Prediction of Post-Collegiate Earnings and Debt

Monica Agrawal, Priya Ganesan, Keith Wyngarden

Stanford University

I. Introduction

Background

The U.S. Department of Education launched College Scorecard in September 2015 as a means of gathering more data on degree-granting institutions, the demographics of college students, and the status of alumni of these institutions [1]. By doing so, the U.S. Department of Education hopes to empower students to make more informed college decisions through a data-driven approach.

Considering the soaring cost of higher education as well as the accompanying rise of student debt, prospective students can greatly benefit from such information. However, College Scorecard has faced scrutiny due to its omission of over 700 colleges, particularly community colleges, in its data set [2]. Hence, applying machine learning to fill in omissions in the data set, particularly related to earnings and debt, and finding correlations between characteristics of colleges and the future success of their alumni has great value to society.

Despite the relevance of machine learning to this issue, fairly little research has been done in this area. Machine learning has been used in several related topics, such as predicting corporate earnings and predicting income based on census data about individuals [3, 4]. However, no research has been conducted on using college data to predict the earnings and debt of its alumni, potentially because higher-education institutions do not condone a solely numbers-based approach to the college selection process.

Goals

We hope to use a variety of machine learning models to make predictions regarding post-college earnings and debt of alumni who were on federal financial aid from various institutions based on factors that reflect the current status of each institution, such as majors and degrees offered, tuition, and admissions rates. Such statistics are easier to obtain than post-college earnings, so our predictions can be used to fill in gaps in the current data set and potentially unearth interesting factors that influence alumni earnings and debt. In addition, alumni earnings can be compared with tuition costs and average student debt to determine the typical interest and length of student loans for a particular school.

Previous Work

As College Scorecard is a newly-released data set and is more comprehensive than past college data sets, not much analysis has been done on College Scorecard or even on the topic of predicting post-collegiate earnings and debt. The most relevant past work in this area was conducted in the late 1980s and early 1990s.

Brewer et al. looked at the effect of college quality on future earnings based on individual and family characteristics of high school students entering into college, and found that elite private institutions had a higher return on investment in terms of future wages [5]. James et al. attempted to predict future earnings (for only male college graduates) using a mix of individual student information, institutional information, individual college experience variables, and labor market variables [6]. They found a general trend that selective private schools on the East Coast generally correlated to higher future earnings, but also found that the college experience variables contributed to the majority of the variance in the data. Hence, they concluded that each individual's college experience, and what each individual makes of the opportunities at his or her college, is the best indicator of future earnings. Lewis C. Solmon, one of the most widely-cited experts in this field, performed a study on what features determine college quality and what impact college quality has on earnings [7]. He used regression analysis to find that variables like college level, average S.A.T. scores, and average faculty salaries drove up alumni earnings the most.

While these papers have made large strides in using machine learning to understand what fuels alumni earnings, and were very careful in avoiding bias with respect to minority communities and other similar factors, they also have some shortcomings. All of these studies were based off of individual alumni data (things like personal and family background, individual major, etc); no one has yet attempted to predict alumni earnings and debt solely based off of anonymized institutional data. Furthermore, these studies focused on the most elite institutions and did not provide analysis on smaller and lesser-known institutions, which are the organizations that could most benefit from a study like ours.

As we were working with a new dataset, there were a number of data quality issues to resolve. These are largely detailed in the following section, but of particular note are metrics that had partially missing data (only some schools had listed values). There is is ample research on missing data problems in machine learning; Marlin (2008) gives an overview of major methods [8]. The most useful family of methods for our dataset is statistical imputation, which is detailed in Rubin (1996) in the context of an overview of multiple imputation [9]. We will return to these papers in the next section.

II. Data and Feature Set Preprocessing

Data

College Scorecard provides a publicly available data set consisting of approximately 2000 metrics for 7805 degree-granting institutions [1]. These metrics include demographic data, test scores, family income data, data about the percentages of students in each major, financial aid information, debt and debt repayment values, earnings of alumni several years after graduation, and more. We chose to focus on the 2011 data set because it was the least sparse data set in the last five years (more future earnings information was available than for more recent years). Our first tasks were to select variables to predict, transform the dataset into pairs of features and prediction variables, and segment the data for evaluation purposes.

