Predicting Student Debt Upon College Graduation

***Abstract*—This project aims to categorize the types colleges in the U.S. that give students the best value for their money. It takes advantage of the College Scorecard Data compiled by the U.S. Department of Education, which provides extensive information about colleges as well as student performance after graduation. We use a machine learning perspective to statistically analyze college data in order to put power in the hands of students who want to join the American higher education system. What factors indicate a school may have a higher rate of debt for its graduating students? Can we categorize universities based on information in the College Scorecard Data?**

I. Introduction

IN OUR MODERN CAPITALIST SOCIETY, which emphasizes productivity and material possessions, higher education is seen as the critical first step on the ladder leading to financial independence and economic stability. But as college tuition prices have skyrocketed in recent years, both for public and private universities, students often leave school with debilitating debt that delays or even prevents any economic progression in the United States. Furthermore, current financial aid models contribute to inequality among students and make understanding the real cost of a college education difficult.

The College Scorecard Dataset was an attempt by the U.S. Department of Education to provide more detailed information to prospective college students, enabling them to make a more informed decision about their choice of college. It contains hundreds of variables associated with some 124,000 college campuses such as admission rate, retention rate, proportion of first generation students and the proportion of students on federal Pell grant funding.

II. Data and Methods

The College Scorecard project was designed to put power in the hands of students and families to compare how well individual postsecondary institutions prepare their students for success by accounting for their own needs and educational goals. The dataset is provided through federal reporting from institutions, data on federal financial aid, and tax information and provides insights into the performance of institutions that receive federal financial aid dollars, as well as the outcomes of the students of those institutions. The U.S. Department of Education has given the public access to the most reliable and comprehensive data on students outcomes at specific colleges, including former students earnings, graduates student debt, and borrowers repayment rates. These data are also available for various sub-groups, like first generation and Pell students. The public nature of this dataset ensures that researchers, policymakers, and members of the public can customize their own analysis of college performance more quickly and easily. We use multiple linear regression as our base model before running PCA to better understand the latent trends in the data and the relationships between different variables. In order to select the best linear model we will employ best subset selection to determine the optimal number of predictors as well as what predictors should be included. We then attempt to improve our regression model using the information from a principal components analysis. In order to present our results in a way that is easier to interpret we also created a regression tree in order to highlight the categories of universities that resulted in the lowest debt upon graduation.

III. Data Exploration and Visualization

Before running any rigorous analysis on the data, we decided to get a feel for what they looked like. One of the most hotly-debated aspects of the American higher education system is whether public or private schools are a better choice for students. To start we generated a simple plot comparing average cost of attendance for public, private for-profit and private non-profit schools (Fig. 1). From the plot we can see that private schools are the most expensive, but it is possible to find private schools that cost less than public ones. Furthermore, this plot does not take financial aid into account. Unfortunately, analysis of financial aid is beyond the scope of these data.

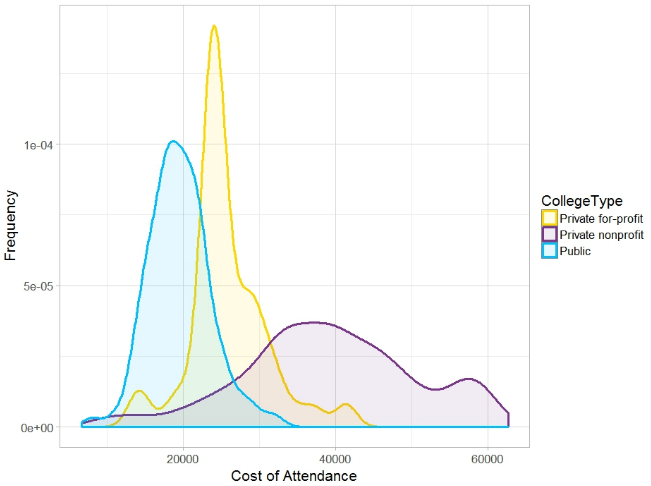


Fig. 1. Average cost of attendance grouped by control type.

