Predicting Graduation Rates

Fred Berendse, Ph.D.

/github.com/fred-b-berendse

/in/fred-b-berendse

fred.b.berendse@gmail.com

Background

Education is touted as the great equalizer that opens doors for the disenfranchised. For that reason, it is imperative to examine how our educational institutions, from pre-K to post-secondary, are facilitating (or hampering) the path toward equity.

Objectives

- Determine the primary factors influencing post-secondary graduation rates of majority and minority groups.
- Utilize regression models and predict graduation rates based on institutional metrics.

Data

The IPEDS database consists of annual survey data collected from post-secondary institutions. Targets are bachelor's degrees completed within 6 years in 2016-17. Features include institution characteristics, admissions data, and student financial aid data. There are 682 institutions with all considered features. Variance inflation factors above 5 were eliminated, resulting in the feature set below.

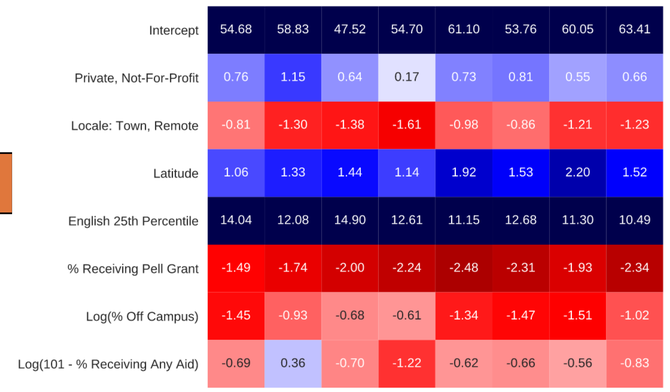
Feature Description	Feature Description
is private, not-for-profit	percent of admissions enrolled
highest degree offered	percent of enrolled students attending full time
is a HBCU	average of normalized ACT English/SAT Verbal 25th percentile
locale: city/suburb/town/rural + large/midsize/small/fringe/distance/remote	percent of students awarded financial aid
institution size	percent of students awarded Pell Grants
longitude of institution	average net price for students awarded aid
latitutde of institution	percent of students living with family off campus
percent of applicants admitted	percent of students living off campus (not with family)

Models

I. Linear Regression with Lasso

Below are the most important coefficients after Lasso regularization:

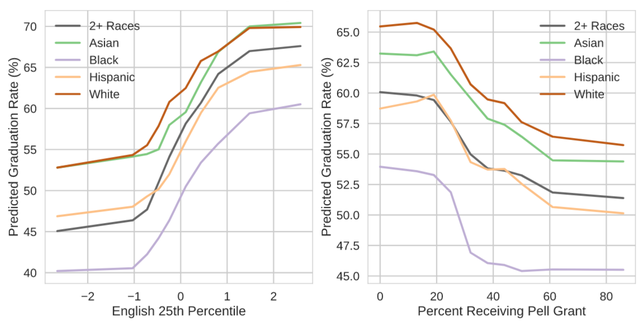
Lasso Regression Coefficients



II. Random Forest Regression

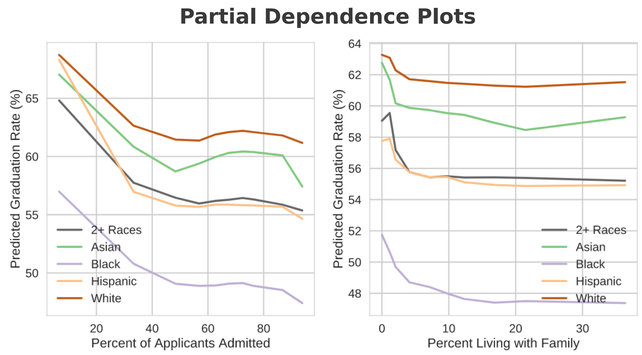
Five-fold cross validation provided a best random forest model with 160 trees. The four most important features affect predicted rates by as much as 20%.

Partial Dependence Plots



Models

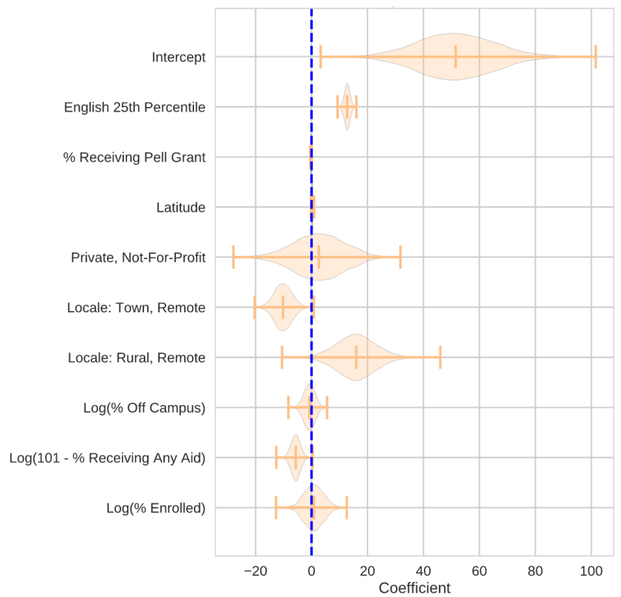
Partial Dependence Plots



III. Markov Chain Monte Carlo

A MCMC regression utilizing the important features from the Lasso regression model was performed. All fits obtained convergence (*i.e.* Gelman-Rubin statistic near 1.00). All racial/aid target groups obtained similar coefficient distributions.

Coefficients: Hispanic Graduation Rate

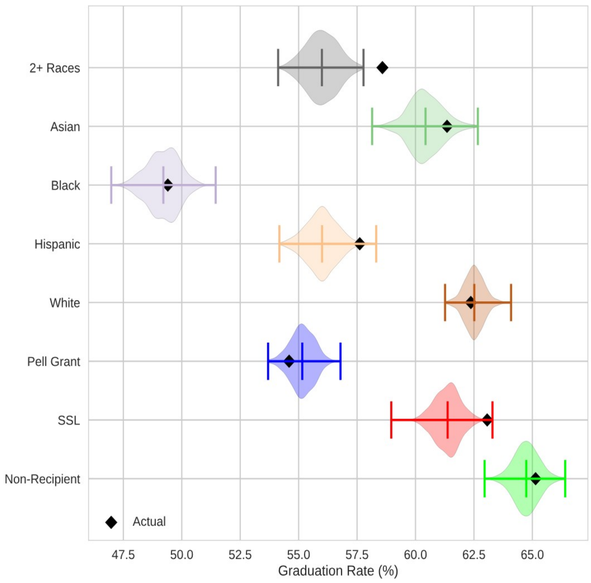


Tech Stack

python · scikit-learn · PyMC3 · NumPy · matplotlib · amazon web services · pandas · Flask · PostgreSQL

Models

Predicted Mean Graduation Rate



Model Comparison

All three models have similar RMSEs with random forest holding a slight edge in performance.

Target	OLS w/Lasso	Random Forest	MCMC
2+ Races	15.5%	14.9%	15.1%
Asian	16.2%	15.5%	15.7%
Black	15.9%	16.0%	16.4%
Hispanic	18.4%	17.1%	17.4%
White	9.4%	9.5%	9.7%
Pell Grant	11.6%	11.3%	11.4%
SSL	14.8%	14.7%	14.9%
Non-recipient	13.9%	14.1%	14.2%

Conclusions

All three models agree that ACT/SAT acceptance benchmarks are positively correlated with graduation rates for all groups. Percentage of students receiving a Pell Grant is negatively correlated with graduation rate in the Lasso and random forest models.