

# Method: Predictors for Institution Graduate Loan Default Rate

## Background

**What makes a certain college a good investment?** Is it producing graduates with less debt? Is it producing graduates who make more money 10 years after finishing? Or is it something less quantifiable, like guiding students into discovering their vocation or completing meaningful coursework?

The **U.S. Department of Education** quantifies college performance on a number of dimensions. **College Scorecard** is an interactive tool designed to help students choose a college based on government data about cost, financial aid, retention, completion and programs. The **data**, spanning 1996-2013, is publicly available in .csv format.

## Summary of Findings

A linear regression model fit to 4 years of observations (2010-2013; n=6080; training set 80%) found higher completion rate, faculty pay, institutional expense per student and admission rate predicted lower default rates, while public control, more Pell Grants, higher cost, larger enrollment, more debt, more federal loans and more full time faculty predicted higher default rates.

## Data & Method

The model was trained on 4 years of data, 2010-2013, with 20% of the data held out as a validation set. Only four-year public and private non-profit colleges and universities with at least 500 students were included in these figures (1666 schools).

### Dependent Variable: Default Rate

While the dataset includes more than 1700 variables, I decided to focus on the **default rate (3 year cohort)** as the best measure of a college's value. It combines debt with post-graduate earnings into one feature, since a student that defaults on their student loans is clearly either buried in debt, not making high enough salary, or both. The median graduate debt dimension in the dataset wasn't viable since it only includes federal loan debt (which caps at \$27,000) and the post-graduate earnings dimension was sparsely populated and therefore not ideal.

Default rate had a right skewed distribution, which I remedied at least partially by taking its square root. See item 1 in the appendix for a comparison of distributions.

### Independent Variable Selection

I cherry-picked about 30 variables out of the dataset that I thought would have the greatest effect on default rate, including factors related to retention, completion, student debt, financial aid and cost. A large percentage of the variables in the dataset were too specific to have much weight - things like "percent of dependent students who died within 3 years at original institution" and "percent of high-income (above \$75,000 in nominal family income) students withdrawn from original institution within 8 years." Additionally, a lot of variables were missing 90-100% of values and thus had to be thrown out (e.g., "percent of students who submitted a FAFSA to at least one college").

Variable	Description
Type	Public or Private Non-Profit
Adm.Rate	Admission Rate
Enrollment	Size of Enrollment

Variable	Description
Avg.Cost	Average Cost
Inst.Expense	Institutional Expense per Student
Avg.Faculty.Pay	Average Faculty Pay
Percent.FTE.Faculty	Percent Full Time Faculty
Pell.Grant.Pct	Percentage of Students with Pell Grants
Completion	Completion Rate (6 years)
Pct.Fed.Loan	Percentage of Students with Federal Loans
Md.Debt.n30	Median Debt of Graduates (suppressed for n<30)

### Missing Value Imputation

I only selected variables with less than 25% missing values. Because the Default.2Y and Default.3Y variables were missing values in alternate observations (rarely both missing in an obs.), I was able to combine the 2 year and 3 year default rates and impute the rest. I then used the **mice** package (Multivariate Imputations by Chained Equations) to impute the remaining missing values using the pmm or predictive mean matching method. See item 2 in the appendix for a breakdown of missing values by variable.

### Model

#### Model Selection

I fit a model with all variables, which found all but institutional expense to be significant, providing an adjusted R-squared value of 0.588,  $F = 631.9$ ,  $p < 2.2e-16$ . Stepwise selection recommended keeping all variables. See item 3 in the appendix for a summary of the model; item 4 for the diagnostic plots and item 5 for the predictions vs. actual observations of the test set.

#### Diagnostics

The diagnostic plots are mostly normal save for the presence schools with default rates of 0 (29 total, a good chunk of them seminaries or bible schools, but also included California Institute of Technology, Claremont McKenna College, Harvey Mudd College and Vassar College).

#### Validation

Comparing the model's predicted values to the actual values in the test set produced an R-squared value of 60.9%.

### Interpretation

Since the dependent variable was log transformed, the coefficients can be interpreted in a slightly different way. Instead of the unit increase in the DV per one unit increase in each variable, the exponentiated (exp) coefficient represents the percent change in the DV per unit change in each variable. Those values are shown below, where log.estimate is the coefficient, exp.estimate is  $\exp(\text{coef})$  and exp.estimate.minus1 is  $\exp(\text{coef}) - 1$ , or the percent change. For example, we can expect a 0.012% increase in default rate with each percent increase in Pell Grants at an institution.

The factors in predicting a lower default rate for a school are completion rate (0.0177% decrease in default rate per one percent increase in completion rate), average faculty pay (0.0104% decrease in default rate per \$1000 increase in salary) institutional expense per student (0.0034% decrease in default rate per \$1000 increase in expense) and admission rate (0.0008% decrease in default rate per one percent increase in admission rate).

