates.





(a) (b)

Figure 2: (a) The proportion of students of different races in different U.S. regions. (b) The average proportions of students in various majors, showing a decline in students studying business-related areas.

B. Discretization and Handling of Missing Values

The median earnings of graduates as reported by the universities exhibited a bell-shaped curve with mean at $34,930 and standard deviation of $8,982 (Figure 3).

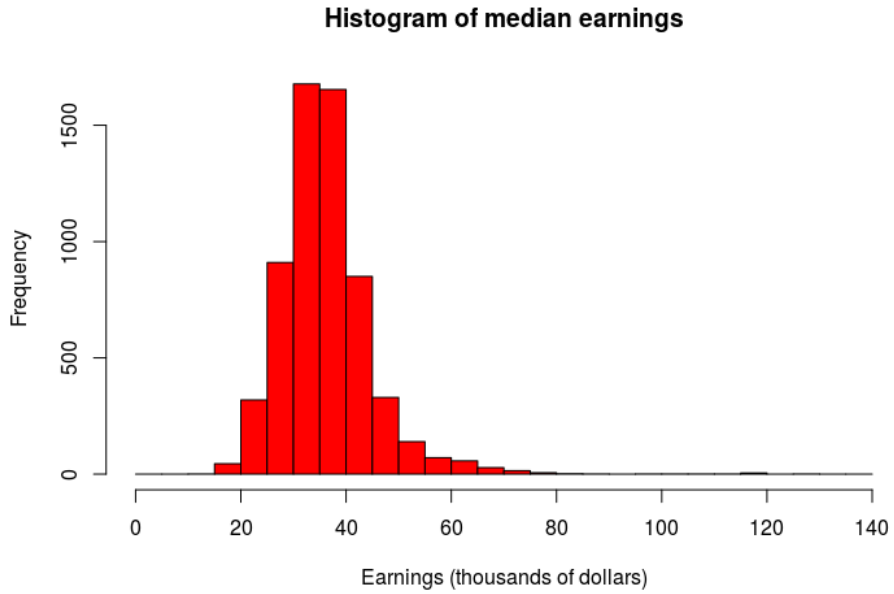


Figure 3: The histogram of median earnings of graduates exhibits a narrow, bell-shaped curve.

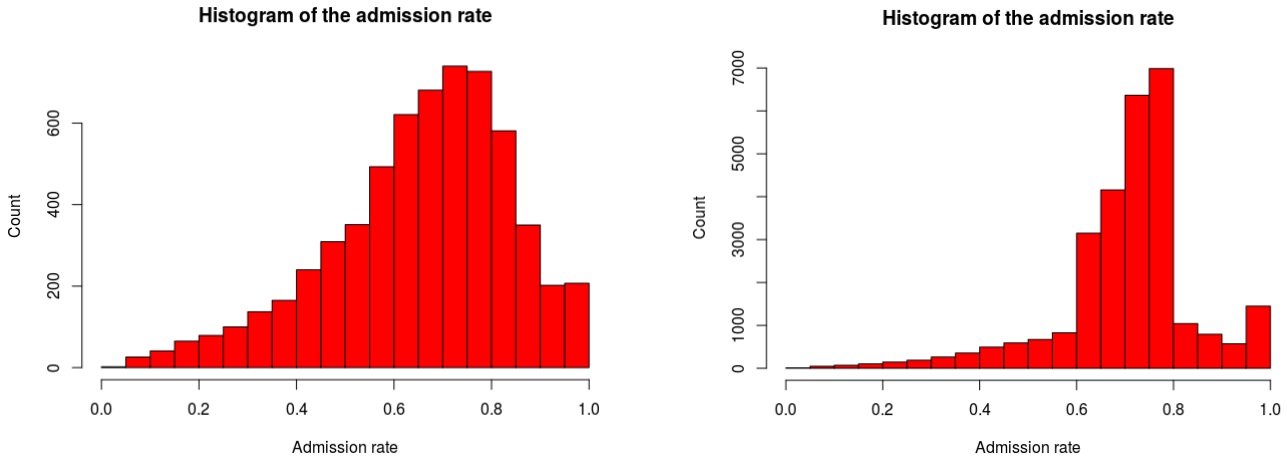
The discretization of median earnings into two and four classes gave balanced class sizes of uneven widths (Table 2).

Number of classes	Class boundaries	Number of observations
2	$0 - $27,300 $27,300.01 - $133,600	14,253 14,028
4	$0 - $21,800 $21,800.01 - $27,300 $27,300.01 - $34,300 $34,300.01 - $133,600	7,072 7,181 7,011 7,017

Table 2: Discretization of the median earnings variable into balanced classes. The value $133,600 is the maximum value for the median earnings variable.

