



# 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

is private, not-for-profit

highest degree offered

is a HBCU

locale: city/suburb/town/rural + large/midsize/small/fringe/distance/remote

institution size

longitude of institution

latitutde of institution

percent of applicants admitted

## Models

### I. Linear Regression with Lasso

Below are the most important coefficients after Lasso regularization:

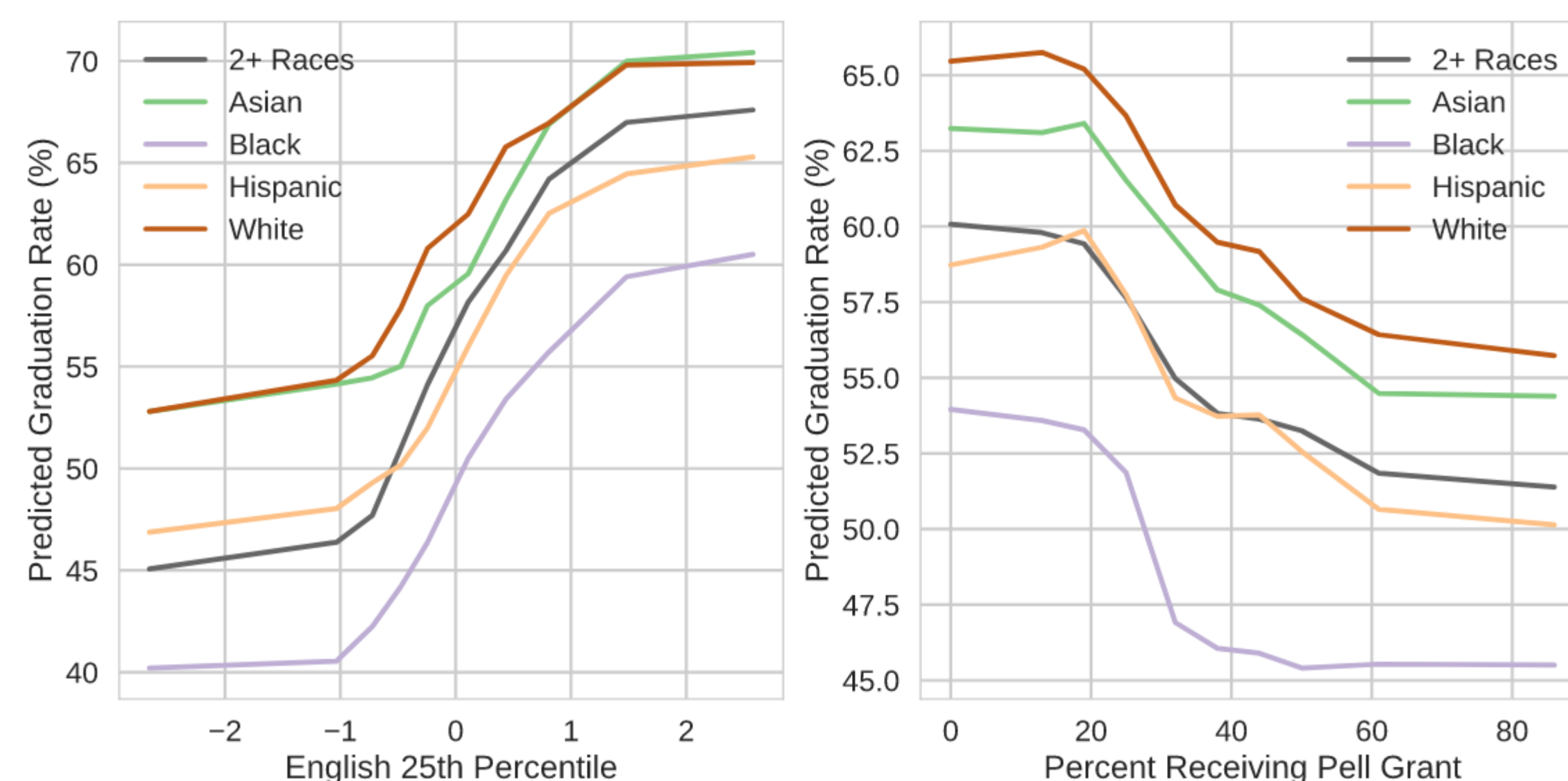
Lasso Regression Coefficients

Intercept	54.68	58.83	47.52	54.70	61.10	53.76	60.05	63.41
Private, Not-For-Profit	0.76	1.15	0.64	0.17	0.73	0.81	0.55	0.66
Locale: Town, Remote	-0.81	-1.30	-1.38	-1.61	-0.98	-0.86	-1.21	-1.23
Latitude	1.06	1.33	1.44	1.14	1.92	1.53	2.20	1.52
English 25th Percentile	14.04	12.08	14.90	12.61	11.15	12.68	11.30	10.49
% Receiving Pell Grant	-1.49	-1.74	-2.00	-2.24	-2.48	-2.31	-1.93	-2.34
Log(% Off Campus)	-1.45	-0.93	-0.68	-0.61	-1.34	-1.47	-1.51	-1.02
Log(101 - % Receiving Any Aid)	-0.69	0.36	-0.70	-1.22	-0.62	-0.66	-0.56	-0.83
	2+ Races	Asian	Black	Hispanic	White	Pell Grant	SSL	Non-Recipient

### II. Random Forest Regression

Partial dependence plots for the four most important features are shown. Higher test benchmarks, fewer Pell Grants, and more selective admissions correlate with higher graduation rates.

