Predicting Post-Collegiate Earnings and Debt

Monica Agrawal, Priya Ganesan, Keith Wyngarden

Stanford ENGINEERING Computer Science

Stanford University

Introduction

The U.S. Department of Education launched College Scorecard (https://collegescorecard.ed.gov) in September 2015 to gather more data on degree-granting institutions, the demographics of college students, and the status of alumni of these institutions.

Data Preprocessing

College Scorecard provides a public dataset of ~2000 metrics for 7805 degree-granting institutions. We chose the 2011 data set because it was the least sparse data set in the last five years.

I. 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 eliminated all features that had non-numerical or non-categorical values; unrelated and unhelpful features; and all features related to debt, earnings, and repayment.

II. Selecting Training and Test Examples

We removed all examples (schools) that were missing the values for our two label variables. From the remaining examples, we set aside 3500 for training, 1000 for development, and the remainder (~800) for testing.

III. Preprocessing Non-Standard Data Values

We turned categorial features into separate indicator features. We also replaced null values with 0s and created an extra indicator feature for each feature that contained null values.

Methodology

I. Handling Privacy-Suppressed Values

We tried two approaches for handling privacy-suppressed values in our data set:

- Remove all features with privacy-suppressed entries
- Statistical imputation using linear regression to predict privacy-suppressed values

II. Feature Selection

We ran sequential forward-based feature selection on 3500 training examples. Features were selected based on their mean-squared error using 10-fold cross-validation.

The top 5 features after statistical imputation are:

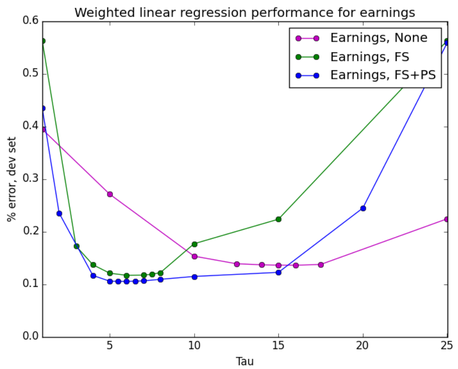
	Earnings	Debt
1	% students who are financially independent and have family incomes between $0-30,000	Predominant degree awarded
2	% degrees awarded in Personal And Culinary Services	Average net price for $0-$48,000 family income
3	Total share of enrollment of undergraduate degree-seeking students who are Asian	% 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

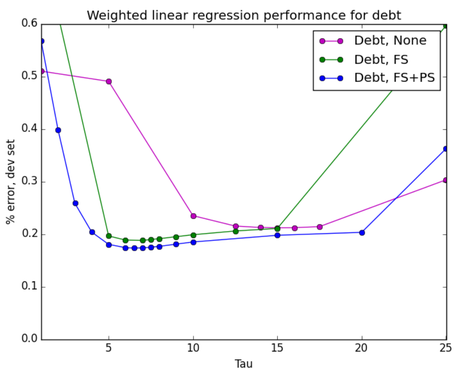
III. Linear Regression

We ran simple linear regression using both feature selection (FS) and statistical imputation (PS). We also did the same with weighted linear regression.

For weighted linear regression, we normalized all of our training data to have mean 0 and standard deviation 1. We used this weighting function:

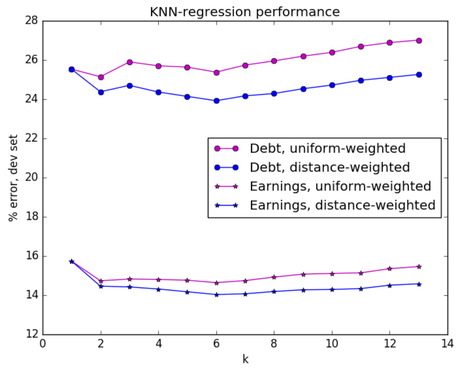
$$w^{(i)} = \exp\left(-\frac{\left\|x^{(i)} - x\right\|_2^2}{\tau^2}\right)$$





IV. KNN Regression

We also used the k-nearest-neighbors model in order to predict debt and earnings, using statistical imputation and the same normalization technique used for weighted linear regression.

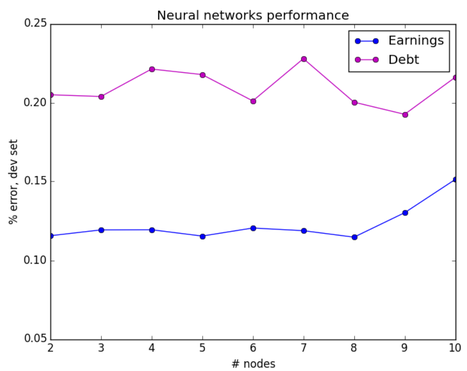


V. Support Vector Machines

We also used a support vector machine to make predictions, using normalization and feature selection. We used L2-regularized L1-loss support vector regression; L2-regularized L2-loss support vector regression yielded similar results. We tuned our regularization parameters on the development set and found 0.000003 and 0.00000007 to be the optimal parameters for earnings and debt, respectively.

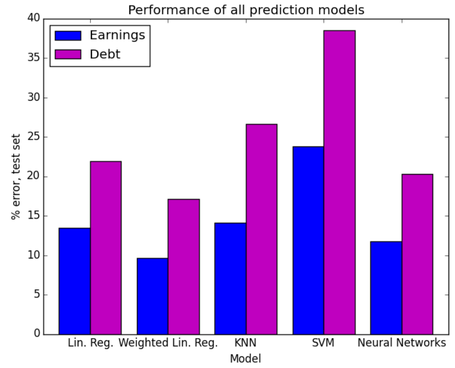
VI. Neural Networks

We used simple neural networks with a singular hidden layer, using the previous feature selection and imputation for privacy-suppressed values. A single hidden layer was chosen since there was insufficient training data to fit a model with more parameters.



Results

The following graph shows the accuracy of each model we tested. Each model has been optimized for the highest accuracy rate, based on the development set.



For the best model, weighted linear regression, the R^2 between predicted and actual values in the test set was 0.9079 for earnings and 0.9221 for debt.

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 iterative process showed that our statistical imputation for the missing values was sufficient to improve 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.

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

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.