Predicting Graduation Rates

Can one accurately predict an institution’s graduation rate for minority and low-socioeconomic-status students based on an institution’s features?

The IPEDS Dataset

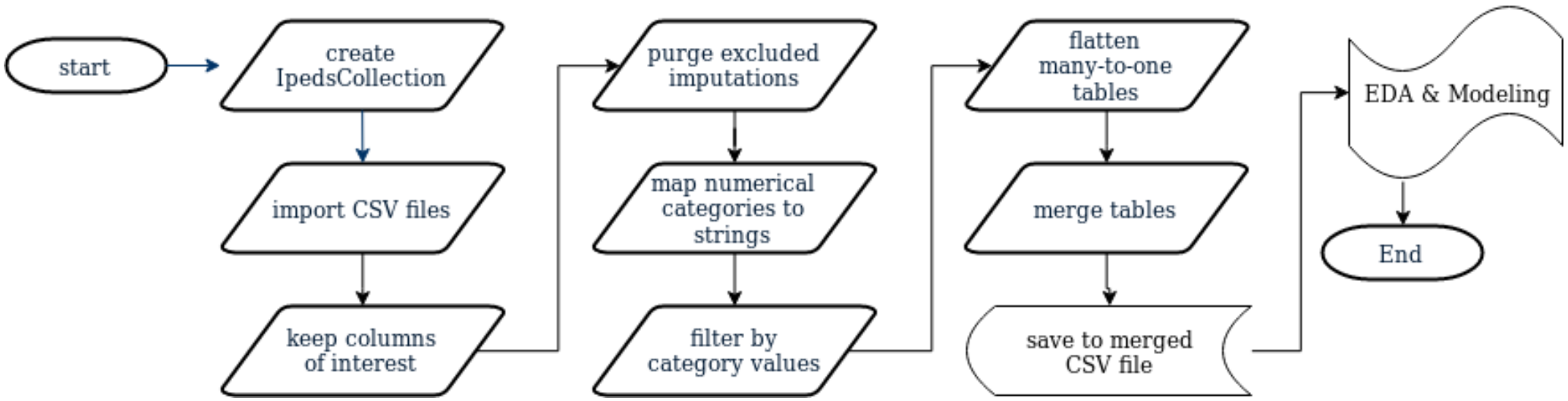
- Collected by the US Department of Education annually
- Over 7000 post-secondary institutions

Available Data	Provisional Release	Final Release
Institutional Characteristics (IC)	2018-19	2008-09 to 2017-18
Pricing and Tuition (IC)	2018-19	2008-09 to 2013-14
Admissions (ADM)	2017-18	2008-09 to 2016-17
Completions (C)	2017-18	2005-06 to 2016-17
12-month Enrollment (E12)	2017-18	2005-06 to 2016-17
Fall Enrollment (EF)	2017	2006 to 2016
Student Financial Aid (SFA)	2016-17	2005-06 to 2015-16
Graduation Rates (GR)	2017	2006 to 2016
Outcome Measures (OM)	2017	2015,2016
Finance (F)	2016-17	2005-06 to 2015-16
Human Resources (HR)	2017-18	2006-07 to 2016-17
Academic Libraries (AL)	2016-17	2014-15,2015-16

The IPEDS Dataset

Table	Rows	Full Rows	Description
HD2017	7153	7153	Institutional characteristics: name, location, locale, highest deg. awarded, etc.
ADM2017	2075	920	Application, admission and enrollment data including SAT/ACT percentiles
GR2017	54714	49981	Number of bachelor’s degree completions in 150% normal time by gender and race/ethnicity
GR2017_PELL_SSL	9116	5557	Number of bachelor's degree completions in 150% normal time by Pell Grant recipients, Subsidized Stafford Loan recipients, and non-recipients
SSL2017	6394	0	Number of students paying in-state/out-of-state tuition and receiving grant/scholarship aid

Data Cleaning Process



Data Cleaning Challenges

- Flattening many-to-one tables before merging
- Merging SAT and ACT benchmark percentiles into one score
- Imputing graduation rates for groups not reported

Data Cleaning Challenges

Laplace Smoothing

$$\frac{p_g + a * \sum p_g}{n_g + a * \sum n_g} \qquad a = 0.01$$

p_g

	gr2mort	graiant	grasiat	grbkaat	grhispt	grnhpit	grwhitt
0	2.0	0.0	0.0	195.0	1.0	0.0	3.0
1	32.0	2.0	76.0	180.0	20.0	1.0	503.0
2	0.0	0.0	0.0	0.0	1.0	0.0	0.0

n_g

	gr2mort	graiant	grasiat	grbkaat	grhispt	grnhpit	grwhitt
0	6.0	1.0	2.0	807.0	6.0	0.0	16.0
1	72.0	4.0	108.0	381.0	36.0	1.0	941.0
2	0.0	0.0	0.0	2.0	1.0	0.0	2.0

Without Smoothing

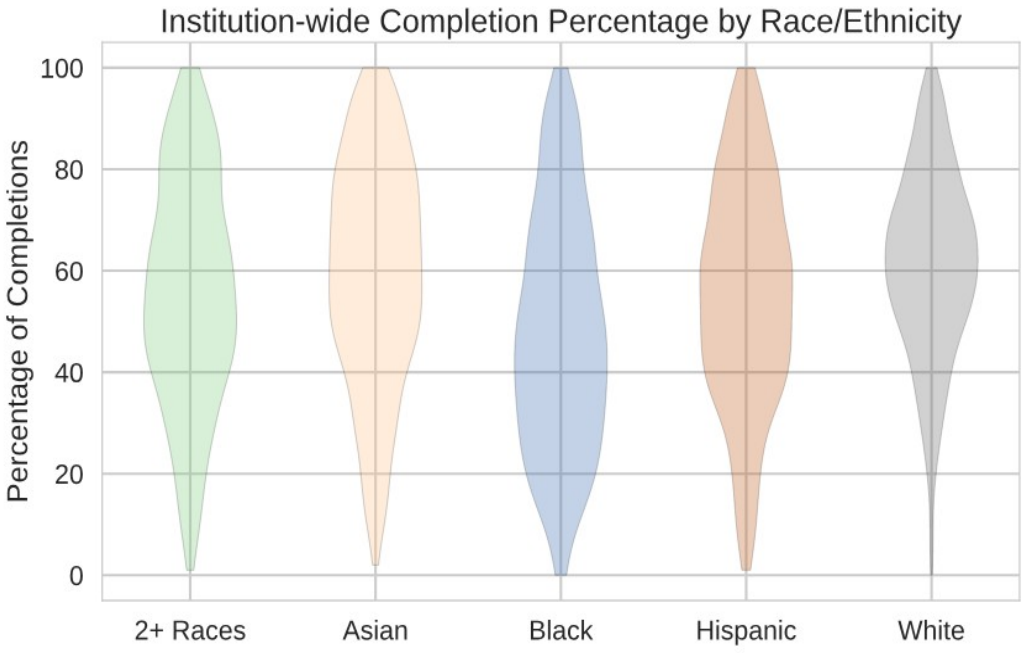
	gr2mort	graiant	grasiat	grbkaat	grhispt	grnhpit	grwhitt
0	33.0	0.0	0.0	24.0	17.0	NaN	19.0
1	44.0	50.0	70.0	47.0	56.0	100.0	53.0
2	NaN	NaN	NaN	0.0	100.0	NaN	0.0

