Project 2

Linear Regression

Predicting Student Loan Debt

Carissa Sosa | Metis Fall 2019

Agenda

01 Background and Goals

02 Methodology and Regression

03 Summary

04 Next Steps

Background

01 Student Loan Debt

The National Student loan debt is now over $1.5 trillion [1]

02 Average & Median

The **average student debt** is $38,390. The median **student debt** is between $10,000 and $25,000[2]

1. Business Insider
2. Forbes

Goals

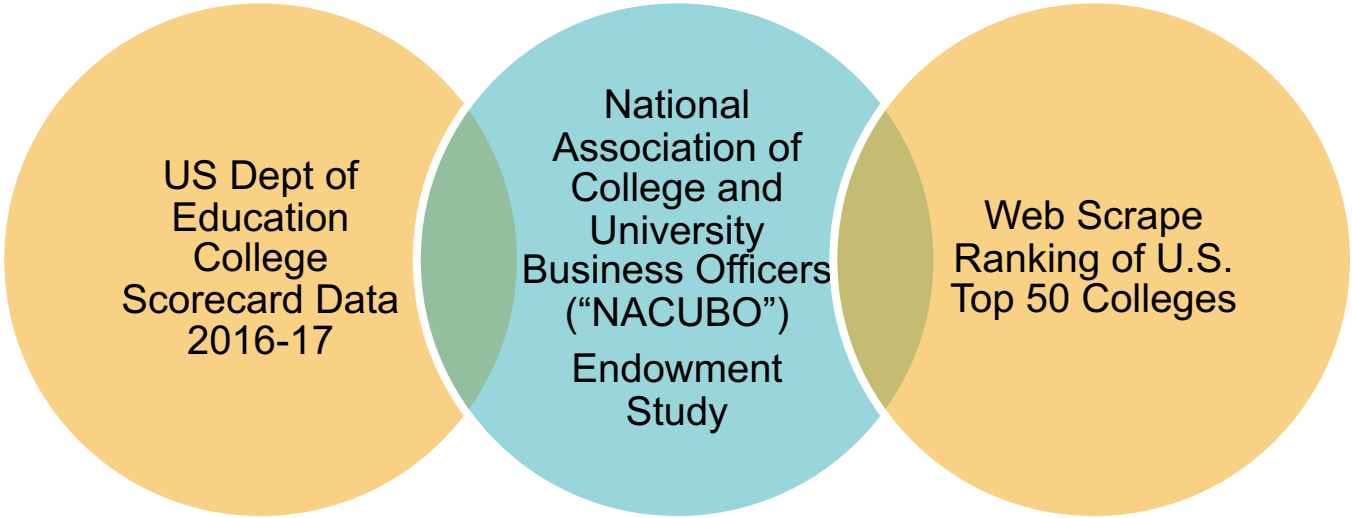
01 **Predict**
Student Loan Debt

02 **Toolkit**

03 **Linear Regression**

04 **Insights**

Data Sources for EDA



Target and Features

Target = Median Student Loan Debt

Tuition (In and Out of State)

Admission Rates

Average SAT Score

Range of Majors

Four Year vs. Two Year/Pub vs. Private

Endowment Value per Student (2017)

