Predicting Student Earnings After College

Miranda Strand MSTRAND@STANFORD.EDU
Tommy Truong TOMMYT@STANFORD.EDU

1. Introduction

Many students see college as an investment to help them earn more and live better lives after graduation. While it is true that college graduates earn more on average than those without a degree, large numbers of students today are graduating with worrying amounts of debt, calling into question the assumption that attending college is always the wisest investment. It has become more important then to understand the factors that contribute to post-graduation earnings and the ability to repay student loans.

A common belief is that the prestige of a university affects the future income of college students. But prestige is likely not the only factor. We looked more closely at some of the other variables associated with a college that could potentially predict the financial future of its students.

Our goal was to create a model that would accurately predict the earnings of a college's graduates given specific features of the college, such as its acceptance rate, average test scores, student body demographics, and the student loans needed to attend. The insights provided by such a model could give incoming college students greater knowledge about the features to consider when choosing a college. The results also yield interesting insights about the American higher education system on the whole.

2. Related Work

There has been interest in the relationship between college education and earnings after college for many decades. In our exploration of the previous literature on this topic, we came across several papers that looked at the effect of college selectivity and quality to earnings of students.

In their paper, Brewer, Eide, and Ehrenberg (Brewer, 1999) built a choice model to determine a students earnings, using the assumption that a student would pick a certain type of college based on their individual characteristics. We thought that using individual characteristics to build this choice model, instead of solely relying on college-specific data, was clever, but we felt that their reduction of colleges into only six classes potentially lost subtle but significant differences between colleges.

Rumberger and Thomas (Rumberger, 1993) similarly considered both individual and college features in studying the impact of three variables on earnings after college: college major, school quality, and student academic performance. We thought that they were astute in using hierarchical linear modeling to address the fact that their data was composed of nested samples of students all in the same few colleges.

In another similar paper, Loury and Garman (Loury, 1995) take a more economics-minded approach to building their model by assuming that students would try to maximize net earnings by picking a college where the marginal product of attending the college would equal its marginal cost. We liked that they considered several potential costs of attending a selective college, such as higher tuition and increased likelihood of failing to graduate, but we felt that their study was limited by their data, which only looked at male students who were either white or black.

Oddly enough, James, Alsalam, Conaty, and To (James, 1989) similarly limited their paper on this issue to male students only. However, we thought they had a clever approach of incrementally building their model by selectively adding more feature sets.

In a different approach, Wachtel (Wachtel, 1976) focused on looking at the effects of increased investment in college in relation to earnings instead of considering a variety of both individual and college features. We thought his concentration on just two expenditure categories, the amount of time spent in college and the amount of money spent per year in college, helped make his paper more targeted and focused. However, Wachtels data was even more limited than the previous papers - it only had information for white, male volunteers for Army training tests.

The major limitation common to all of these papers is that they relied on data collected almost twenty years before their publication, and the scope of the college

data they collected is considerably smaller than that of the College Scorecard dataset that we used in our project. This Scorecard dataset has only recently been made publicly available by the government, which explains why we did not find literature making use of this dataset. We thus have the privilege of working with a large, up-to-date, comprehensive dataset, with many more features than any dataset in the literature we researched.

3. Dataset and Features

In 2013, the US Department of Education matched information from the college financial aid system with federal tax returns of the graduates of those colleges, creating the College Scorecard dataset, a wealth of information intended to help students and families make the best decisions about where to attend college. For the almost 8,000 colleges included, there are over a thousand fields, including demographics about the students at each college, the degrees and majors offered, the cost and average loans taken out, students test scores, admission rates, and more, matched with statistics for rates of repayment of student loans, and the distributions of graduates incomes over the course of the ten years following graduation.

Not all of the data was relevant to our task. We chose the mean income 10 years after graduation as our response variable, and eliminated the many other fields pertaining to post-graduation income, as well as those describing the loan repayment patterns and death rates of graduates. We then focused primarily on a set of 32 features provided by the US Department of Treasury, including gender, age, ethnic, and income demographics of students.

To gain an understanding of the data, we ran Principal Components Analysis (PCA) on the scaled and centered features of each college to reduce them to a visualizable number of dimensions. PCA works by projecting the data onto a k dimensional subspace in which the basis vectors for the subspace are the top k eigenvectors of the original data. This serves to maximize the variance of the projections onto the subspace, preserving as much as possible of the data's original variance.

After performing PCA, we plotted our reduced data points in two and three dimensions. To visualize the relationship between the reduced features, and post-graduation earnings, we scaled each college's point by the mean income of its graduates and colored it according to whether that mean income was above or below the average for all colleges. Looking at the resulting graph, we found that the first principal component of the features, which captured about one third of the overall variance in the data, also seemed to capture some of the variation in post-grad earnings, as shown in Figure 1.