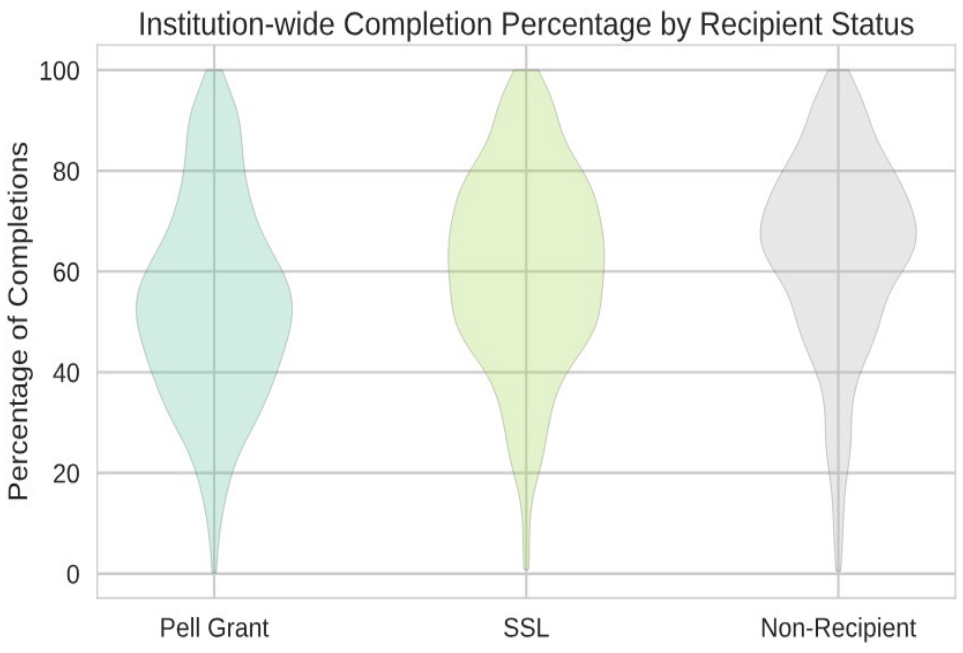
With Smoothing

	gr2mort	graiant	grasiat	grbkaat	grhispt	grnhpit	grwhitt
0	28.0	21.0	19.0	24.0	21.0	24.0	21.0
1	46.0	52.0	68.0	47.0	55.0	56.0	53.0
2	20.0	20.0	20.0	0.0	96.0	20.0	0.0

Exploratory Data Analysis

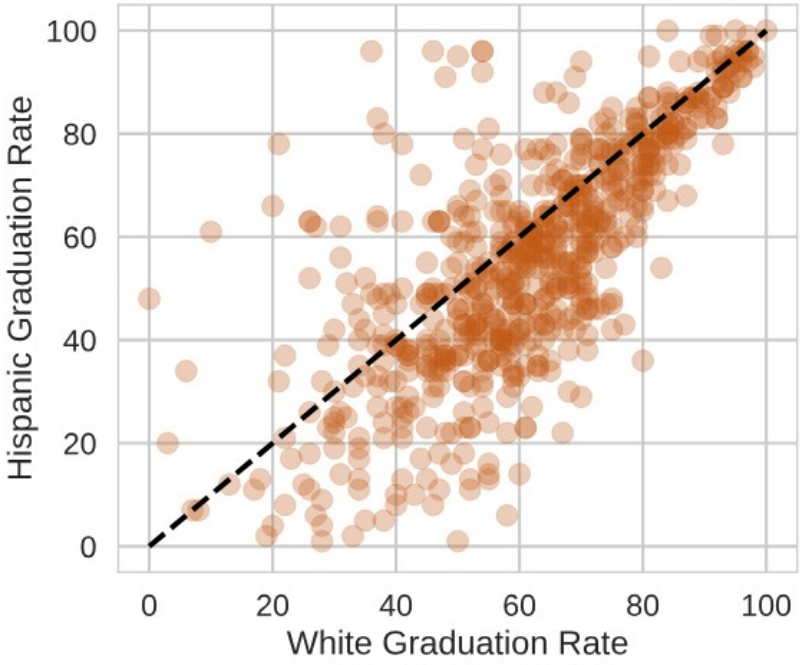
Lower mean completion percentages for some groups

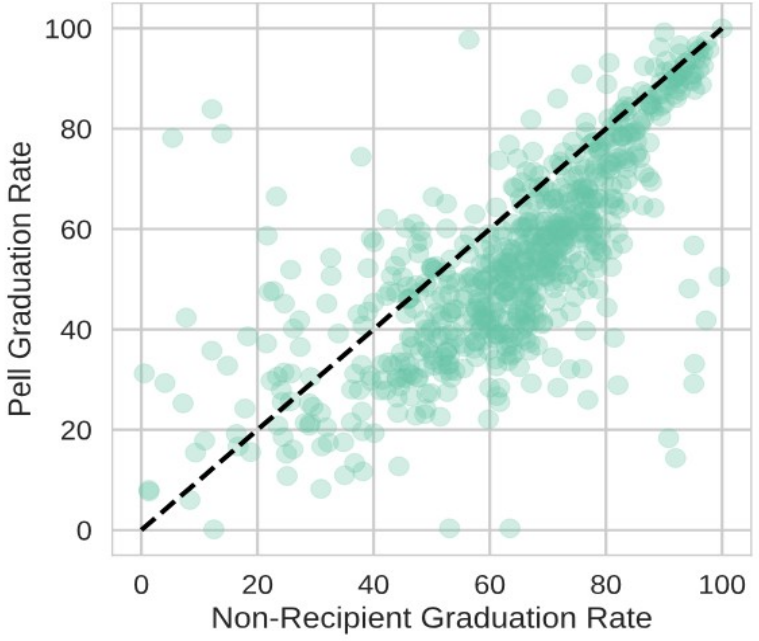




Exploratory Data Analysis

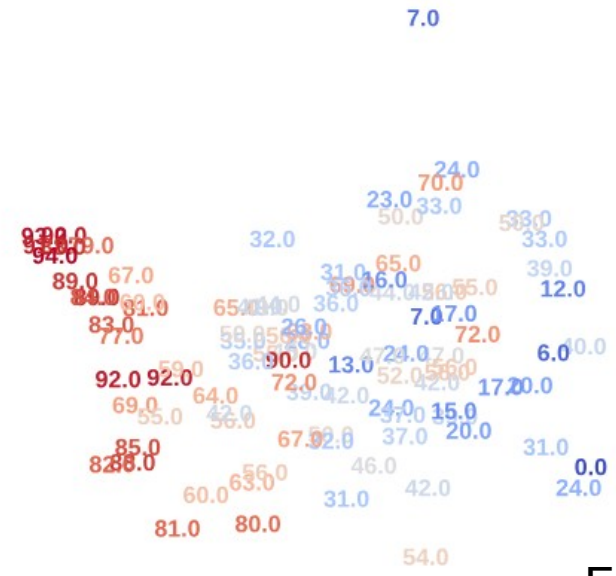
Strong correlation of graduation rates between different groups