From here, we decided to look at median earnings ten years after matriculation (Fig. 2). The plot shows that earnings are similar for most public and private schools, which indicates that attending a public school may be a more fiscally responsible decision for students who are struggling financially. For many schools, tuition is less expensive at a public rather than a private school, which points to students at public universities graduating with less debt. Before seeing this plot we had largely ignored private for-profit universities. However, it is also interesting to note that the median earnings distribution for private for-profit schools is distinctly bimodal, unlike the distributions of median earnings for public and private non-profit schools.

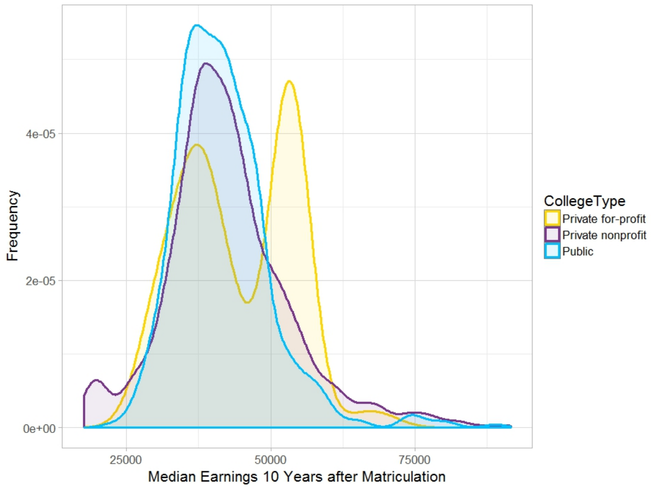


Fig. 2. Median earnings ten years after matriculation grouped by control type.

Can we pull out a list of the specific schools that provide the highest earnings after matriculation? (Fig. 3) Examining the boxplot, we note that many of these schools are private, meaning that although on average earnings after matriculation have similar distributions for public and private schools, most of the schools with the top quartile earnings in the United States are private.

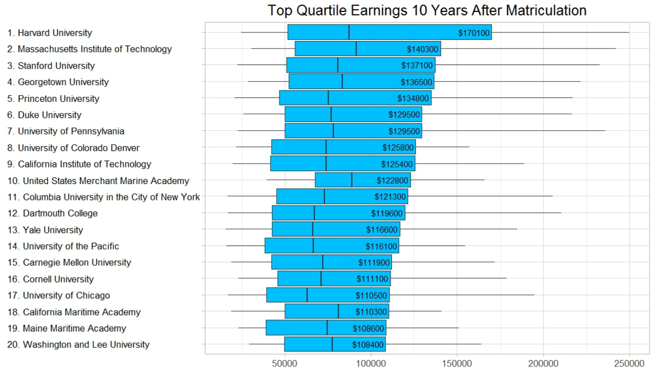
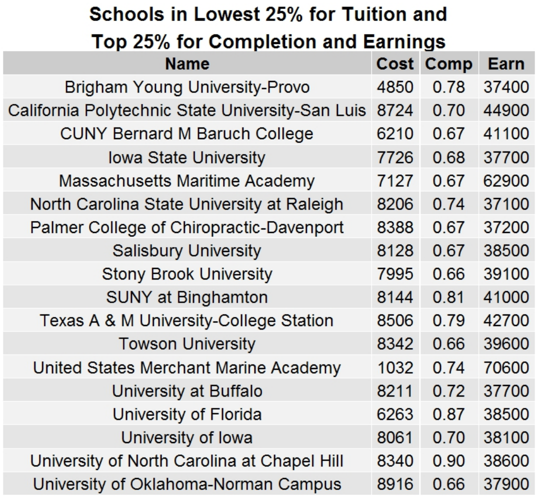


Fig. 3. Median earnings ten years after matriculation grouped by control.