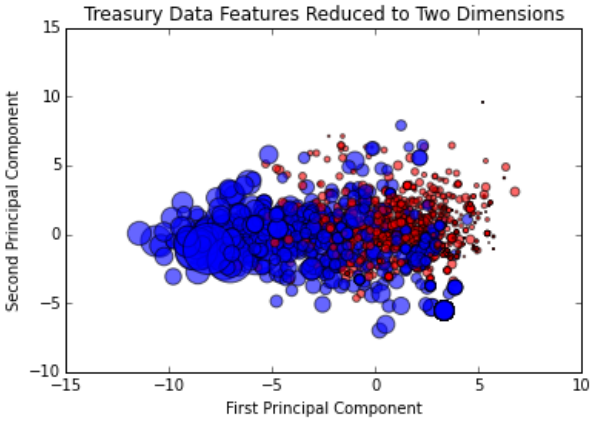

Figure 1. Colleges with mean graduate income above the average are shown in blue; those below the average are in red. Points are scaled according to the magnitude of the mean graduate income.

Along with the features from the treasury data, we then added the admission rates of the schools, and the midpoint SAT scores of their students. This reduced the size of the dataset even further, but running PCA on the augmented set of features, we found again that the first principal component captured much of the variation in future earnings.

4. Methods

We sought to perform a regression on students' mean income ten years after graduation. To do so, we began with linear regression, which fits a coefficient vector θ so as to minimize the residual sum of squares

$$\frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2$$

where each $x^{(i)}$ is a training example (vector of college features, with $x_0 = 1$ for the intercept), and $y^{(i)}$ is its response (mean post-graduate income). By viewing each data point as a row of a matrix X, linear regression can also be solved using the normal equations, $\theta = (X^T X)^{-1} X^T \vec{y}$, which correspond to setting the derivative of the original least squares cost function to 0. But from this equation, we can see that least squares will suffer when the features are collinear. In the case of perfect collinearity, $X^T X$ is not even invert-

ible. A nearly singular X^TX will still cause increased variance in the model.

We knew that many of our features were likely to have collinearities. For example, Pell Grants are awarded based on family income, so the percentage of students receiving Pell Grants would undoubtedly be correlated with the mean household income of students' families. It is almost certain, too, that less obvious correlations exist among the different demographic statistics of schools.

To make the model more robust to collinearity, we introduced a degree of bias to the regression, imposing a penalty term constraining the norm of the coefficient vector. Ridge regression penalizes the squared L_2 norm, with the cost function

$$\frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2 + \lambda||\theta||_2^2$$

The solution to the normal equations then becomes $\theta = (X^TX + \lambda I)^{-1}X^T\vec{y}$, resolving the previous need to invert a singular matrix.

The Lasso, similarly, introduces a penalty term, but it uses instead the L_1 norm

$$\frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2 + \lambda||\theta||_1$$

which has the advantage of performing a type of feature selection by forcing coefficients to be 0, giving a sparse solution.

In general, given the multicollinearity of our data, we found greater success with models that perform an inherent feature selection. In addition to the Lasso, we also tried Random Forest regression. Random Forests work by building a series of decision trees on the training data.

A decision tree is formed by partitioning the data one variable at a time. These partitions are made by choosing a region, a prediction, and splitting point in order to produce the largest decrease in the residual sum of squares. To make a prediction on a new datapoint, we find the the partition that the point lands in, and predict the mean value of training points in that space.

In a Random Forest, we make a series of decision trees; to predict, we take the mean prediction from all of them. In forming each tree, we also choose a random subset of features to consider at each step. The result is that we build uncorrelated trees, making the model more robust to multicollinearity in the data.

Table 1. An example of MSE using the treasury, admission, and SAT features.

Model	Training Set MSE	CV Set MSE
Baseline	115232806.61	145807576.00
Linear Regr.	31778703.19	40891468.87
Ridge Regr.	35726792.71	39095726.91
Lasso	35485976.76	35465442.07
Random Forest	6881504.09	39358187.26

5. Experiments

5.1. Regression Models

To measure accuracy of our models, we used hold-out cross validation. We set aside a random 30% of our data and calculated the cross-validation set error on this data as an estimate of the generalization error. We compared this error to the baseline of computing the average post-graduate mean income, and predicting that for every college.

As we looked at the financial aid data from the Department of Treasury, combined with the SAT score and admission rate statistics–a set of 36 features–we faced a substantial problem of missing data. Many colleges were lacking a large number of fields, either due to unavailable data or privacy concerns. To start, we removed these data points. But in doing so, we reduced the size of our data set immensely from 7804 to 289 colleges.