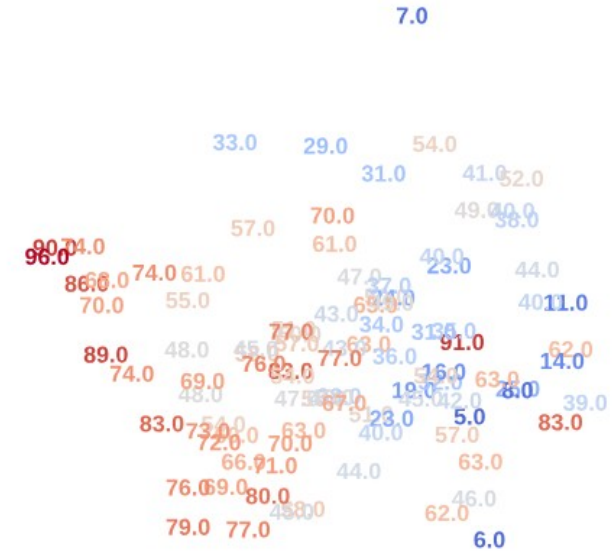


Exploratory Data Analysis

Principal Component Embedding: Black Graduation Rate



Principal Component Embedding: Hispanic Graduation Rate



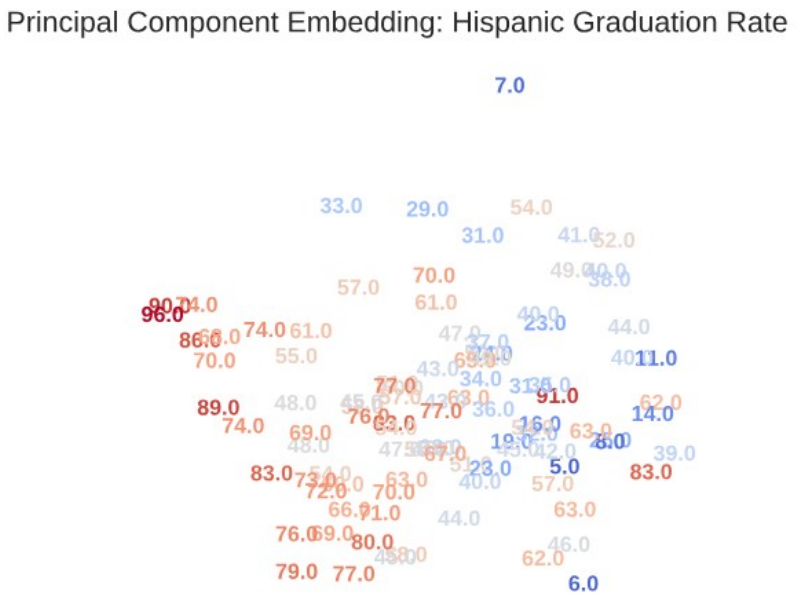
First PCA axis weighted by:

