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.

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 categorical 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.

IV. 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:

\[ w(i) = \exp(-\|\mathbf{x}_i - \mathbf{\bar{x}}\|^2) \]

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 0.0000003 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 because there was insufficient training data to fit a model with more parameters.