Selecting Features and Prediction Values

We chose two values for our prediction variables – the median postgraduate debt and the median postgraduate earnings for alumni 6 years after graduation. We then went through several steps to prune the full feature set to an initial feature list.

We first eliminated all features that had non-numerical/categorical values (primarily school name). Additionally, we removed unrelated features that should not be used to make predictions, such as features that provided the number of students in different data collection cohorts.

We also removed all features related to debt, earnings, and repayment. All metrics in these categories are highly correlated with the two we chose to predict, so they would be weighted very strongly compared to other features and would hurt the ability of our models to generalize to schools without any of this information available, which is the motivation for this project.

Finally, after the preprocessing steps listed above and the non-standard data value processing described below, we removed all features (mostly null-indicators and unused categories) which had only one value for all examples, as they offer no predictive power.

Preprocessing Non-Standard Data Values

Some features in the data set were categorical fields; we chose to turn each category into separate indicator features. Many values in the data set were listed as "NULL", and a portion of these were meaningful (for example, indicating the absence of a binary feature) rather than indicative of missing data. In order to transform the nulls into usable numeric values while preserving the original meaning of the nulls, we replaced each null value with 0 and created an extra feature for each feature that contained null values. This new feature used 1s and 0s to indicate whether the value in the previous feature was null or non-null. For categorical fields that contained null values, we created just one null indicator feature in addition to the category indicators described previously.

Handling Privacy-Suppressed Values

All values in our dataset that were computed using data from fewer than 30 students were listed as "Privacy-Suppressed". Privacy-suppressed values are more common for smaller schools than larger schools and many privacy-suppressed values occurred in potentially useful metrics. One approach for handling these values was to simply remove all features with any privacy-suppressed entries. However, discarding hundreds of features in this fashion, especially for features with a low percentage of privacy-suppressed values, was undesirable.

In Marlin's overview of approaches to missing data, alternatives to case deletion (the above strategy) include mean imputation (setting missing values to the mean of observed values), regression imputation (learning regression models based on observed values), and the class of multiple imputation solutions (sampling multiple values from a simpler/generalized model over observed values and running analyses on each for later aggregation) [8]. We determined that mean imputation was not appropriate in this case, since many features of schools vary significantly based on school size, degree level, and so on.

We implemented regression imputation by training a linear regression model (with ordinary least squares cost function) to the fully-observed features with respect to each feature with privacy-suppressed values. To avoid training these models with limited data, we imposed a requirement that imputed features must have missing data for less than 30% of schools. We then replaced the missing values with predictions of the appropriate model. This is a single imputation method (though since the model cost function is convex, it is very similar to multiple imputation methods with this same choice of model). As noted by Rubin, multiple imputation methods capture variability of the data that is lost with single imputation [9]. Future work might involve using more generalized models for imputation, such as a mixture of Gaussians, and running multiple imputation.

Selecting Training and Testing Examples

We removed all examples (schools) that were missing the values for our two label variables: median postgraduate debt and median postgraduate earnings 6 years after graduation (among the provided options of 6, 8, and 10 years post-graduation, 6 years had the least sparse data). From the remaining examples, we set aside 3500 for training, 1000 for development, and the rest (∼800) for testing.

III. Prediction Models and Methodology

Linear Regression

We pose our learning task as a regression problem: given a processed list of features for a school, we would like to predict real values for that school's students' median debt at graduation and median earnings 6 years after graduation. Linear regression is a natural choice of baseline model for regression problems, so we first ran simple linear regression on the full feature set (including imputation of privacy-suppressed features), using our 3500 training examples and 1000 development examples. The performance of this baseline was 12.97% mean absolute percent error (average of the absolute values of percent error made on each soon) on the development set for earnings and 20.20% for debt. In addition to tuning the number of privacy-suppressed features to include in the feature set, we saw two avenues for lowering this error: pruning the feature space and enabling our model to learn nonlinear relationships between the features and earnings/debt.