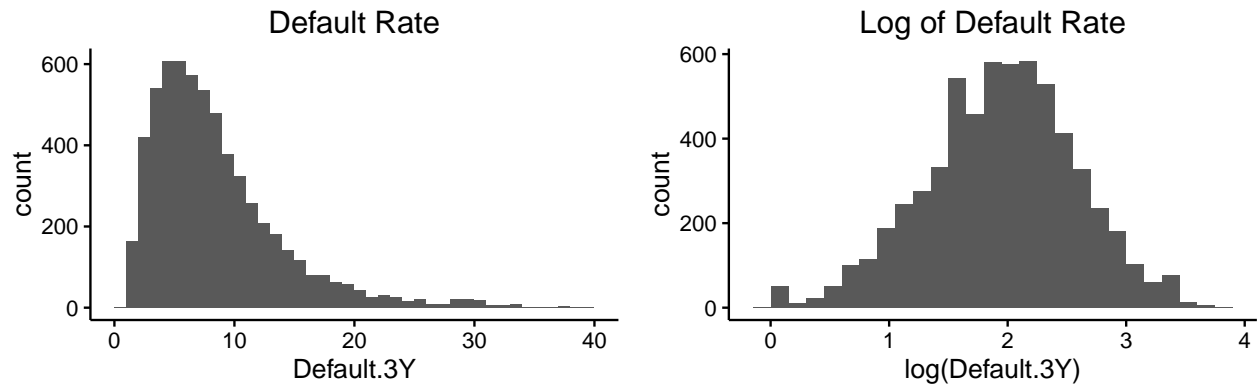
Factors predicting higher default rates are public control of the school (0.1789% increase in default rate), Pell Grant percentage (0.0121% increase in default rate per one percent increase in Pell Grants), average cost of attendance (0.0066% increase per \$1000 increase in cost), enrollment size (0.0057% increase per increase of 1000 students), median debt (0.004% increase per \$1000 increase in debt), percentage of students with

federal loans (0.0023% increase per \$1000 increase) and percent full time faculty (0.0012% increase per one percent increase).

coef	log.estimate	exp.estimate	exp.estimate.minus1
(Intercept)	2.0281995	7.6003892	6.6003892
TypePublic	0.1646620	1.1789946	0.1789946
Pell.Grant.Pct	0.0120842	1.0121575	0.0121575
Avg.Cost	0.0065928	1.0066145	0.0066145
Enrollment	0.0056889	1.0057051	0.0057051
Md.Debt.n30	0.0040559	1.0040641	0.0040641
Pct.Fed.Loan	0.0023820	1.0023849	0.0023849
Percent.FTE.Faculty	0.0012166	1.0012173	0.0012173
Adm.Rate	-0.0008543	0.9991461	-0.0008539
Inst.Expense	-0.0034614	0.9965446	-0.0034554
Avg.Faculty.Pay	-0.0104773	0.9895774	-0.0104226
Completion	-0.0179466	0.9822135	-0.0177865

## Appendix

### 1. Default rate transformation



### 2. Missing values

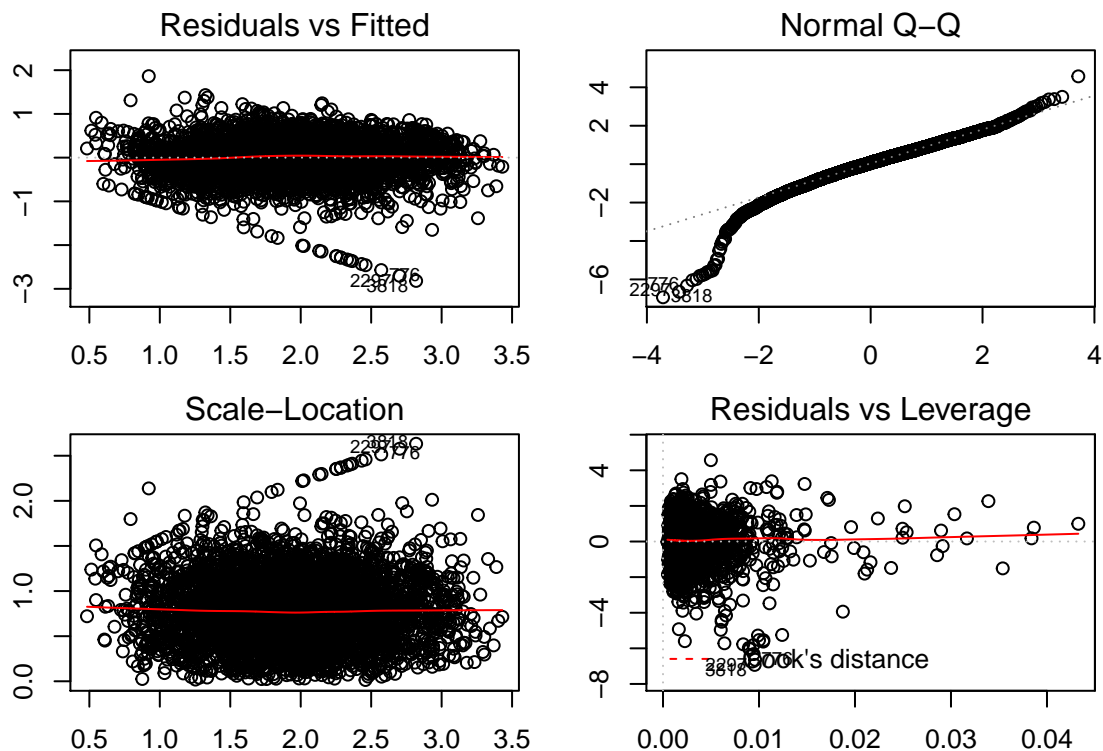
Variable	Imputed
Admission Rate	529
Completion Rate	165
Median Debt	25
Institutional Expense per Student	31
Avg. Faculty Pay	31
Percent Full Time Faculty	15
Default Rate	8
Percent of Students with Pell Grants	4