- Benchmark SAT/ACT scores
- Percentage of students receiving a Pell Grant

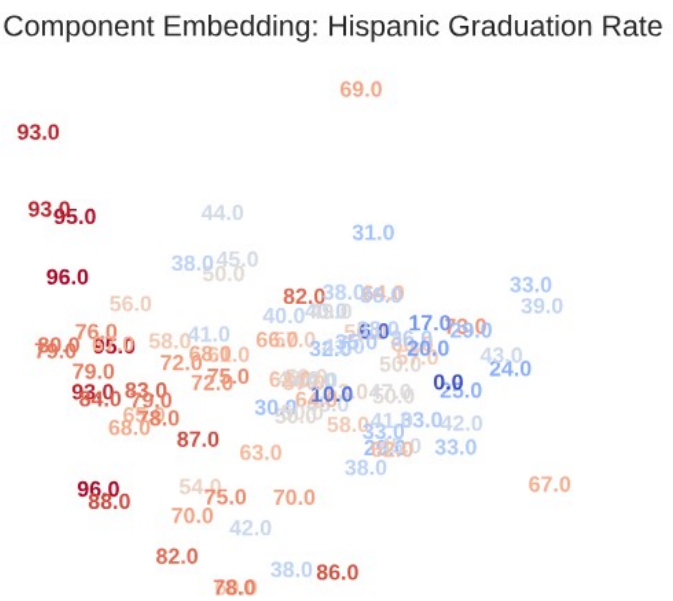
Exploratory Data Analysis

- Laplace smoothing affects separation of Asian graduation rates in PCA

With Smoothing



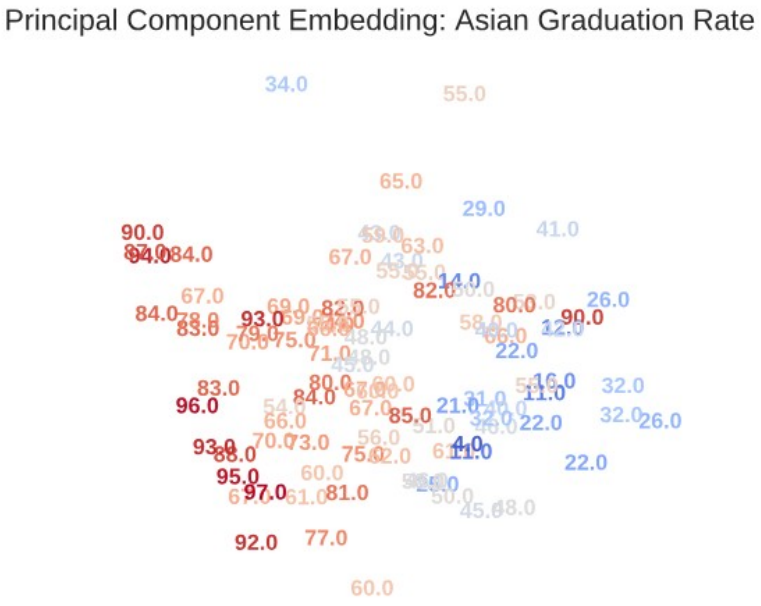
Without Smoothing



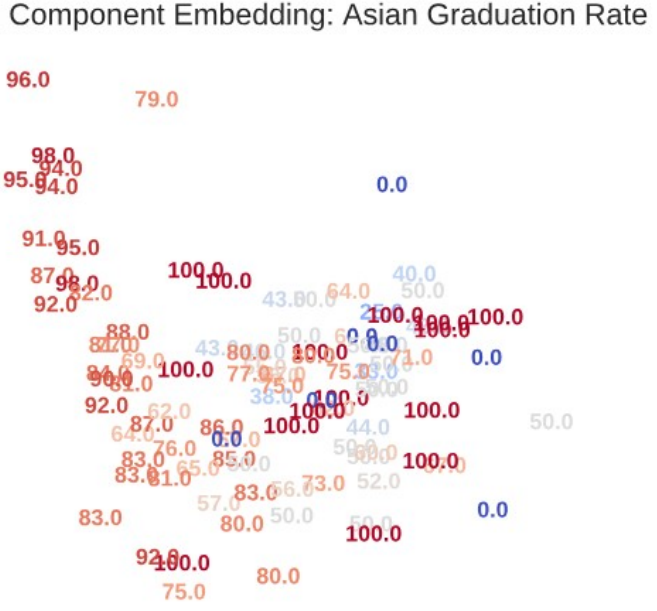
Exploratory Data Analysis

Laplace smoothing affects separation of Asian graduation rates in PCA

With Smoothing



Without Smoothing



Linear Regression Modeling

Variance Inflation Factors used to remove co-linear features (cutoff = 5)

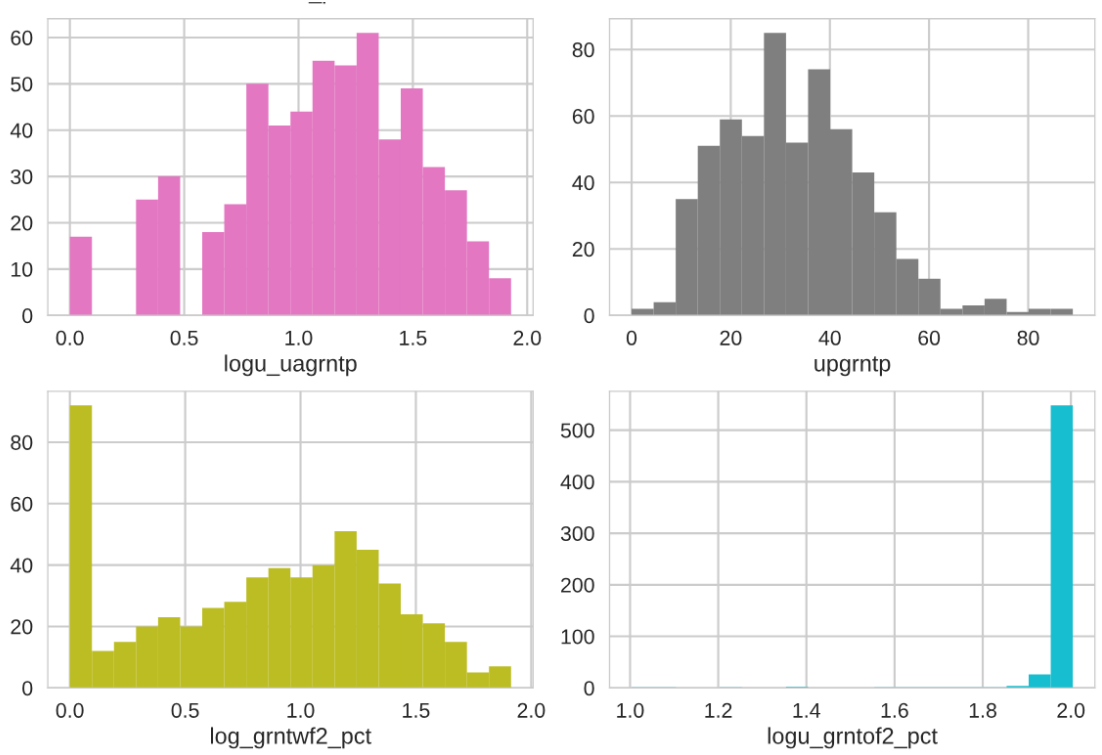
Surviving Features

Feature	VIF
Latitude	1.09
Longitude	1.14
% Enrolled Full Time	1.20
% Enrolled	1.29
% Admitted	1.42

Feature	VIF
% living off campus (w/ family)	1.44
% living off campus (w/o family)	1.32
English 25th percentile	2.81
% awarded any aid	1.84
% awarded Pell Grant	3.17

Linear Regression Modeling

Some features have very asymmetric distributions even after log transformations.