Feature Selection

After data preprocessing and statistical imputation of privacy-suppressed values, 599 features remained. This is a large number of features in comparison to the training set size of 3500 schools, especially as we moved from simple linear regression to more complex models. We therefore explored the use of feature selection to shrink the number of input features.

To select the most important features to keep, we ran sequential forward-based feature selection on our 3500 training examples, using our median earnings prediction variable and median debt prediction variable in turn to evaluate and select the most relevant features [10]. Features were selected based on their mean-squared error, using 10-fold-cross-validation, and selection was terminated at the point where the prediction error stabilized. This procedure yielded 170 features for earnings prediction and 165 features for debt prediction, with 70 features in common.

The top 5 features yielded after running statistical imputation and feature selection were:

	Earnings	Debt
1	% financially independent students with family incomes between $0-30,000	Predominant degree awarded
2	% degrees awarded in Personal And Culinary Services	Average net price for $0-$48,000 family income
3	Total share of enrollment of Asian undergraduate degree-seeking students	% first-generation students withdrawn from original institution within 3 years
4	% students who are financially independent	% with status unknown within 3 years at original institution
5	Predominant degree awarded	% who transferred to a 2-year institution and withdrew within 2 years

Table 1: *Top features for median earnings and debt.*

We also tried using PCA on the school/feature data matrix to transform the data into a smaller set of uncorrelated model inputs. After full optimization under each approach, a model using PCA performed only slightly worse than a model using forward-based feature selection. However, the use of PCA for feature selection would require collection of data for all features when adding new schools to the dataset, since the principal components need to be recomputed when the data matrix grows. By contrast, after running feature selection on the existing dataset, adding new schools to the dataset requires collecting data only for the selected feature subset. If too many new schools were added, the feature selection results may become outdated. However, since the number of examples in our task is limited by the number of colleges in the United States, and since the initial dataset is fairly comprehensive and the rate of school closures/openings is low compared to the total number of institutions, this is not a major concern.

Locally Weighted Linear Regression

To capture local nonlinearities between the features and debt/earnings, we added local weighting to the cost function for our linear regression model. Using the Euclidean norm, our weight function for a training example $x^{(i)}$ with respect to an input example x was:

$$w^{(i)} = \exp\left(- \frac{||x^{(i)} - x||^2}{\tau^2} \right)$$

To make the Euclidean distance (the norm in the equation above) meaningful, we standardized features to zero mean and unit variance prior to computing the weights. The parameter τ in the weighting function above was tuned on the development set data for various other model choices (feature selection, inclusion of privacy-suppressed values).

Figures 1 and 2 show the results of tuning τ for each output variable and model. We found that the best linear regression model on the development set used local weighting, feature selection, and imputation of privacy-suppressed values.