Age of Entry

% Received Pell Grants

Top 50 Schools

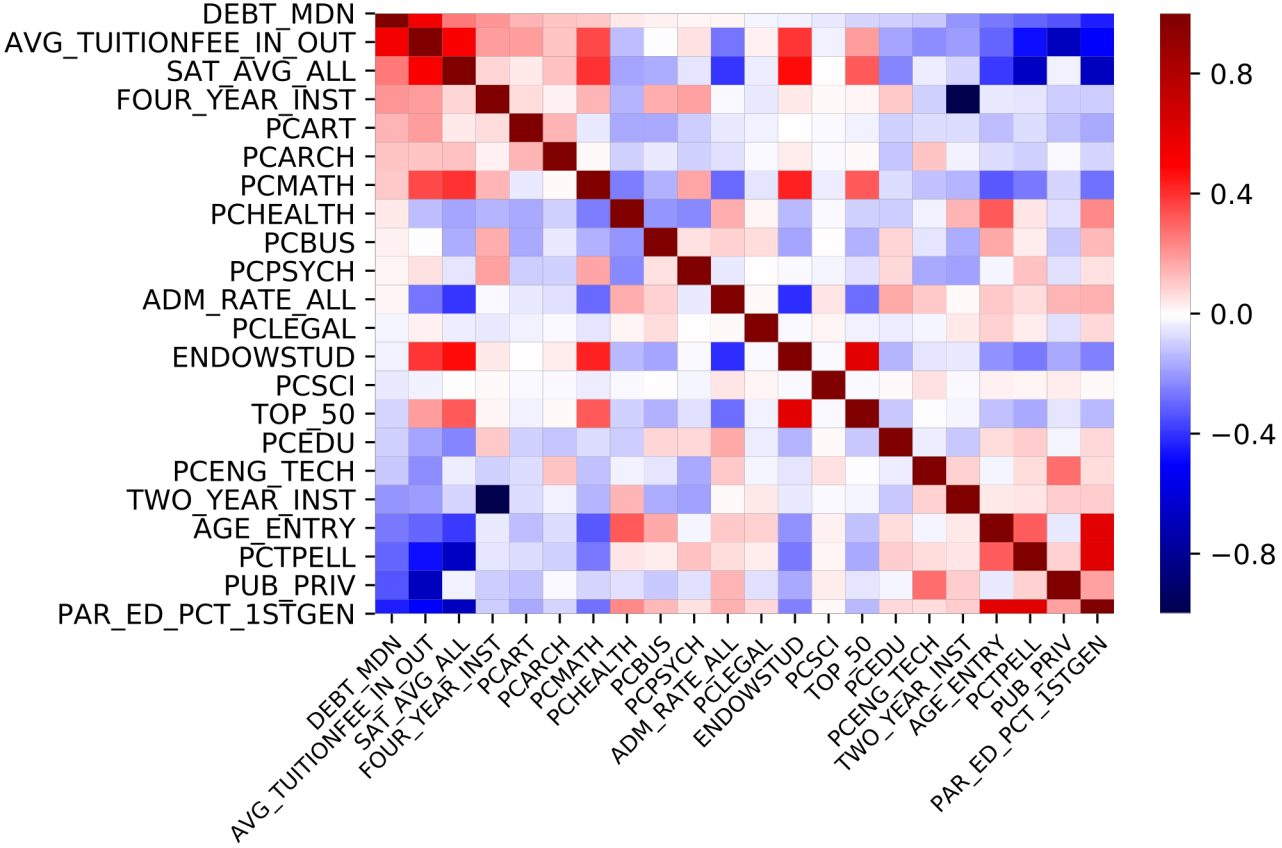
% of First Generation College Students

OLS Regression

Dep. Variable:	DEBT_MDN	**R-squared:**	0.466
Model:	OLS	**Adj. R-squared:**	0.457
Method:	Least Squares	**F-statistic:**	55.92
Date:	Thu, 10 Oct 2019	**Prob (F-statistic):**	1.21e-158
Time:	17:16:04	**Log-Likelihood:**	-12396.
No. Observations:	1305	**AIC:**	2.483e+04
Df Residuals:	1284	**BIC:**	2.494e+04
Df Model:	20		
Covariance Type:	nonrobust		

Omnibus:	18.522	**Durbin-Watson:**	1.737
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	30.350
Skew:	-0.077	**Prob(JB):**	2.57e-07
Kurtosis:	3.731	**Cond. No.**	6.15e+15