This led us to question what schools provide top post-matriculation earnings for their students while simultaneously having a low cost of attendance? To do this we used data from 2011, which is the latest year that had median earnings recorded. Then we identified those schools that were in the lowest quartile for cost of attendance and the highest quartile for median earnings after matriculation as well as for six year completion rate. The results from our findings are detailed in the figure below (Fig. 4).



Schools in Lowest 25% for Tuition and
Top 25% for Completion and Earnings

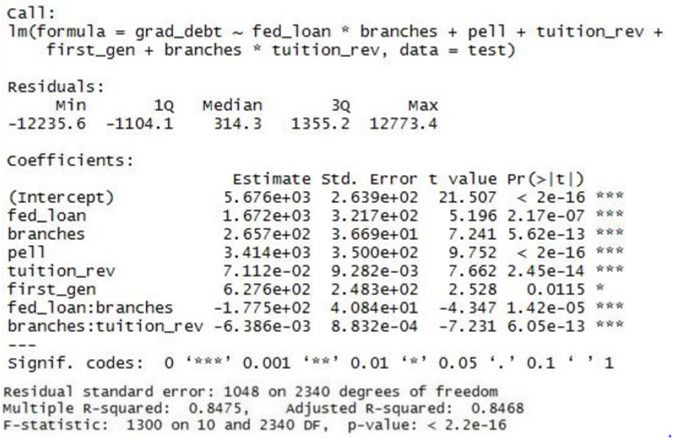
Name	Cost	Comp	Earn
Brigham Young University-Provo	4850	0.78	37400
California Polytechnic State University-San Luis	8724	0.70	44900
CUNY Bernard M Baruch College	6210	0.67	41100
Iowa State University	7726	0.68	37700
Massachusetts Maritime Academy	7127	0.67	62900
North Carolina State University at Raleigh	8206	0.74	37100
Palmer College of Chiropractic-Davenport	8388	0.67	37200
Salisbury University	8128	0.67	38500
Stony Brook University	7995	0.66	39100
SUNY at Binghamton	8144	0.81	41000
Texas A & M University-College Station	8506	0.79	42700
Towson University	8342	0.66	39600
United States Merchant Marine Academy	1032	0.74	70600
University at Buffalo	8211	0.72	37700
University of Florida	6263	0.87	38500
University of Iowa	8061	0.70	38100
University of North Carolina at Chapel Hill	8340	0.90	38600
University of Oklahoma-Norman Campus	8916	0.66	37900

Fig. 4. Median earnings ten years after matriculation grouped by control type.

IV. Machine Learning Techniques

After exploring the data through various graphs and tables, we started the cleaning process. In order to do this we were somewhat slaves to the dataset, as we could only use those variables that had enough values to actually run analysis on. Variables with too many missing values had to be excluded so that our results would not be too colored by bias. Our base model for this project was a multiple linear regression. From the available variables we thought it would be most interesting to predict debt upon graduation. To begin we split the dataset into training and test sets. Then, using exhaustive best subset regression on the training data we determined that the best number of predictors was 4, as adding more predictors after this did not result in a significant reduction in error. Then we selected the top four predictors: the proportion of students on federal loans, whether the school was public, private for-profit or private non-profit, the proportion of students on federal Pell grants and median faculty salary. Running a linear model using these predictors on the test set was not successful–the adjusted R^2 value was only about 0.15 for our model. This made sense because our original exploration of the data showed that there was a lot of collinearity in the predictor variables.

