Predicting Graduation Rates

Fred Berendse, Ph.D.

/github.com/fred-b-berendse

in /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 necessary 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 institutional, admissions, and student financial aid data. There were 682 institutions with all considered features. Variance inflation factors above 5 were eliminated, resulting in the feature set below.

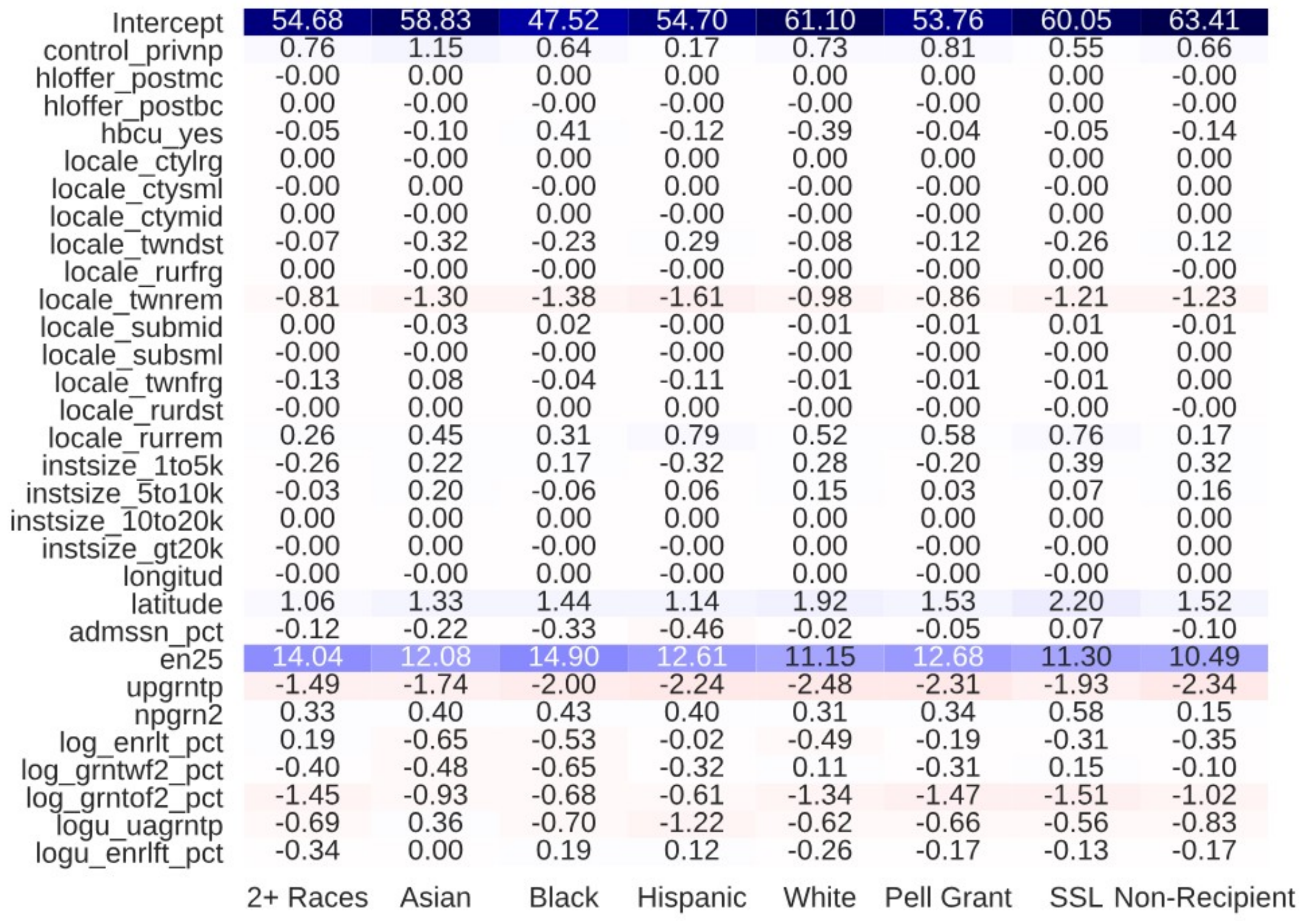
Feature	Description
control_privnp	is private, not-for-profit
hloffer	highest degree offered
hbcu_yes	is a HBCU
locale	city/suburb/town/rural + large/midsize/small/fringe/distance/remote
instsize	institution size
longitud	longitude of institution
latitude	latitutde of institution
admssn_pct	percent of applicants admitted
enrlt_pct	percent of admissions enrolled
enrltft_pct	percent of enrolled students attending full time
en25	average of normalized ACT English/SAT Verbal 25th percentile
uagrntp	percent of students awarded financial aid
upgrntp	percent of students awarded Pell Grants
npgrn2	average net price for students awarded aid
grntwf2_pct	percent of students living with family off campus
grntof2_pct	percent of students living off campus (not with family)