Collinearity



LassoCV

01 **Train – Val - Test**
60 - 20 - 20 train/val/test

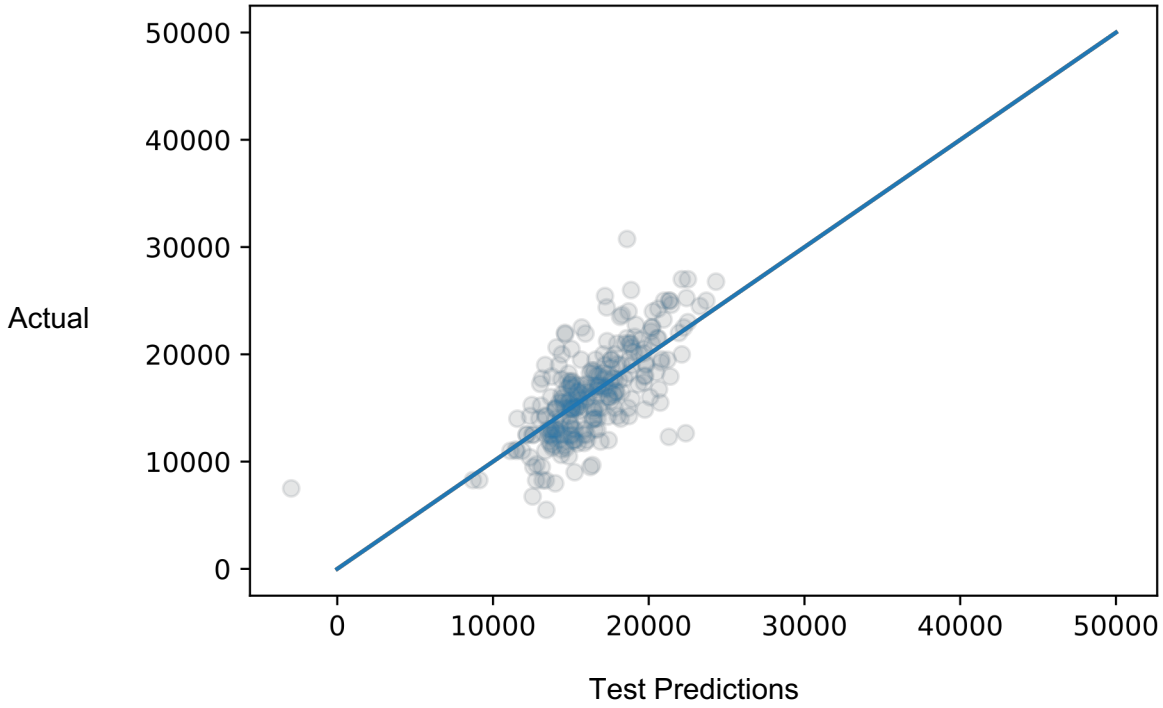
02 **Lasso Model Alpha = 28.66067**

03 **Set collinear features to 0**

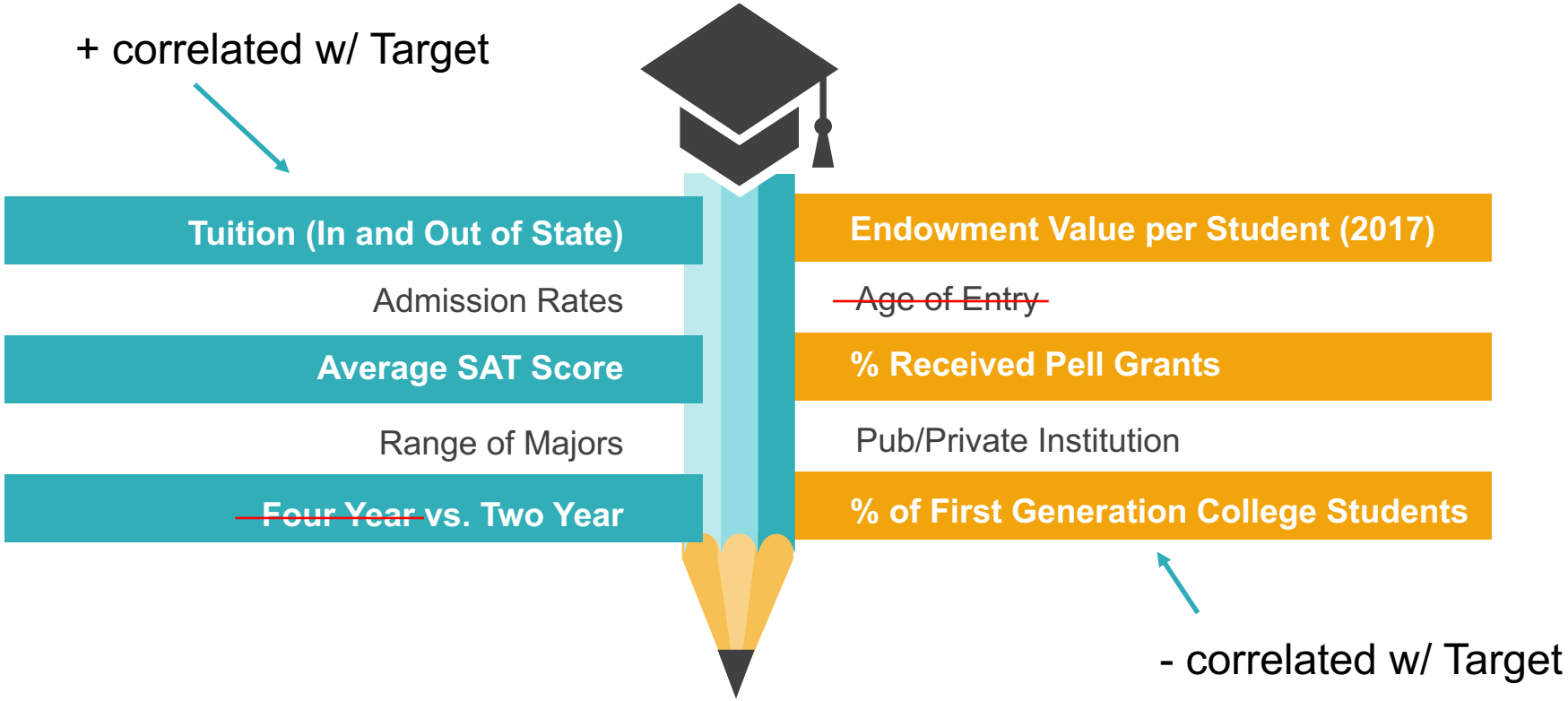
04 **R^2 on Test Set 0.475**

LassoCV

MAE: $2,337



LassoCV - Insights



Ridge Regression and CV w/ KFold