### 3. Summary of linear model

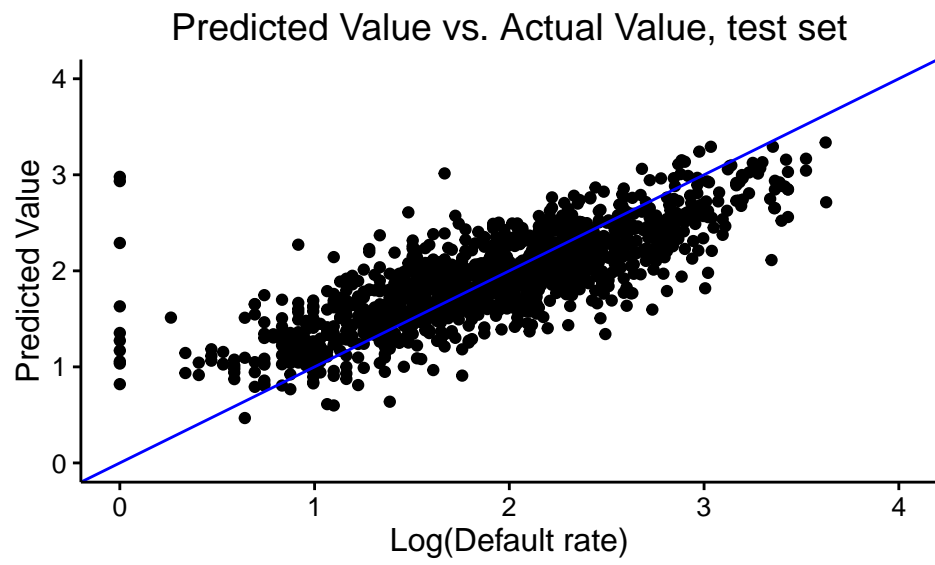
```
##
## Call:
## lm(formula = log(Default.3Y) ~ Type + Adm.Rate + Enrollment +
##     Avg.Cost + Inst.Expense + Avg.Faculty.Pay + Percent.FTE.Faculty +
##     Pell.Grant.Pct + Completion + Pct.Fed.Loan + Md.Debt.n30,
##     data = train)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81899 -0.22724  0.02664  0.25882  1.86326
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.0281995   0.0687595   29.497 < 2e-16 ***
## TypePublic      0.1646620   0.0218595    7.533 5.89e-14 ***
## Adm.Rate       -0.0008543   0.0003719   -2.297 0.021658 *
## Enrollment      0.0056889   0.0011236    5.063 4.28e-07 ***
## Avg.Cost        0.0065928   0.0010763    6.125 9.77e-10 ***
## Inst.Expense    -0.0034614   0.0009275   -3.732 0.000192 ***
## Avg.Faculty.Pay -0.0104773   0.0049752   -2.106 0.035264 *
## Percent.FTE.Faculty 0.0012166 0.0002466    4.933 8.36e-07 ***
## Pell.Grant.Pct  0.0120842   0.0005523   21.882 < 2e-16 ***
## Completion     -0.0179466   0.0005598  -32.058 < 2e-16 ***
## Pct.Fed.Loan    0.0023820   0.0004609    5.169 2.45e-07 ***
## Md.Debt.n30     0.0040559   0.0019756    2.053 0.040128 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4087 on 4852 degrees of freedom
## Multiple R-squared:  0.5889, Adjusted R-squared:  0.588
## F-statistic: 631.9 on 11 and 4852 DF,  p-value: < 2.2e-16
```

#### 4. Linear model diagnostics



#### 5. Linear model validation



6. R script