After discretization of the dependent variable median earnings, the two strategies for the handling of missing values were executed. First, missing values for the independent variables were imputed with the means of each of those variables. In a second, independent approach, missing values for independent variables were imputed with the mean value from each median-earnings class.

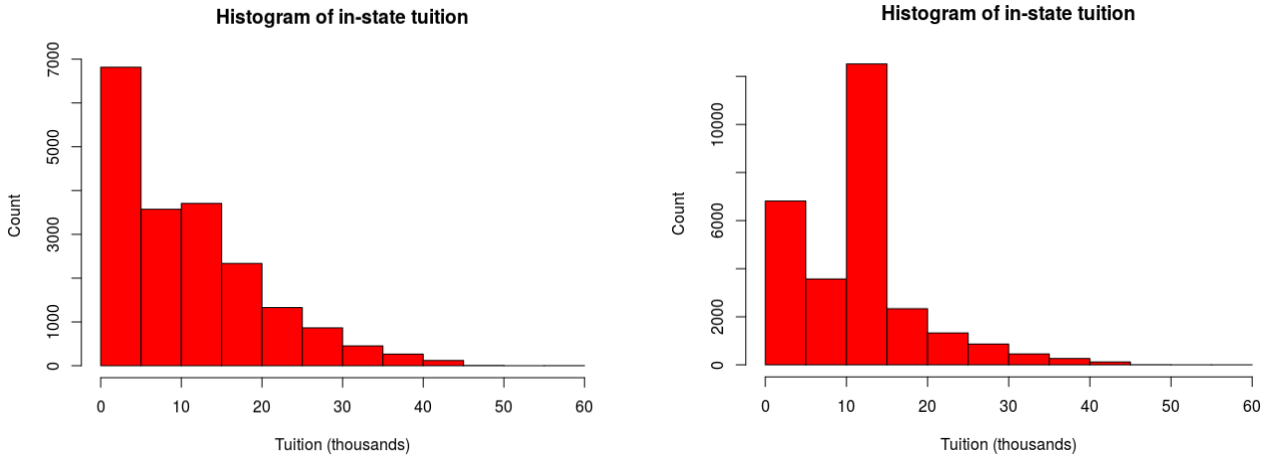
In both cases, the effect of imputing the missing values with mean values was to severely change the distribution of the independent variables. For example, for the admission rates of universities, the histograms reveal a dramatic shift in distribution after imputing the class means (where there are four different classes) for the missing values (Figure 4).



(a) (b)

Figure 4: The distribution of the admission rates of universities (a) prior to, and (b) after the imputation of missing values with the class mean (four-class case) admission rate after discretization of the median-earnings variable. The original distribution has been highly distorted.

The problem demonstrated in Figure 4 is even more pronounced when the missing values are replaced with the mean for the entire variable, without consideration of class (Figure 5).



(a) (b)

Figure 5: The distribution of in-state tuition of universities (a) prior to, and (b) after the imputation of missing values with the mean in-state tuition for all universities. The imputation strategy leads to a drastically changed distribution of the tuition variable.

Figure 4 and 5 highlight the inherent difficulties in imputing a large number of missing values: the resulting distributions are markedly different than the original ones, introducing bias into the sample of observations.

C. Regression Techniques

The first and most easy-to-interpret model was a linear regression model. Simple linear regression of median earnings onto the independent variables revealed interesting trends (Figure 6).

First, the p-values are below 0.05 for almost all of the independent variables, indicating evidence of an association between those independent variables and the dependent variable median earnings. Even more interesting are the coefficients of some of the variables: first, decreasing the admission rate by 1% leads to an average decrease in the predicted median earnings by nearly $4,000. Also, increasing the math SAT score leads to an increase in the predicted median earnings, but the opposite is true for increasing the verbal SAT score. Likewise, a decrease in the student body size leads to an increase in median earnings. Moreover, increasing the proportion of students in computer science or business majors leads to increased median earnings, while the opposite is true for biology and literature.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.909e+00  1.654e+00   2.363  0.01821 *
admit         -3.998e+00  8.348e-01  -4.789 1.77e-06 ***
satmath        1.237e-01  6.432e-03  19.227  < 2e-16 ***
satverbal     -7.398e-02  6.566e-03 -11.267  < 2e-16 ***
tuitionin     -1.178e-01  3.918e-02  -3.006  0.00268 **
tuitionout     2.716e-01  4.640e-02   5.853 5.47e-09 ***
pricecombined  1.956e-01  3.397e-02   5.759 9.52e-09 ***
studentbody   -2.620e-05  2.796e-05  -0.937  0.34888
cs             1.576e+01  3.901e+00   4.040 5.52e-05 ***
bio           -1.455e+01  3.251e+00  -4.475 8.00e-06 ***
math           3.963e+00  1.304e+01   0.304  0.76116
business       5.092e+00  1.062e+00   4.797 1.71e-06 ***
lit           -3.598e+01  5.425e+00  -6.632 4.05e-11 ***
```

Figure 6: Output of linear regression, indicating the p-values and coefficients of a linear model.

The test mean squared error ("MSE") for the linear regression model of Figure 6 was 48.56. On the other hand, when the median earnings were regressed onto only the "pricecombined" and "admit" variables, the test MSE increased to 69.05. This increase reveals that the larger model, with more independent variables, better captures the complexity of the factors affecting median earnings.

One of the advantages of lasso regression is its ability to use a penalty term to drive a model's coefficients to zero with minimal increase in test MSE. After using lasso regression on the data, the coefficients for verbal SAT score, in-state tuition, and the proportion of math majors were driven to zero (Figure 7). Lasso regression's test MSE is very close to that of linear regression, indicating that we can get the same predictive power out of a model that doesn't have these terms in it.