Models

I. Linear Regression with Lasso

Five-fold cross validation was used to optimize the Lasso shrinkage parameter to 0.08.

Lasso Regression Coefficients

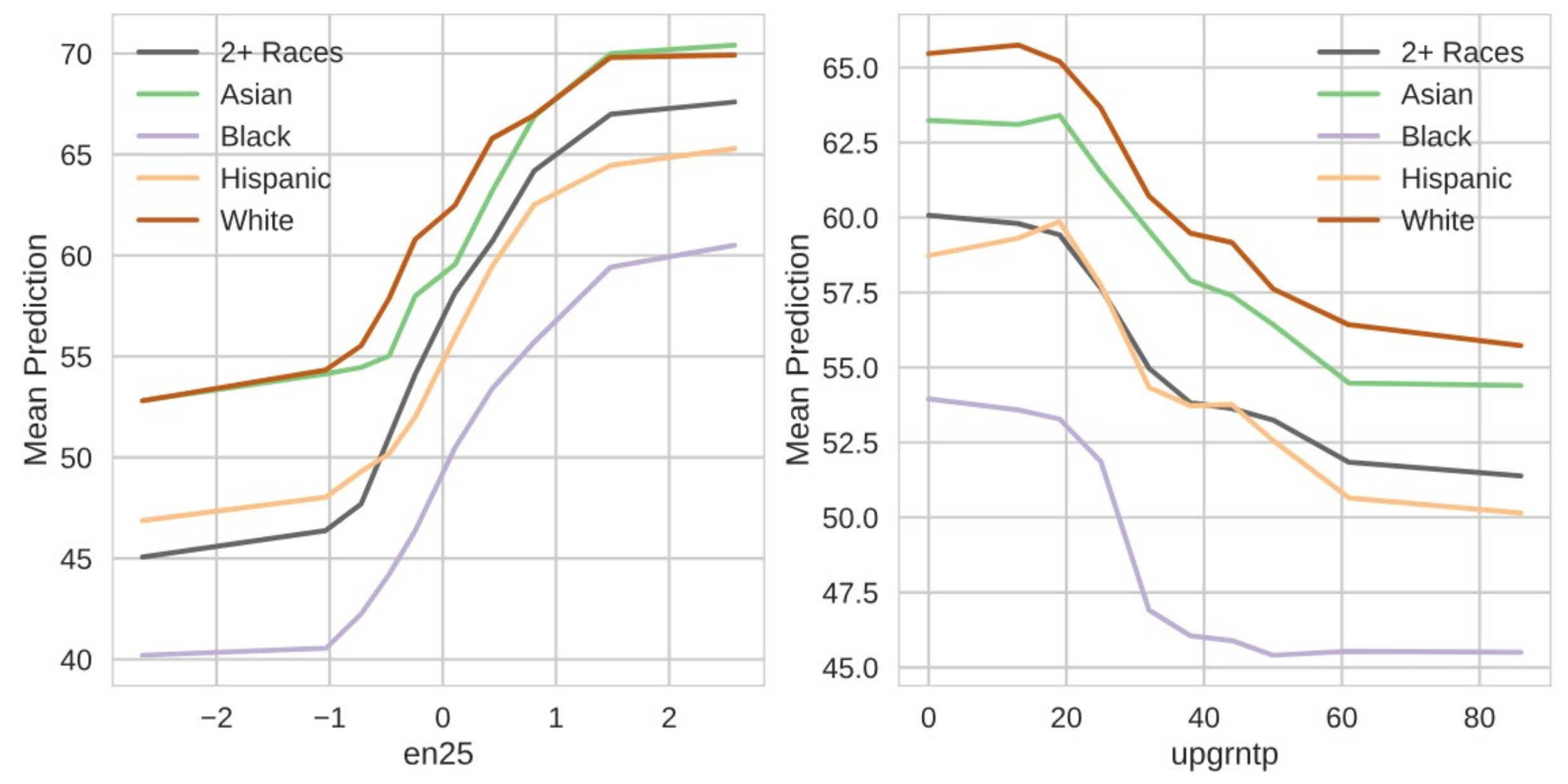


	2+ Races	Asian	Black	Hispanic	White	Pell Grant	SSL	Non-Recipient
Intercept	54.68	58.83	47.52	54.70	61.10	53.76	60.05	63.41
control_privnp	0.76	1.15	0.64	0.17	0.73	0.81	0.55	0.66
hloffer_postmc	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00
hloffer_postbc	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00
hbcu_yes	-0.05	-0.10	0.41	-0.12	-0.39	-0.04	-0.05	-0.14
locale_ctylrg	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00
locale_ctysml	-0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00
locale_ctymid	0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00
locale_twndst	-0.07	-0.32	-0.23	0.29	-0.08	-0.12	-0.26	0.12
locale_rurfrg	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00
locale_twnrem	-0.81	-1.30	-1.38	-1.61	-0.98	-0.86	-1.21	-1.23
locale_submid	0.00	-0.03	0.02	-0.00	-0.01	-0.01	0.01	-0.01
locale_subsml	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00