Partial Dependence Plots



### Feature Description

percent of admissions enrolled

percent of enrolled students attending full time

average of normalized ACT English/SAT Verbal 25th percentile

percent of students awarded financial aid

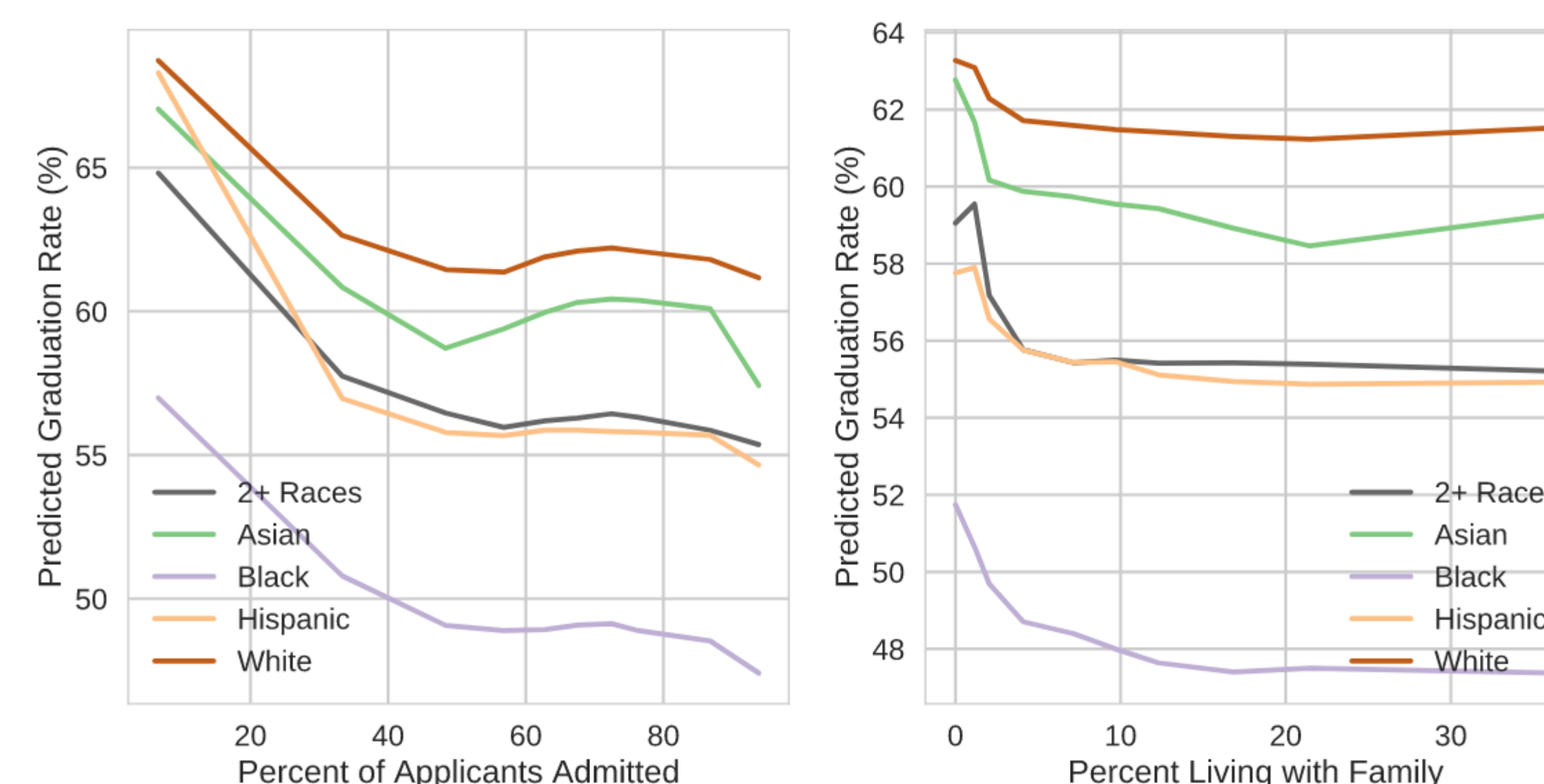
percent of students awarded Pell Grants

average net price for students awarded aid

percent of students living with family off campus

percent of students living off campus (not with family)