Even a simple linear regression on the treasury, admission and SAT data fared significantly better than the baseline, as shown in Table 1. For context, note that a Mean Squared Error (MSE) of 40891468.87 is a mean difference of $6394.64 between the predicted and actual mean incomes, which is about 15% of the average mean income. As expected from the multicollinear features, though, the linear regression model appeared to have very high variance. The MSE of the training set tended to be about ten million dollars lower than that of the cross-validation set. Re-running the model with different choices of training and cross-validation sets also resulted in changes to the MSE on the order of ten million.

Ridge Regression and the Lasso both improved on the cross validation set error. To choose values for the penalty term multipliers, we ran many trials. For the Lasso, the best multipliers seemed to be around 20; for Ridge Regression, they were about one half. In particular, the Lasso's ability to perform feature selection seemed helpful. With the best choice of hyper-

Table 2. An example of MSE using just the treasury features.

Model	Training Set MSE	CV Set MSE
Baseline	115232806.61	145807576.00
Ridge Regr.	23541870.69	30063602.47
Lasso	23584437.60	29858251.28
Random Forest	2114891.30	23049251.34

parameter, 9 out of the 36 features were eliminated, including the age of students upon college entry, percentages of their marital and veteran statuses, and some logarithmic transformations of family income.

We still seemed to face a problem of variance, though, even in the penalized models. Except for with Lasso, there was a high discrepancy between training set and cross validation error, and changing these sets still resulted in substantial changes to the MSE. To address the variance issue, we needed a smaller set of features, or a larger set of training examples. Given the number of NULL and PrivacySuppressed data points that we had removed, these two goals could actually go hand in hand sometimes.

Since the SAT and admission rate data were missing for a majority of schools, removing those features allowed us to expand the size of our dataset from 289 to 1,664 colleges. On the larger training set with fewer features, all of the models performed better, though still with some variance. The Random Forest regressor stood out in paticular with the best results and the least variance between trials.

5.2. Feature Selection

One of the more interesting aspects of our project was identifying the most important features of a college that determine student earnings. To achieve this, we used two feature selection methods to see which features were the most important predictors in our model.

Since the Lasso performed so well, we decided to use it in conjunction with recursive feature elimination to identify its five most important features. Recursive feature elimination first trains the Lasso on all features, prunes the features with the lowest learned weights, then recursively trains and prunes on the smaller set of features until only a few are left. The top five features here were percentage of students who received a Pell grant, percentage of dependent students, percentage of female students, percentage of first-generation students, and percentage of students who sent FAFSA applications five or more schools.

The Random Forest regressor also performed well in predicting earnings, and it also conveniently assigns importances to features automatically, based on which features were used to make splits in the decision trees. The five highest-ranked features here were percentage of students who received a federal loan for college, the midpoint SAT scores of the college for each of reading, math, and writing, and the college's admission rate.

We then plotted each of these individual features against mean earnings after graduation. A few of the results were as expected; there was an obvious positive correlation between SAT scores and mean earnings. Students with higher SAT scores are higher-achieving and can attend more selective and distinguished schools, and thus earn more after graduation. Additionally, there were negative correlations between admission rate, percentage of Pell recipients, and first-generation students. Schools with lower admission rates can be more selective and admit high-achieving students. Students who receive Pell grants and first-generation students typically come from poorer or less-educated family backgrounds, and thus will tend to earn less after graduation due to the challenges of moving out of an economic class.

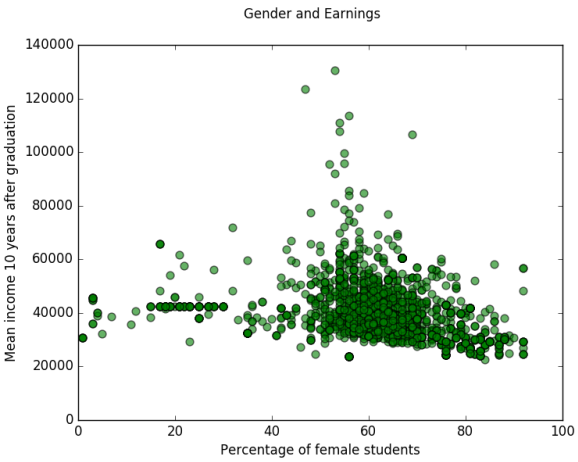

Figure 2. The colleges with highest student earnings typically had a roughly even split between the two genders.

However, there were interesting results that we did not expect. We found that the schools with higher post-graduate earnings typically had a nearly even split between male and female students (see Figure 2). We believe that this is the case because more prestigious schools will have many applicants and are thus more able to admit an evenly-split class of qualified students.