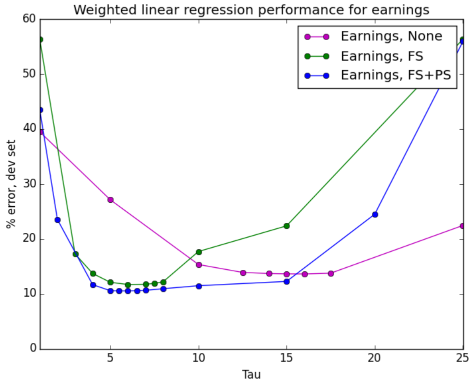

Figure 1: *τ values plotted against percent error for median earnings.*

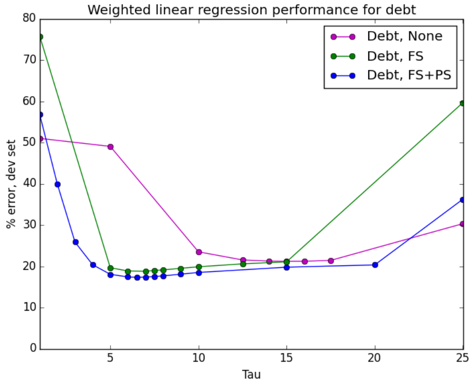

Figure 2: *τ values plotted against percent error for median debt.*

KNN Regression

We also used the non-parametric k-nearest-neighbors model in order to capture nonlinearities in prediction of debt and earnings, using imputation of privacy-suppressed values and the same data standardization technique used for weighted linear regression. The KNN algorithm predicts debt and earnings as a weighted combination of debt and earnings of an input's k nearest (defined here as Euclidean distance) neighbors. The weighting schemes tried were uniform weights and weights proportional to inverse distance.

Figure 3 shows the performance of KNN regression on the development set across values of k. Inverse-distance weighting outperformed uniform weighting, giving evidence that school with similar graduate earnings and debt are clustering in our feature space, but KNN with optimal k had higher error than the best weighted linear regression model.

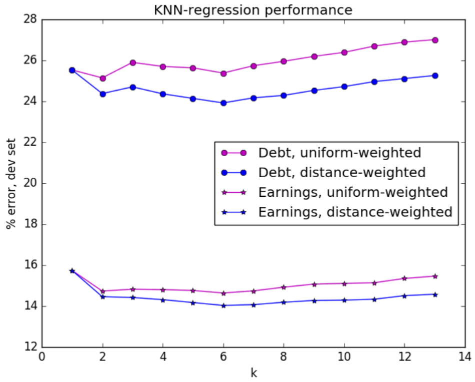

Figure 3: *k values plotted against percent error for median earnings and debt.*

Capturing Nonlinearities Among Features

Lastly, we explored using models that can automatically capture nonlinear relationships among the variables, in addition to nonlinearities between the variables and outputs.

First, we used a support vector machine with data standardization and feature selection to make predictions. We used the RBF kernel and L2-regularized L1-loss support vector regression; L2-regularized L2-loss support vector regression yielded similar results. We tuned our regularization term coefficients on the development set and found 0.000003 and 0.00000007 to be the optimal parameters for earnings and debt, respectively.

We also trained simple neural networks with a single hidden layer, using the previous feature selection and imputation for privacy-suppressed values [11]. A single hidden layer was chosen because there was insufficient training data (number of schools) to fit a model with more parameters without significant overfitting. The network is trained using the Levenberg-Marquadt algorithm for minimization with the logistic function as the activation function, and it uses a randomly held-out set from the training set as a validation set and ceases training when improvement on the held-out set plateaus.

The number of nodes in the neural network was tuned by examining the performance on the development set; results were mostly consistent for networks up to 10 nodes, after which the network suffered an overfitting problem. A hidden layer with 4 nodes performed optimally for debt with 19.36% error, and a hidden layer with 6 hidden nodes performed optimally for earnings with 11.74% error.

IV. Results

Tables 2 and 3 show the test set performance of the optimized (with respect to the development set) model from each class for earnings and debt. The primary error metrics were mean absolute percent error, which penalizes errors of different sizes and directions equally, and RMSE (root mean squared error), which penalizes larger deviations superlinearly. For the best model under both metrics, weighted linear regression, the R^2 measure between predicted and actual values in the test set was 0.9079 for earnings and 0.9221 for debt. Our absolute error is lower for debt than for earnings, but since the dollar amounts for debt are typically lower than those of earnings, we have a higher percentage error for debt prediction.