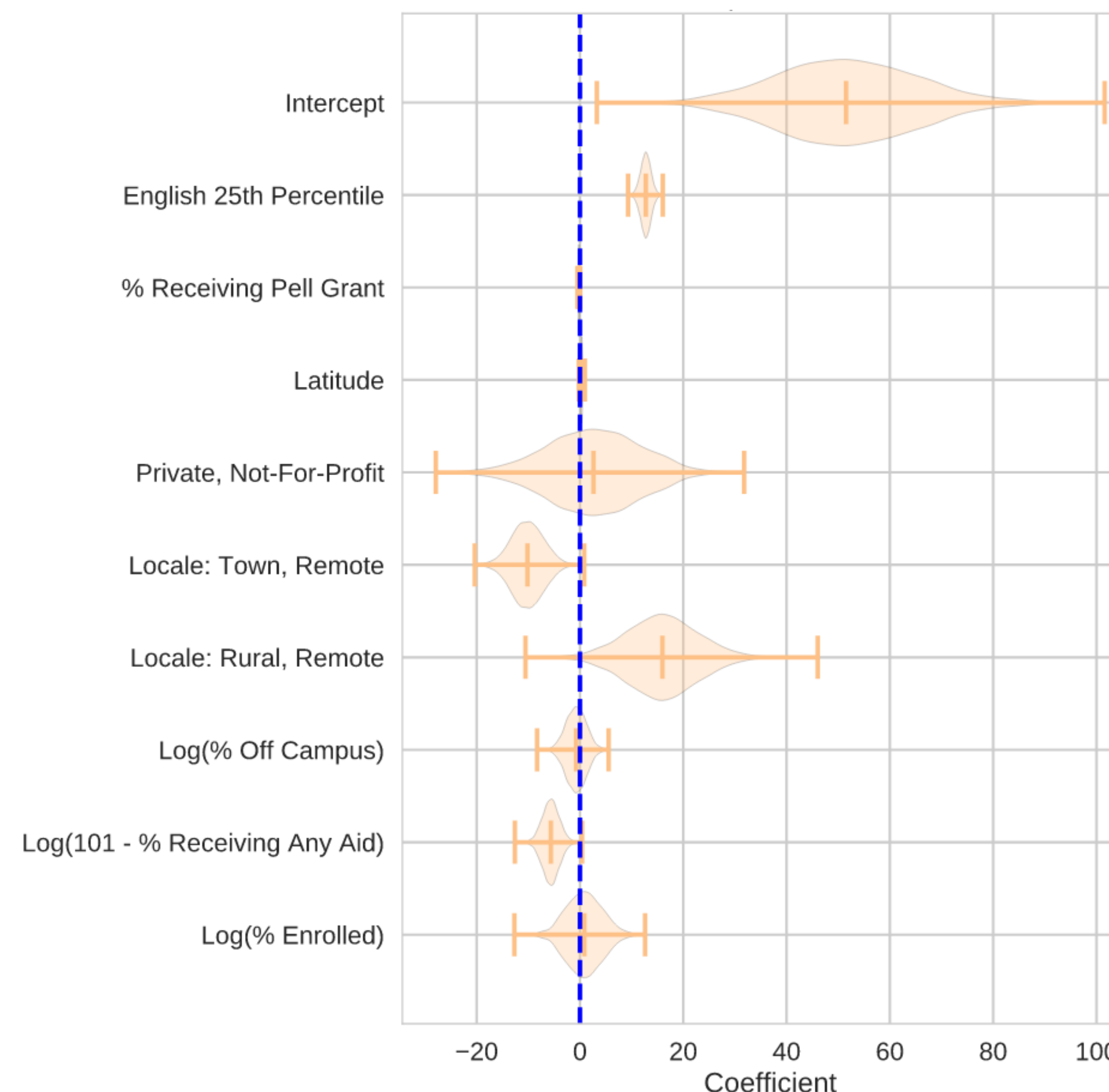
Partial Dependence Plots



### 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