locale_twnfrg	-0.13	0.08	-0.04	-0.11	-0.01	-0.01	-0.01	0.00
locale_rurdst	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00
locale_rurrem	0.26	0.45	0.31	0.79	0.52	0.58	0.76	0.17
instsize_1to5k	-0.26	0.22	0.17	-0.32	0.28	-0.20	0.39	0.32
instsize_5to10k	-0.03	0.20	-0.06	0.06	0.15	0.03	0.07	0.16
instsize_10to20k	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
instsize_gt20k	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00
longitud	-0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00
latitude	1.06	1.33	1.44	1.14	1.92	1.53	2.20	1.52
admssn_pct	-0.12	-0.22	-0.33	-0.46	-0.02	-0.05	0.07	-0.10
en25	14.04	12.08	14.90	12.61	11.15	12.68	11.30	10.49
upgrntp	-1.49	-1.74	-2.00	-2.24	-2.48	-2.31	-1.93	-2.34
npgrn2	0.33	0.40	0.43	0.40	0.31	0.34	0.58	0.15
log_enrlt_pct	0.19	-0.65	-0.53	-0.02	-0.49	-0.19	-0.31	-0.35
log_grntwf2_pct	-0.40	-0.48	-0.65	-0.32	0.11	-0.31	0.15	-0.10
log_grntof2_pct	-1.45	-0.93	-0.68	-0.61	-1.34	-1.47	-1.51	-1.02
logu_uagrntp	-0.69	0.36	-0.70	-1.22	-0.62	-0.66	-0.56	-0.83
logu_enrlft_pct	-0.34	0.00	0.19	0.12	-0.26	-0.17	-0.13	-0.17