	Mean Percent Error	RMSE
Linear Regression	13.42%	5346
Weighted Lin. Reg.	9.60%	4002
KNN	14.11%	6509
SVM	23.80%	9488
Neural Networks	11.93%	4824

Table 2: *Error for earnings across all models.*

	Mean Percent Error	RMSE
Linear Regression	21.94%	3778
Weighted Lin. Reg.	17.08%	3078
KNN	26.60%	4951
SVM	38.47%	10264
Neural Networks	22.12%	3694

Table 3: *Error for debt across all models.*

V. Discussion

Overall, much of the variance in earnings and debt information was in fact captured by the static school data provided in College Scorecard. Our incremental model selection process showed that regression imputation of privacy-suppressed values improved overall performance. In addition, local weighting helped adapt linear regression to nonlinear relationships between school characteristics and graduate debt/earnings. The number of training examples is limited by the number of schools, but feature selection helped constrain the complexity of our models in this setting. In addition, test set performance was very similar to development set performance, so optimizing our model parameters through the development set did not lead to excessive overfitting.

Support vector regression did worse than all other models, even after optimization of regularization parameters. The test set performance was only marginally worse than errors for the training and development sets, so overfitting was not an issue. This indicates that learning decision boundaries in our kernelized feature space is not very helpful for the values we want to predict.

Several selected features relate to socioeconomic backgrounds of the student population. The College Scorecard data set included earning and debt data subdivided by background, but most of this data was privacy-suppressed. For future work, partnering with the U.S. Department of Education to gain access to this data could help provide more accurate or individualized estimates.

If we examine the predictions made by weighted linear regression for median earnings, approximately 40% of the test set schools had predictions within $1,000 of the true value, and almost 90% of schools had predictions within $5,000 of the true value, meaning that the function did well for the majority of schools. However, ten of the schools had absolute percent errors above 50%; in examining these schools, the majority only instructed specialized skills, e.g. cosmetology, massage therapy. Therefore, it seems like the current algorithm has trouble extending to trade schools, in which future debt and earnings may be best characterized by a different set of features than those emphasized by College Scorecard.

We also explored whether our best model generalized well outside of our training, development, and test sets by running it on the 2371 schools missing earnings or debt information. Though we have no benchmark to calculate accuracy for predictions on these schools, we qualitatively examined the schools with the highest predicted earnings. The Columbia College of Nursing had the highest predicted earnings; the rest of the top ten included two other health care-related schools, six law schools, and the Hawaii Technology Institute. Since health care-related schools dominated the top earnings schools list for our labeled data (schools which had earnings information already in the dataset), their presence in the predictions for the unlabeled data is expected. However, the labeled data had no law schools and law schools made up only 1.5% of the unlabeled data, indicating that our features were general enough to correctly predict high earnings for law school graduates.

With below 10% average error for earnings data, our best weighted linear regression model could be used to fill in gaps in the current College Scorecard data set, given a proper disclaimer.

References

[1] U.S. Department of Education. *College Scorecard*, 2015.

[2] The Washington Post. *Hundreds of Colleges Missing from Obama's College Scorecard*, 2015.

[3] Michael Kamp, Mario Boley, Thomas Gartner. *Beating Human Analysts in Nowcasting Corporate Earnings by using Publicly Available Stock Price and Correlation Features*, 2013.

[4] Center for Machine Learning and Intelligent Systems. *Census Income Data Set*, 1996.

[5] Dominic J. Brewer, Eric R. Eide, Ronald G. Ehrenberg. *Does It Pay to Attend an Elite Private College? Cross-Cohort Evidence on the Effects of College Type on Earnings*, 1999: The Journal of Human Resources.

[6] Estelle James, Nabeel Alsalam, Joseph C. Conaty, Duc-Le To. *College Quality and Future Earnings: Where Should You Send Your Child to College?*, 1989: The American Economic Review.

[7] Lewis C. Solmon. *The Definition of College Quality and Its Impact on Earnings*, 1975: The National Bureau of Economic Research.

[8] Benjamin M. Marlin. *Missing Data Problems in Machine Learning*, 2008: University of Toronto.

[9] Donald B. Rubin. *Multiple Imputation After 18+ Years*, 1996: Journal of the American Statistical Association.

[10] Machine Learning and Statistics Toolbox Release 2015b, The MathWorks, Inc., Natick, Massachusetts, United States.

[11] Neural Network Toolbox Release 2015b, The MathWorks, Inc., Natick, Massachusetts, United States.