```
(Intercept)    1.033911e+00
admit         -1.178067e+00
satmath        5.783661e-02
satverbal      .
tuitionin      .
tuitionout     1.183743e-01
pricecombined  7.961140e-02
studentbody    3.808343e-09
cs             1.227716e+01
bio           -3.704095e+00
math           .
business       4.417756e+00
lit           -3.417594e+01
```

Figure 7: Output in the R IDE for model shrinkage using lasso regression. Some coefficients are reduced to zero (indicated by periods).

Polynomial regression of the median earnings onto the independent variables was then performed, with models up to (1) second-order terms and (2) third-order terms of each independent variable. As shown in Table 3, increasing the complexity of the model and adding more terms

improved the test MSE, indicating that the linear models are worse at capturing the true relationship between the median earnings and the independent variables.

All regression models using the dataset with imputation of missing values had higher test MSE compared to simply removing rows with missing values (Table 3). This suggests that the strategies used to impute missing values distorted the true relationship between median earnings and the independent variables, and that these imputation strategies are inadequate for this dataset.

Regression Technique	Test MSE (missing values not imputed)	Test MSE (missing values imputed)
linear	48.56022	94.08526
ridge	48.93166	94.09596
elastic net	48.94044	94.09955
lasso	48.94994	94.10378
polynomial (degree 2)	45.66559	**92.84564**
polynomial (degree 3)	**45.11173**	361.93108

Table 3: Test MSE for all regression techniques, with and without imputation of missing values. The lowest test MSE for each missing value strategy is shown in bold.

To further assess the effect of imputation of missing values, the adjusted R^2 values were compared for linear and polynomial regression in the cases where missing values were, or were not, imputed (Table 4). The imputation of missing values led to decreases in this statistic, indicating that the disruption of the underlying distributions of the independent variables resulting from the imputation of missing values led to worse model fits (specifically, decreases in the proportion of the variability in model earnings explained by regression onto the independent variables).

Regression Technique	Adjusted R^2 (missing values not imputed)	Adjusted R^2 (missing values imputed)
linear	0.401	0.211
polynomial (degree 2)	0.450	0.241
polynomial (degree 3)	0.470	0.253

Table 4: Assessments of model fit for linear and polynomial regression approaches. Imputation of missing values led to a decrease in adjusted R^2 for linear and polynomial regression models.

D. Classification Techniques

Various classification techniques were employed on the data, so as to identify the best-performing approach and compare their effectiveness (Table 5). The tree-bsaed methods of bagging, random forest, and Adaboost outperformed naive Bayes, LDA and QDA approaches.

Classification Technique	Error rate (2 classes)		Error rate (4 classes)	
	no imputation of missing values	missing values imputed	no imputation of missing values	missing values imputed
naive Bayes	0.3354037	0.06937275	0.5718862	0.49038965
LDA	0.2969669	0.05566731	0.5496368	0.27540958
QDA	0.3266437	0.06251522	0.5614421	0.35228067
bagging	0.2693425	0.02197383	0.5023611	0.07922838
random forest	0.2720668	0.02243743	**0.4966764**	**0.07891408**
Adaboost	**0.2615661**	**0.02050847**	0.5357201	0.08687812

Table 5: Estimated test error rates of classification techniques (lowest in bold).

The results reveal that the classifiers perform better for the 2-class case than for the 4-class case, and that imputation of the missing values leads to deteriorating classification performance. The variable importance plot for the random forest classifier, without imputation of missing values and for the two-class case, reveals that under both metrics of importance, the SAT math score and the student body size were the most important factors (Figure 8), and exceeded the importance of cost.

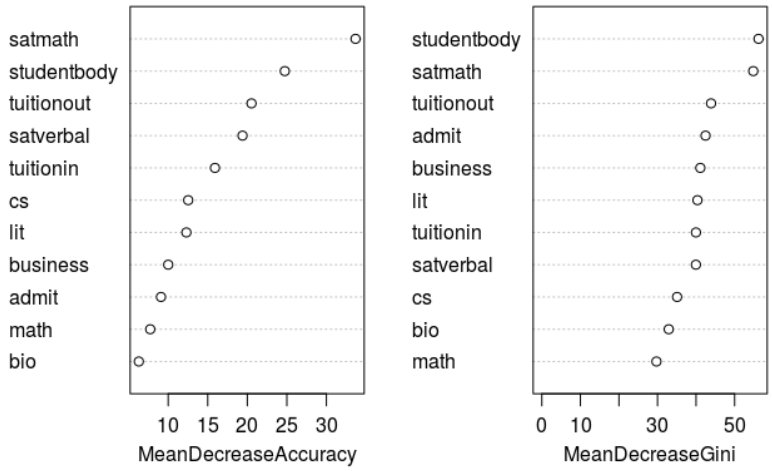


Figure 8: Example variable importance plots generated by the random forest classifier, for the 2-class case in which missing values were not imputed.

V. Pros and Cons of the Strategies