II. Random Forest Regression

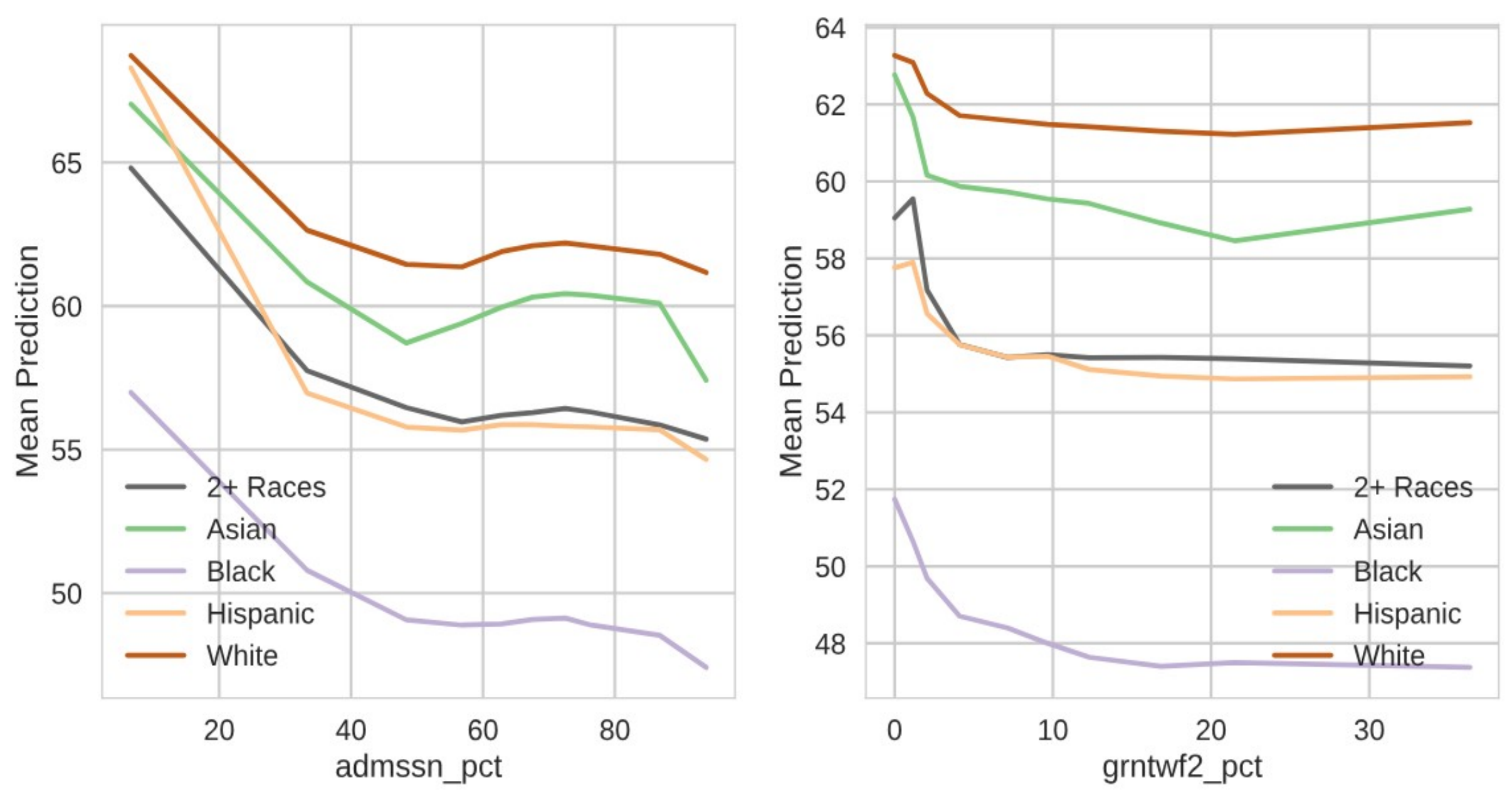
Five-fold cross validation provided a best RF model with 160 trees, a MSE split criterion, 2 or more samples per split, and considered $\sqrt{n}$ features per split.