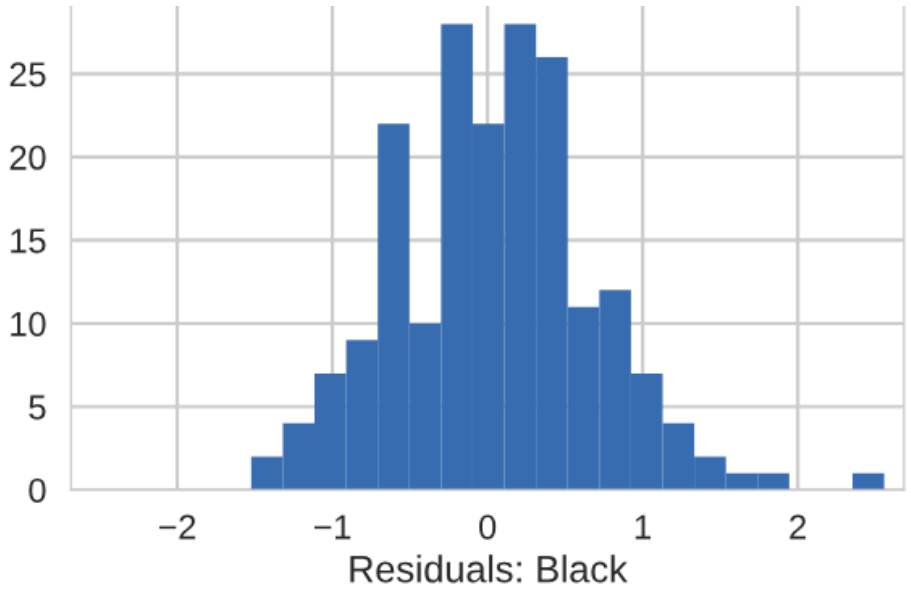
Linear Regression Modeling

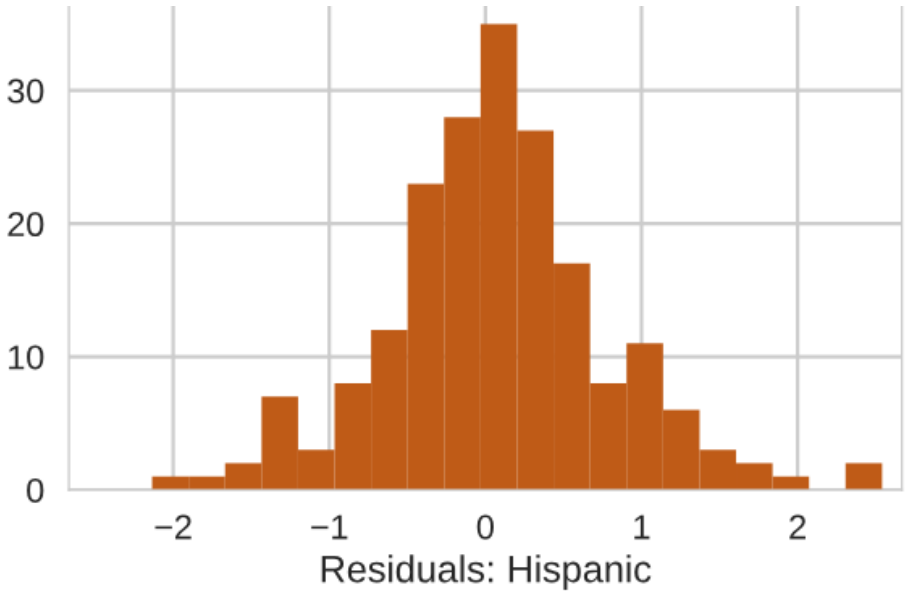
Regression Metrics

Target	Train R^2	Test R^2	RMSE
2 or More Races	0.56	0.44	15.4
Asian	0.49	0.41	17.2
Black	0.59	0.52	15.4
Hispanic	0.52	0.38	16.6
White	0.70	0.60	10.9
Pell Grant	0.67	0.55	12.5
Sub. Stafford Loan	0.57	0.46	14.1
Non-Recipient	0.54	0.39	15.4

Linear Regression Modeling

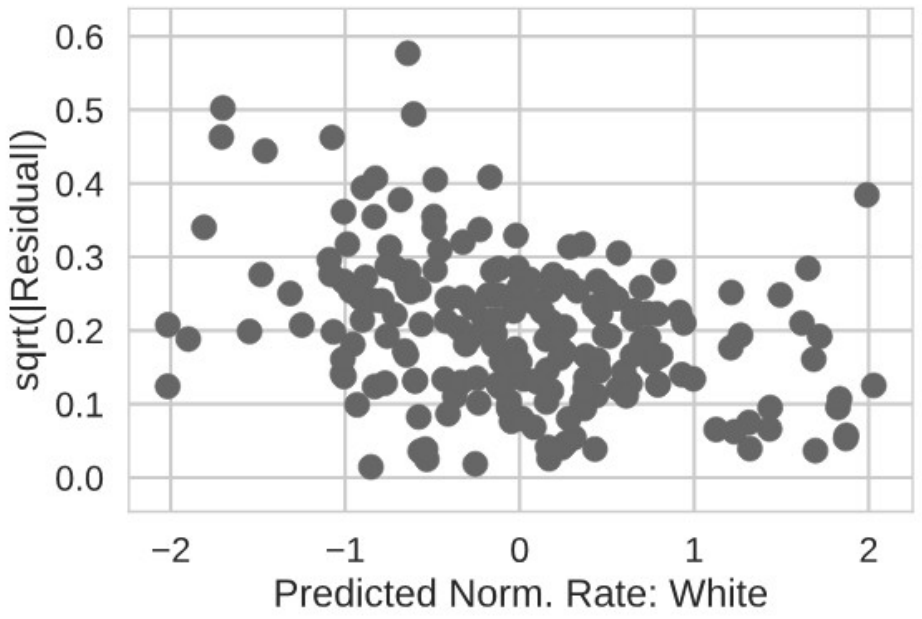
Residuals have normal-like distributions