The strategy of using a variety of regression and classification techniques offered the advantage of producing multiple, independent lines of evidence regarding the relationship between median earnings of university graduates and the independent variables. For example, linear regression yielded evidence of both the existence and the nature of the relationship between certain independent variables and the median earnings. Likewise, comparing linear and polynomial regression (particularly, the growth in adjusted R^2 and the lower test MSE associated with polynomial regression) indicated that the factors influencing median earnings may be too complex for a simple linear model. Further, lasso regression was useful for eliminating the coefficients associated with some independent variables.

Likewise, the use of multiple classification techniques provided evidence that tree-based approaches are superior for this problem than probabilistic approaches like naive Bayes.

The employed strategies of imputing missing values, however, were clearly shown to be inadequate for this problem. Neither of the strategies led to reduced test MSE (for regression) or improved test error (for classification). Moreover, the large decreases in adjusted R^2 provide evidence that the use of these imputation strategies led to significantly worse model fits.

The strategies for discretization of the dependent variable led to insights about the relationship between the number of classes and classification performance. Across all classifiers, having only two classes instead of four led to improvement in the test error of each classifier. However, this improvement is offset by the fact that binary classifiers are arguably less useful, due to less discriminating power, than four-class classifiers.

Lastly, the literature research and careful review of the documentation proved useful to the results of this project, in that my understanding of the data set and its inherent limitations allowed me to better understand the results. Moreover, the use of Tableau for visualizing the data (even without performing any statistical analysis) revealed trends in the data that were relevant to my research question, for example the rising cost of postsecondary education and the decreasing earning power of college graduates and the continuing gender inequity of earnings.

VI. Conclusions and Future Directions

This project provided an opportunity to experience firsthand some of the challenges in working with large, real-world data sets, such as the potential difficulties associated with a large number of missing values and the complexity that can characterize the variable of interest (in this case, the median earnings of college graduates). Such complexity can lead to difficult decisions in terms of model fit, and requires a practitioner to temper his or her own expectations. It was useful to think of each statistical technique as producing evidence to support a possible answer to my research question.

This project revealed several possible areas of future work. Exploration of other approaches for imputing missing values, perhaps involving generation of classifiers for the missing values based on other available variables, might lead to improvements in the resulting models. Likewise, increasing the complexity of the models by experimenting with the addition of more variables might lead to models with better test set performance and increased model fit, indicating that those models better reflect the complicated nature of the median earnings of college graduates. Supplementing the data set with external sources, such as other databases maintained by the National Center for Education Statistics, might ameliorate the missing value problem and provide more data to improve the resulting models.

References

1. "Using Federal Data to Measure and Improve the Performance of U.S. Institutions of Higher Education," U.S. Department of Education, January 2017 (published at: <https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf>), accessed March 10, 2017.
2. "Better Information for Better College Choice & Institutional Performance," U.S. Department of Education, January 2017 (published at: <https://collegescorecard.ed.gov/assets/BetterInformationForBetterCollegeChoiceAndInstitutionalPerformance.pdf>), accessed March 10, 2017.
3. Kaggle web site (<https://www.kaggle.com/kaggle/college-scorecard>), accessed March-May 2017.