Partial Dependence Plots



Models

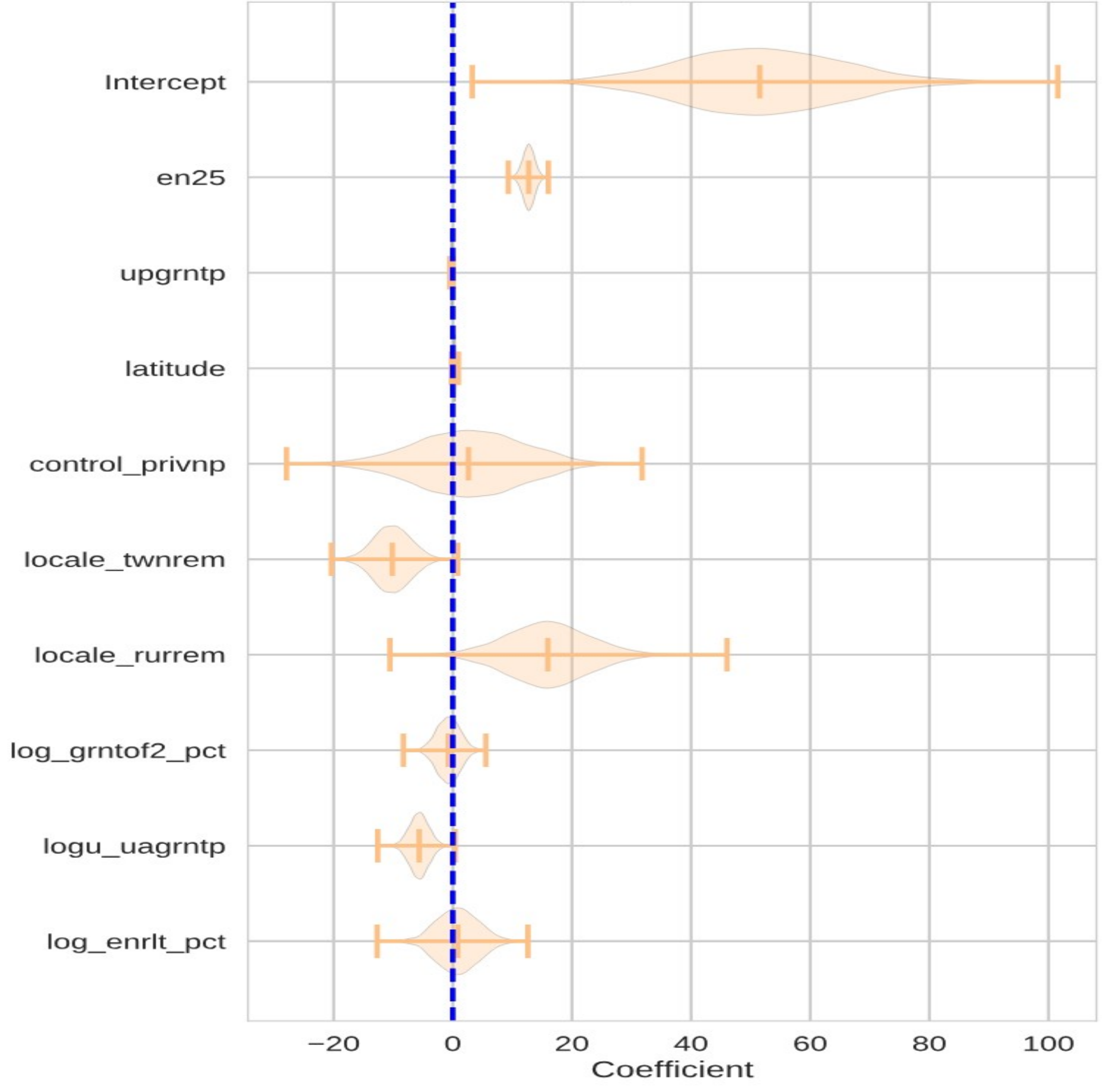
Partial Dependence Plots



III. Markov-Chain Monte Carlo

A MCMC regression utilizing a subset of features from the Lasso regression model was performed. The model utilizes 2000 draws plus a burn-in of 500 draws on 4 chains. 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