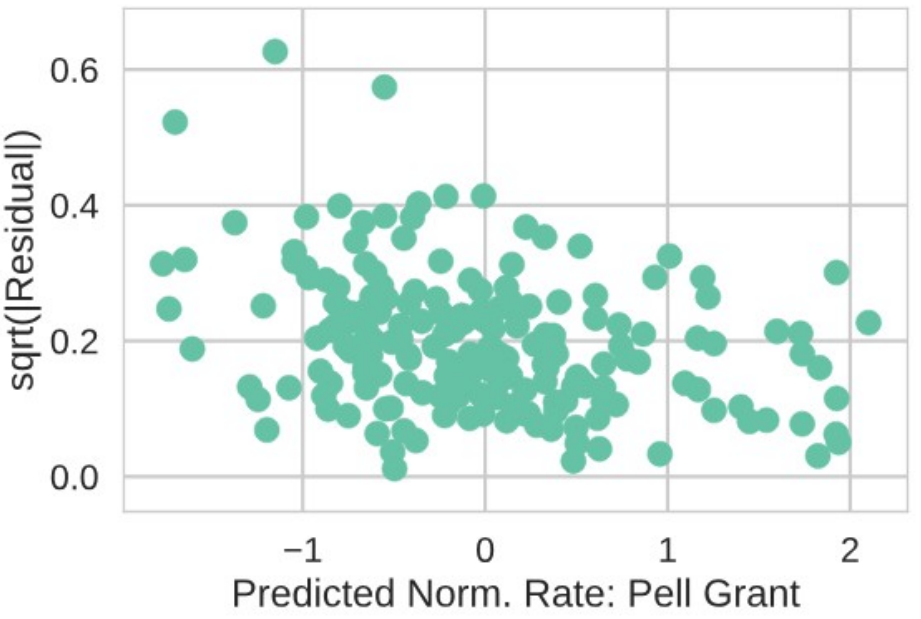




Linear Regression Modeling

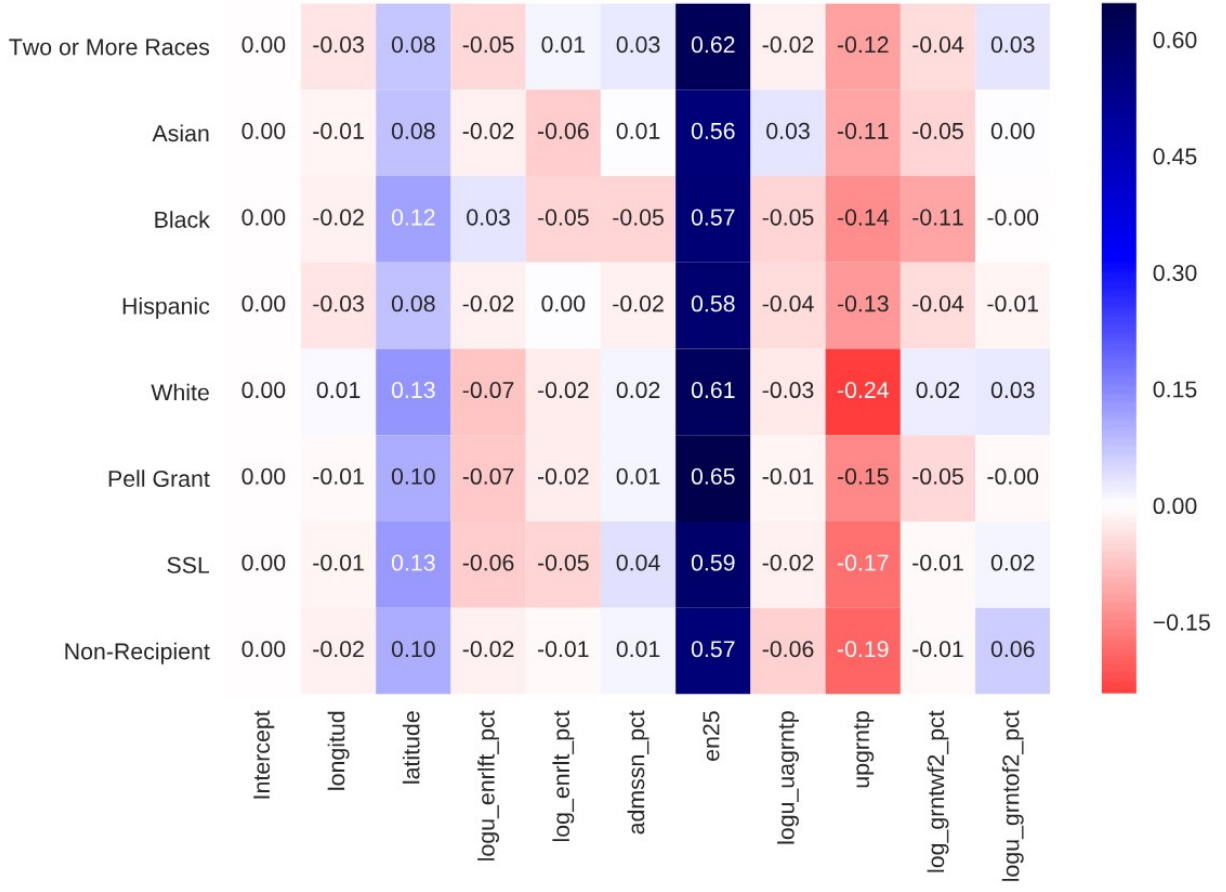
Homoscedacity tested with spread-location plots





Linear Regression Modeling

Normalized coefficients



Lasso Regularization

sklearn's LassoLarsCV

- 5-fold cross validation
- 500 maximum iterations

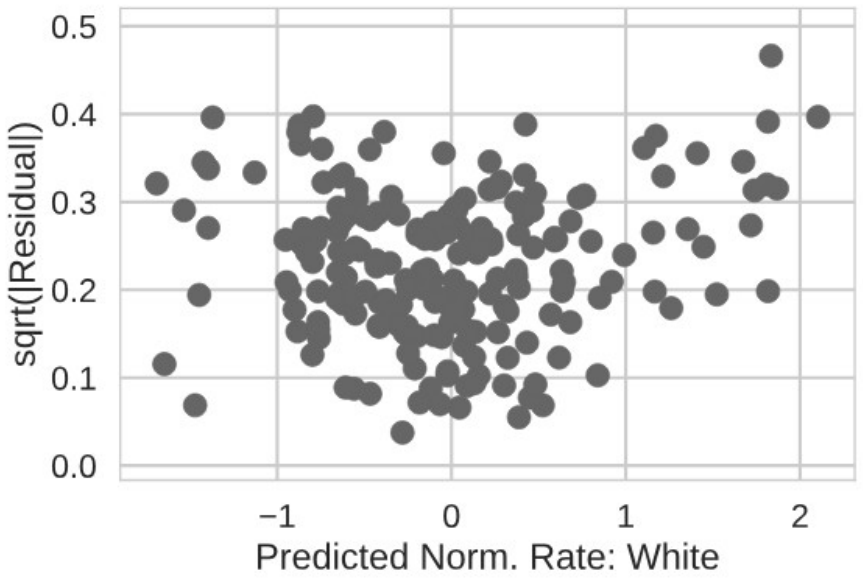
Best Hyperparameter Results

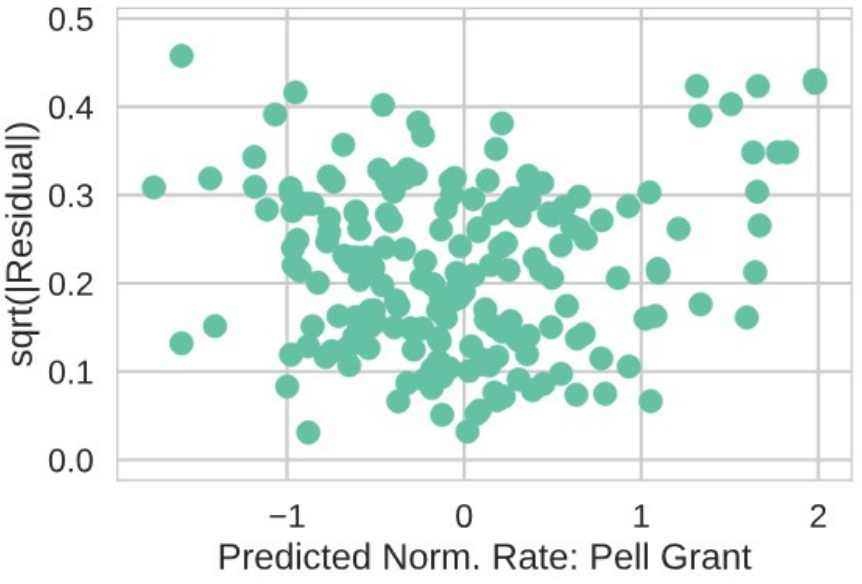
- Shrinkage parameter α = 0.002
- Converged within 10 iterations

Target	Test R^2
2 or More Races	0.44
Asian	0.41
Black	0.52
Hispanic	0.38
White	0.60
Pell Grant	0.55
Sub. Stafford Loan	0.46
Non-Recipient	0.39