From here we went on to run a PCA on the dataset to see if we could better identify the collinear predictors of graduating debt. The biplot of the data showed the clusters to be separated at least in part by the proportion of students on federal loans and the number of branches the college has. Using the information from the PCA about the relationship between variables, we ran a new regression model. Our new model took advantage of collinear predictors by using interaction terms between these variables. This regression model had an adjusted R^2 of 0.84 and was therefore a great improvement over the original model. Our individual predictors were also more significant than the predictors used in the original model. Figure 5 on the next page shows the R-output from this model.



```
Call:
lm(formula = grad_debt ~ fed_loan * branches + pell + tuition_rev +
    first_gen + branches * tuition_rev, data = test)

Residuals:
     Min       1Q   Median       3Q      Max
-12235.6  -1104.1    314.3   1355.2  12773.4

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.676e+03  2.639e+02  21.507  < 2e-16 ***
fed_loan              1.672e+03  3.217e+02   5.196 2.17e-07 ***
branches              2.657e+02  3.669e+01   7.241 5.62e-13 ***
pell                  3.414e+03  3.500e+02   9.752  < 2e-16 ***
tuition_rev           7.112e-02  9.282e-03   7.662 2.45e-14 ***
first_gen             6.276e+02  2.483e+02   2.528   0.0115 *
fed_loan:branches    -1.775e+02  4.084e+01  -4.347 1.42e-05 ***
branches:tuition_rev -6.386e-03  8.832e-04  -7.231 6.05e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1048 on 2340 degrees of freedom
Multiple R-squared:  0.8475,    Adjusted R-squared:  0.8468
F-statistic:  1300 on 10 and 2340 DF,  p-value: < 2.2e-16
```

Fig. 5. R-output from our second improved regression model.

We also produced a regression tree to identify variables that are the most statistically significant for predicting debt upon graduation. There are five variables that are used in the tree construction, the four predictors used in our original linear region and the highest degree available from the institution. Figure 6 shows the final pruned tree. Next, we ran random forests, bagging and boosting in order to make sure that our most important variables were stable. Since our data has a high variance, our test MSE turns out to be very big, but we still see a decrease in our test MSE from bagging to random forests, which indicates that random forests yielded an improvement over bagging in this case. The results show that across all of the trees considered in the random forest, the most important variables were the proportion of students on federal loans, the number of branch campuses, the proportion of students on federal Pell grants and average tuition revenue per student. Interestingly, the trees in our random forest all chose a different set of predictors than our original linear model. Boosting indicated the same four predictors of debt upon graduation. However, when we used the boosted model to predict debt upon graduation, the test MSE is bigger than the test MSE from random forest and bagging.

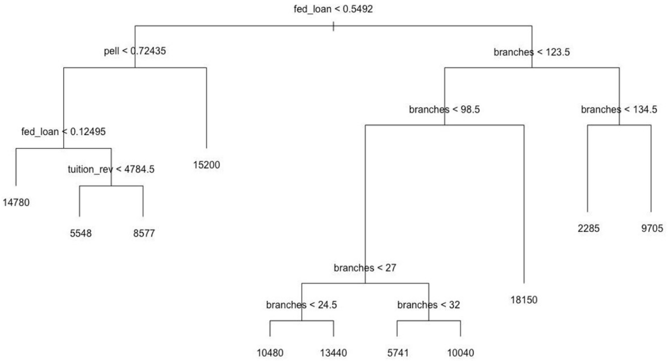


Fig. 6. Our final pruned regression tree to predict graduate debt.

V. Discussion and Future Work

It would be interesting to combine the results of our machine learning analysis with the plots and interactive map we created at the start of the project. Having visual aids as well as mathematical evidence would help prospective students to better understand what indicators they should focus on when choosing a university. Given more time, we would like to make our analysis more interactive in order to better appeal to the audience of high school students interested in attending college. We would also like to spend more time analyzing the clusters that were visible in our PCA–what differences are there between the schools in these clusters and how does the best predictive model vary from cluster to cluster. Using other principal components we would also like to try to identify more potential clusters in the dataset. We would also like to explore why the distribution of median earnings was bimodal for private for-profit colleges. Looking at this group of colleges might be fruitful as well, given that generally fairly little is known about them. At the end of the day, we were able to use a number of machine learning methods on this dataset but it was simply too extensive and detailed for us to do it full justice in the time we had for this project.