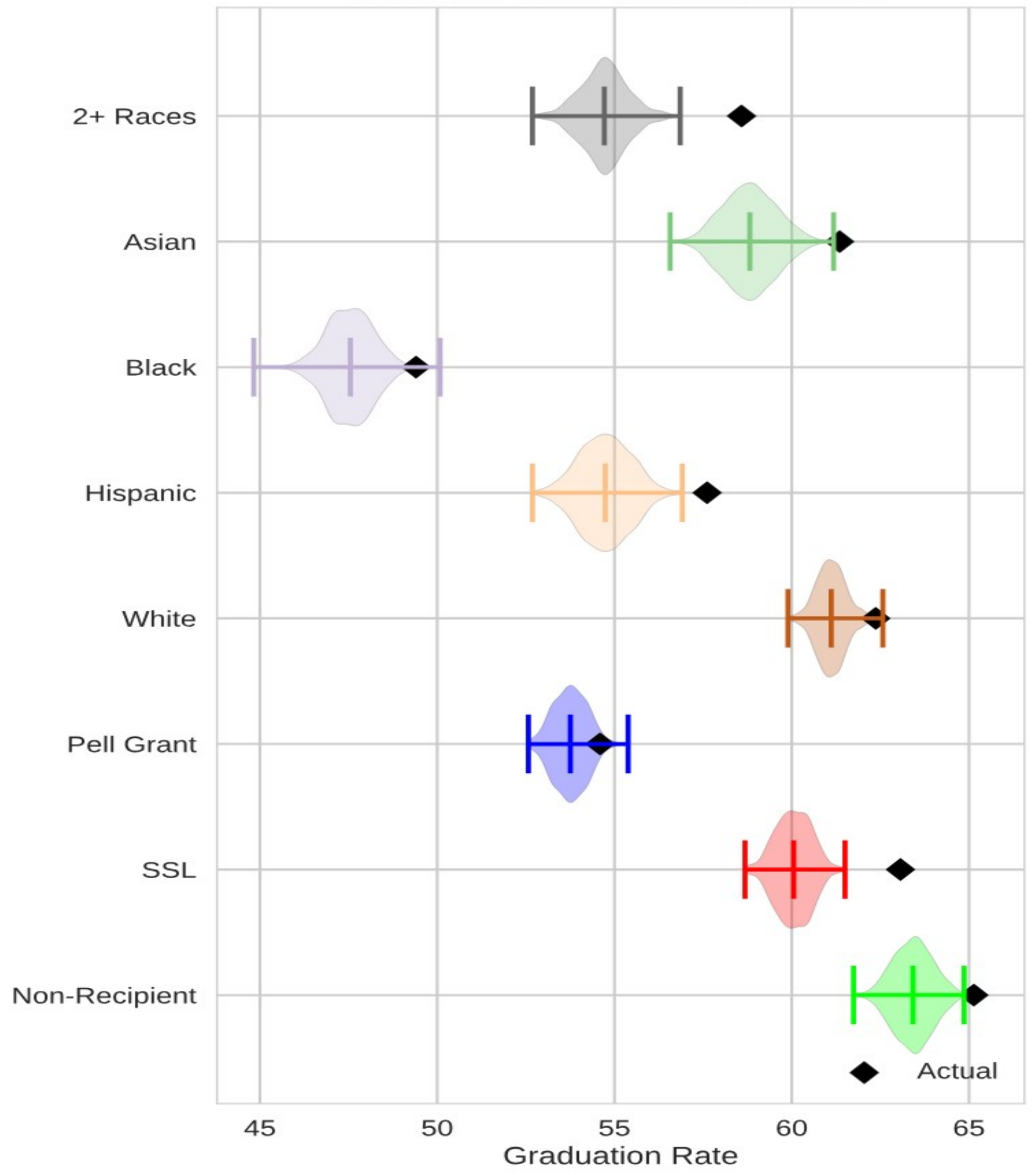
pandas

Flask

PostgreSQL

Models

Mean Predicted 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	9.71
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 on the most important features consistent across all race and aid status groups:

- English 25th acceptance percentile
- Percent of students awarded Pell grant
- Percent applicants admitted