In addition, schools with many students who submitted more than five FAFSA applications tended to have higher post-graduate earnings (see Figure 3), which at first glance seems to contradict the Pell grant trend we observed. We then reasoned that students who took the time to apply to many colleges tend to be more ambitious and high-achieving, and schools with many of these students must be attractive enough to convince students to attend their school instead of the other schools they applied to.

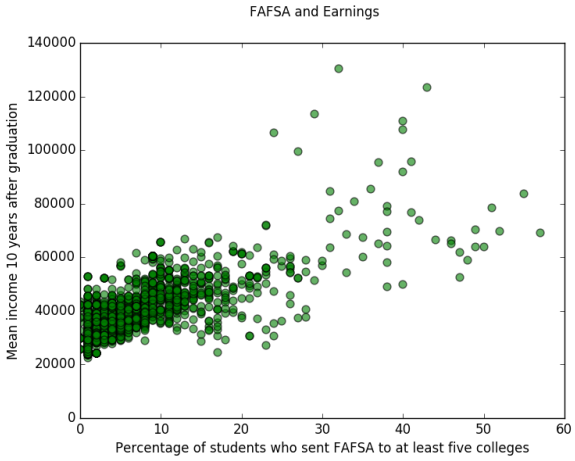

Figure 3. Colleges with many students who sent FAFSA apps to many colleges tended to have higher earnings.

5.3. Imputation

One of the most challenging aspects of working with the Scorecard data was handling the missing (NULL or PrivacySuppressed) values. One common and simple strategy to handle this is to throw out examples with missing values, but this could potentially cause models to miss out on valuable information available from the non-missing values in these examples. In our case, many colleges were missing at least one feature, so performing this strategy reduced the number of training examples available from around 5000 to around 1000.

We thus tried to use imputation to substitute missing values with estimated values. We replaced missing values with the mean of the present values for that particular feature. The benefit of doing this is that it preserves the sample mean for each feature, and more importantly, it allows the model to train and make predictions on examples that are missing features, increasing our sample size and making our model more robust to incomplete data. However, imputation adds noise and makes it harder to observe correlations between variables because significantly different training examples would both be given the same value for the same missing feature.

Table 3. Performance without and with imputation.

Model	MSE W/O Imp.	MSE With Imp.
Baseline	145807576.00	168041319.68
Ridge Regr.	30063602.47	111845543.08
Lasso	29858251.28	110042792.78
Random Forest	23049251.34	79082666.98

Performing imputation increased our sample size back to around 5000 schools (we still left out schools missing values for our response variable) but reduced the performance of our regression models, as shown in Table 3. This is understandable given that some features, such as SAT midpoint scores, were missing values for over 6000 schools; for these features, the estimated imputed values overwhelmed the actual observed values. However, the model can still make a decent prediction on a new example with missing values, whereas without imputation this would not have been possible.

6. Conclusion and Future Work

For our project, we used the College Scorecard dataset to build a model that could predict the earnings of a colleges students after graduation. We also gained insight into what characteristics of a college are important in determining the earnings of their students.

We used a few different regression algorithms and found that Lasso and Random Forests yielded the lowest mean squared errors. We believe that these two algorithms performed the best because they both perform a type of feature selection (which reduces high variance); Lasso uses regularization to force the coefficients of the least useful features to 0, while Random Forests assigns importances to features when using them to make splits in decision trees. This property allows these two in particular to perform well on our large dataset that contains hundreds of features.

If we had more time for future work, we would like to develop better ways of visualizing the dataset. The sheer number of features and colleges contained in the data makes it hard to grasp, and it would be worthwhile to create an application that can project the data onto custom features or components to yield visible insights on the relationship between college and earnings. We might also try more unsupervised approaches to group similar colleges together, thereby providing possible alternatives to attending a specific college.

References

Brewer, D. J., Eide E. R. Ehrenberg R. G. Does it pay to attend an elite private college? cross-cohort evidence on the effects of college type on earnings. *The Journal of Human Resources*, 34(1):104–203, 1999.

James, E., Alsalam N. Conaty J. C. To D.-L. College quality and future earnings: Where should you send your child to college? *The American Economic Review*, 79(2):247–252, 1989.

Loury, L. D., Garman D. College selectivity and earnings. *Journal of Labor Economics*, 13(2):289–308, 1995.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rumberger, R. W., Thomas S. L. The economic returns to college major, quality and performance: A multilevel analysis of recent graduates. *Economics of Education Review*, 12(1):1–19, 1993.

Wachtel, P. The effect on earnings of school and college investment expenditures. 58(3):326–331, 1976.