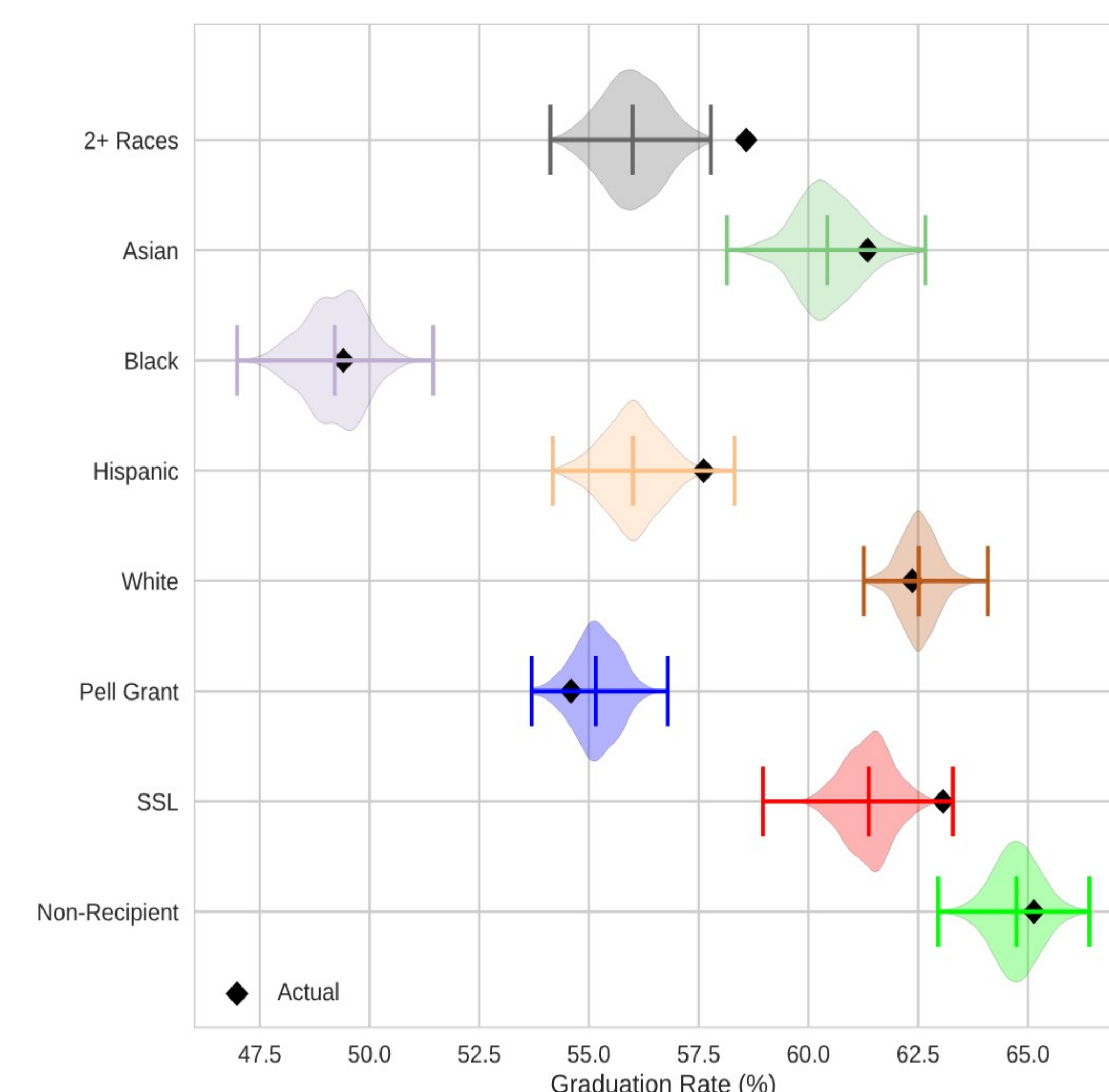


## Tech Stack



## 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.