Lasso Regularization

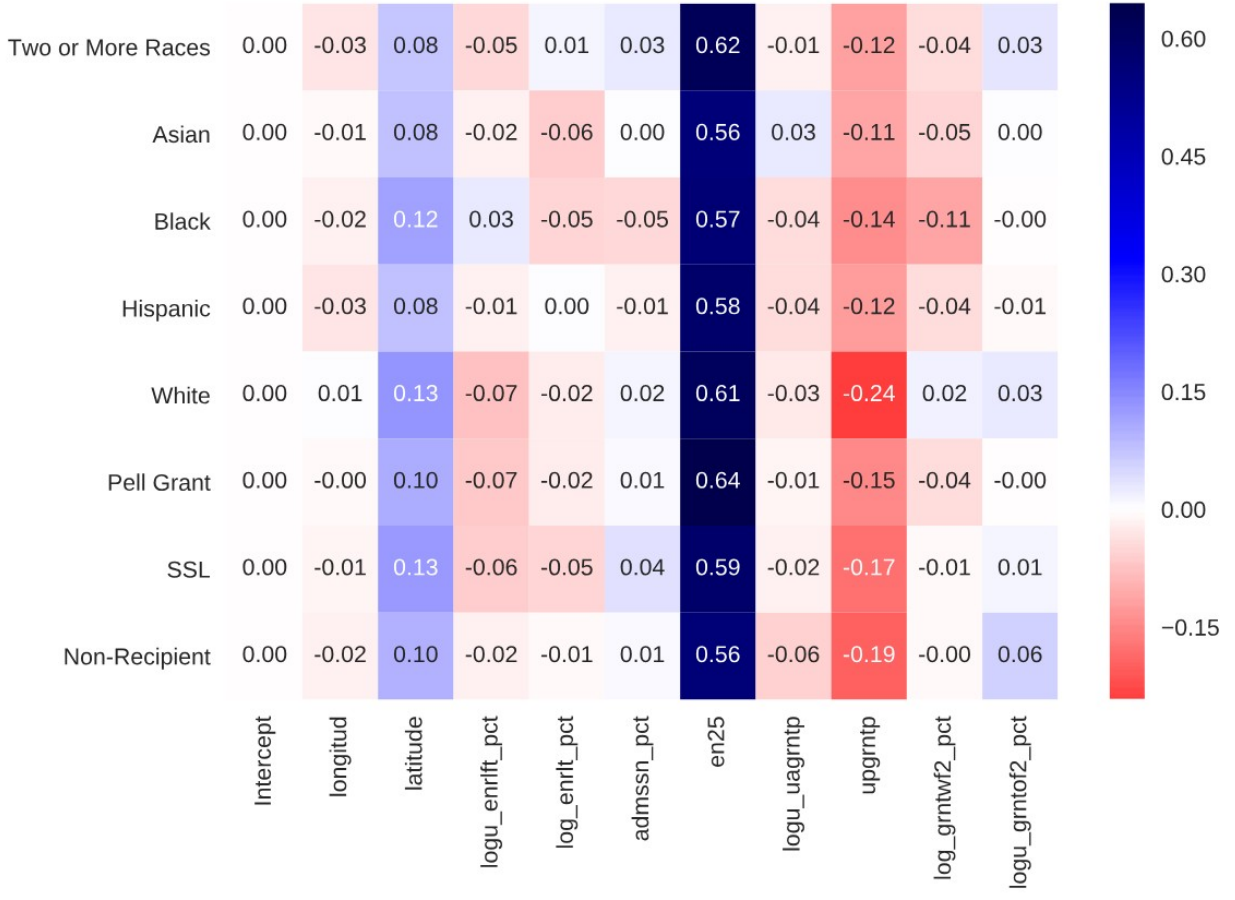
Improved spread-location plots





Lasso Regularization

Normalized coefficients



Random Forest Modeling

sklearn's RandomForestRegressor + RandomizedSearchCV

- 5-fold cross validation

Hyperparameter	Baseline	Search Range	Best Model
number of trees	100	100 - 1000	600
criterion for split	MSE	MSE, MAE	MAE
Min # for split	2	2, 5, 10, 20	5
Min # per leaf	1	1, 2, 5	2
Max features per split	sqrt(n)	n, sqrt(n)	sqrt(n)

Random Forest Modeling

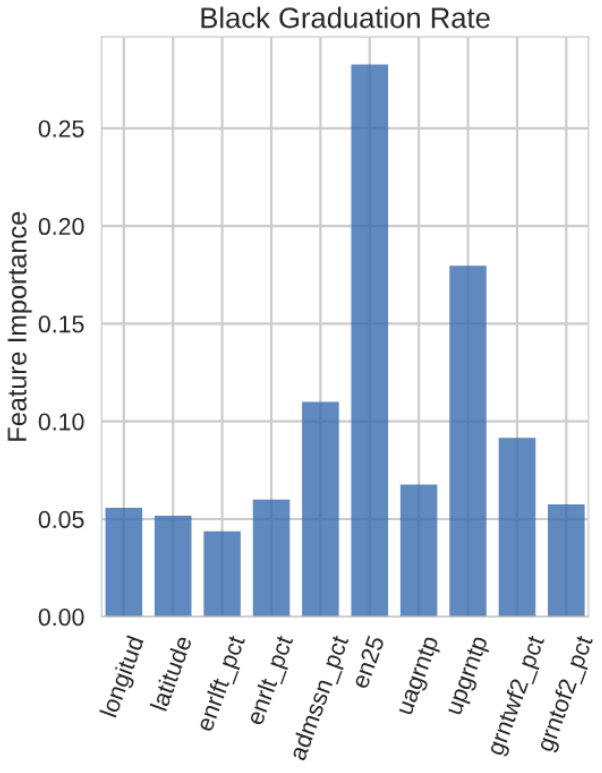
Model Metrics

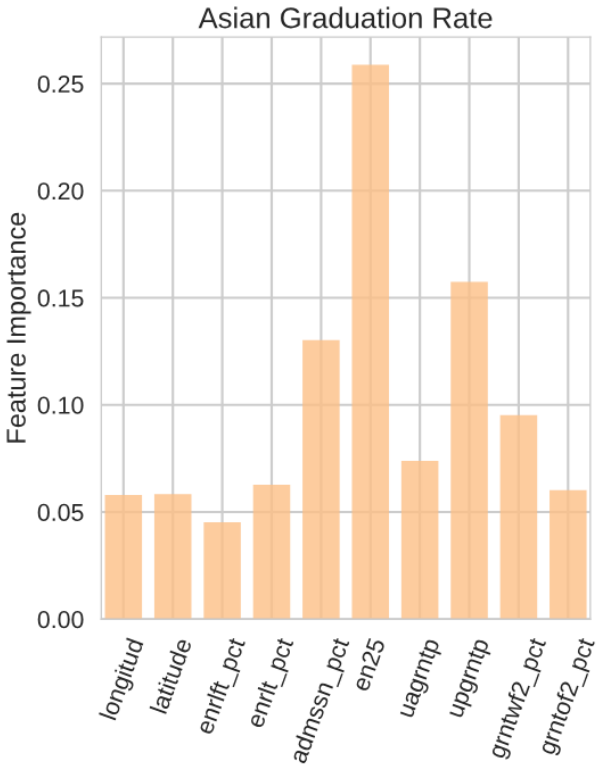
Target	Test R^2	Test RMSE
2 or More Races	0.44 (0.42)	15.30 (15.72)
Asian	0.38 (0.38)	17.27 (17.31)
Black	0.56 (0.55)	14.92 (14.86)
Hispanic	0.36 (0.32)	16.65 (17.23)
White	0.63 (0.62)	10.52 (10.70)
Pell Grant	0.57 (0.55)	12.33 (12.23)
Sub. Stafford Loan	0.48 (0.47)	13.84 (14.21)
Non-Recipient	0.39 (0.37)	15.34 (15.70)