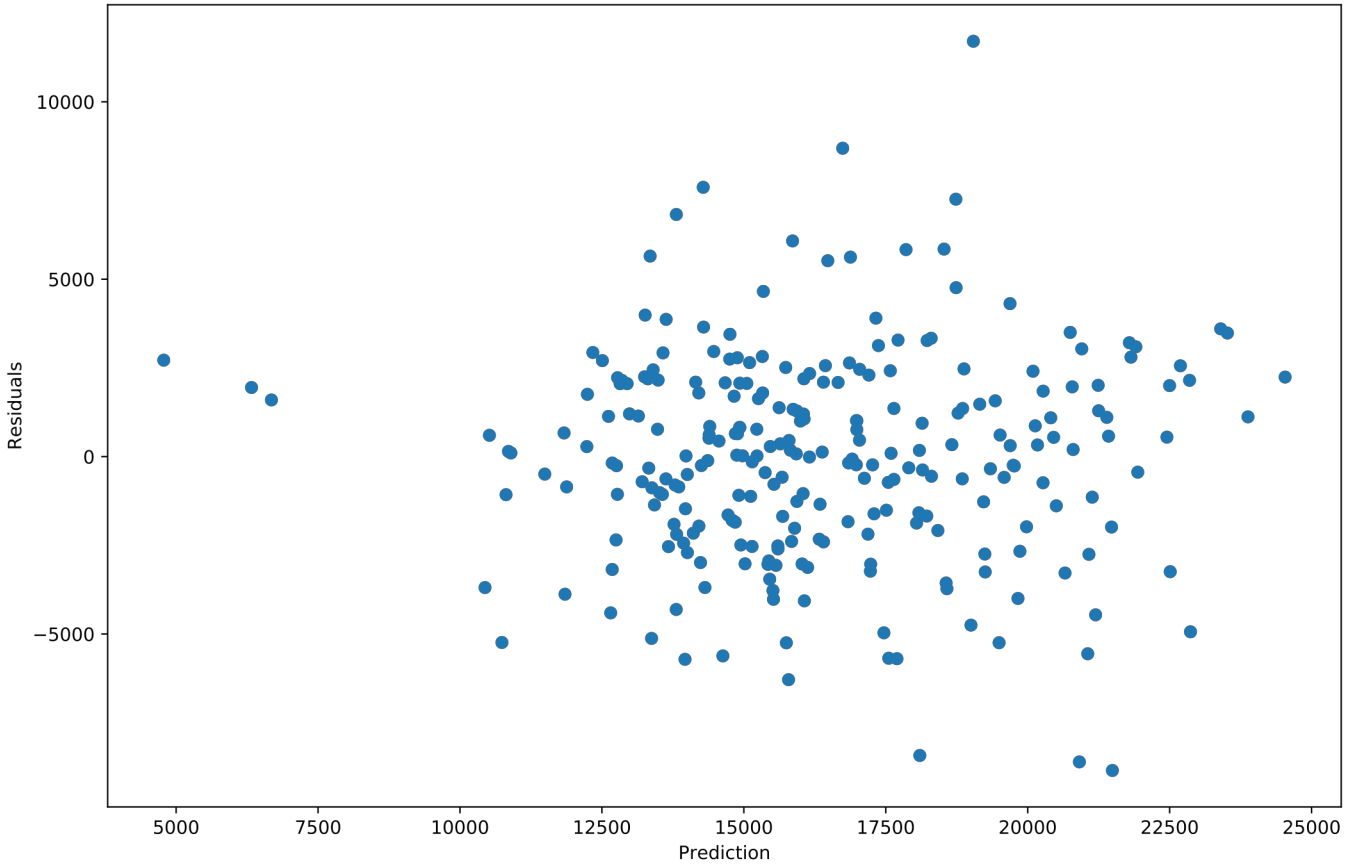
01 **Train – Val - Test**
60 - 20 - 20

02 **Ridge Regression R^2 = 0.479**

03 **CV with KFold R^2 = 0.434**

04 **Ridge Residuals vs Predicted**

Ridge Regression



Summary

01

Predict Student Loan Debt

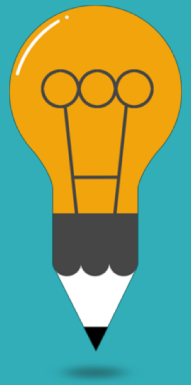
Can account for 47.9% of the variability in student loan debt with current features and Ridge.