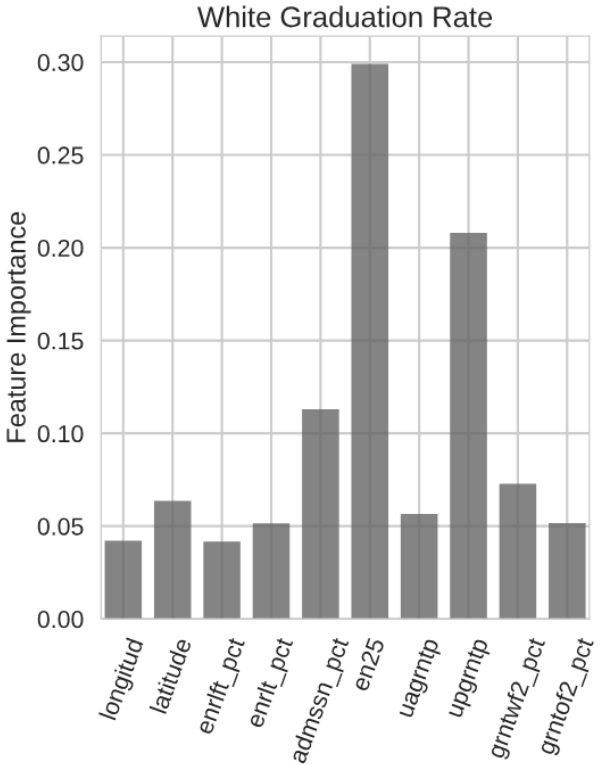
Baseline model results in parentheses

Random Forest Modeling

Feature Importance

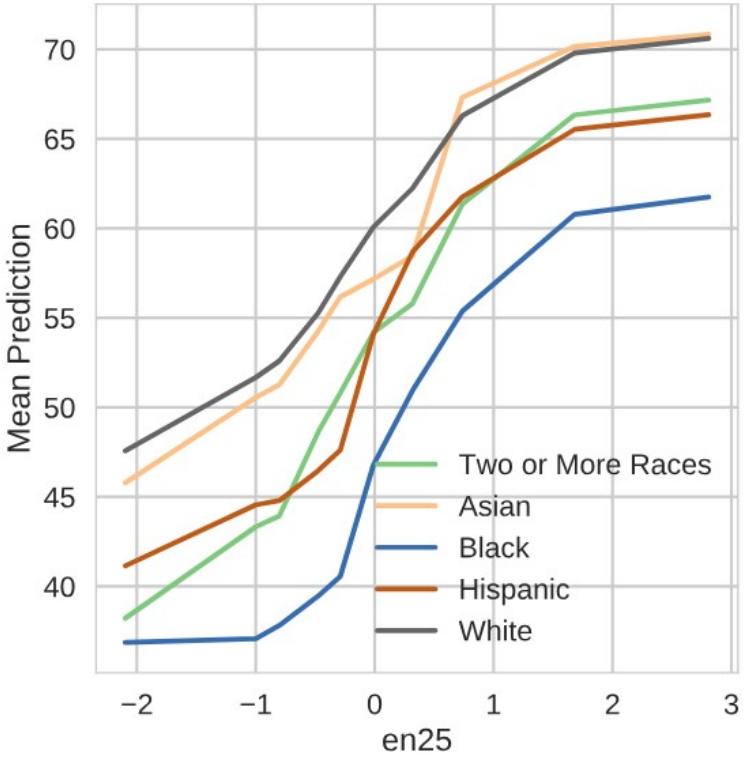


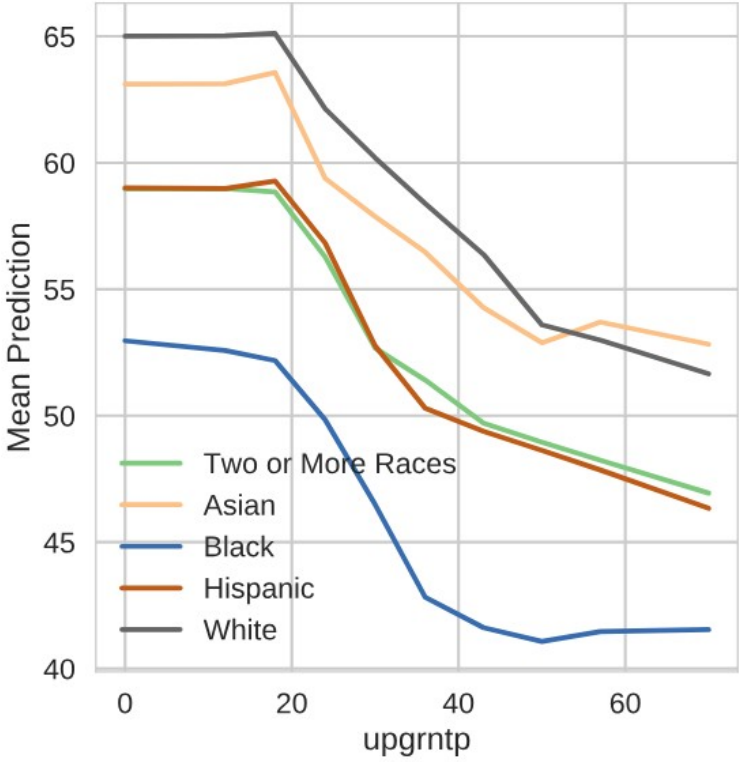




Random Forest Modeling

Partial Dependence





Conclusions

- R^2 values 0.4 – 0.6 for all models
- Most important predictors:
  - SAT/ACT benchmarks (positive correlation)
  - percent of undergraduates receiving Pell Grant (negative correlation)

Next Steps

- Incorporate institutional categorical data
- Add degree categories as an additional feature