Next Steps

More Data

More Feature Engineering

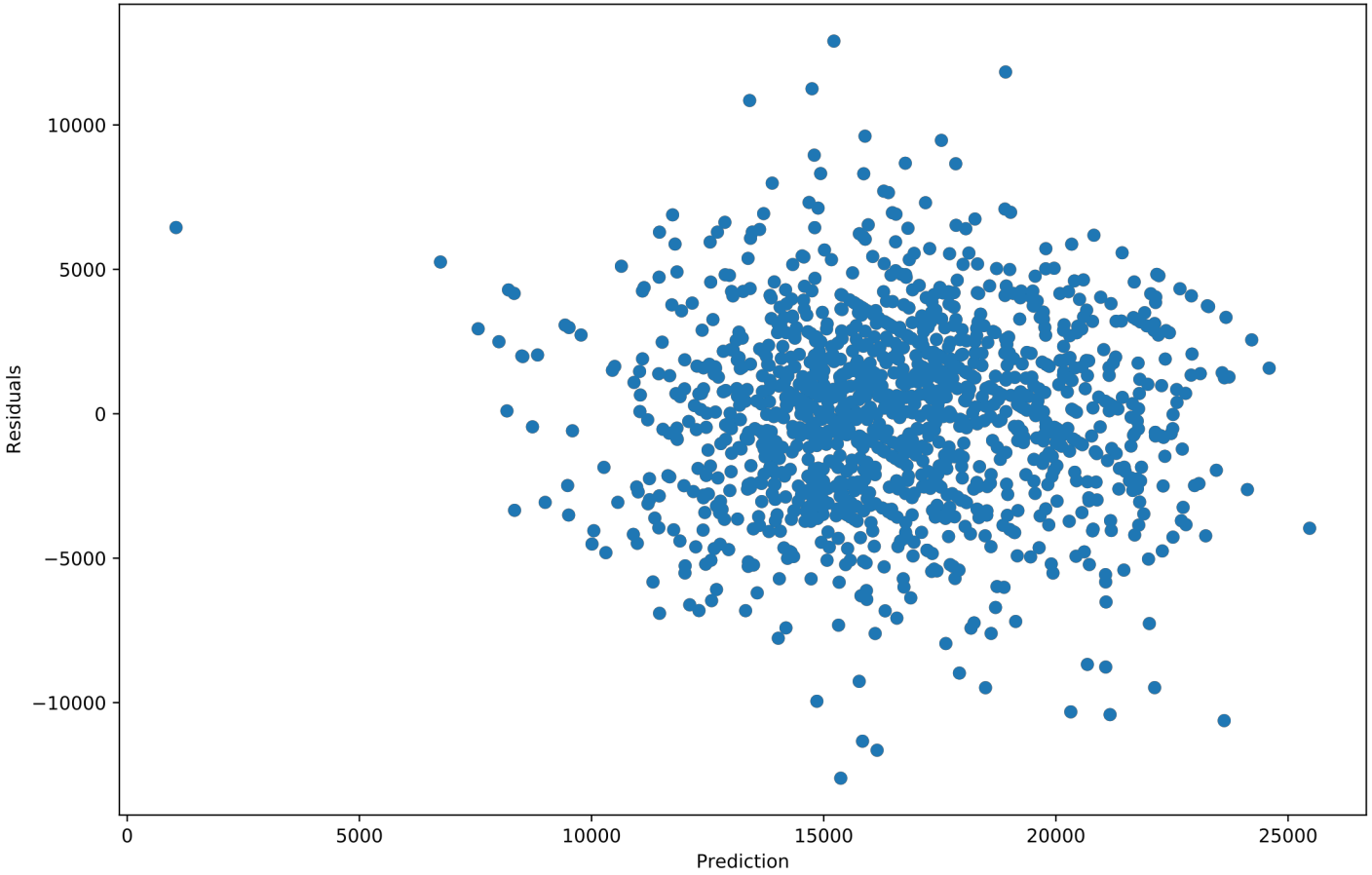
Inform



Questions?

Appendix

Prediction vs. Residual Initial



Lasso CV Coefficeints

```
[('SAT_AVG_ALL', -165.03886167415465, -1.2702785598506723),
 ('PCART', 0.0, 0.0),
 ('PCARCH', 249.4520077100926, 13849.152049193422),
 ('PCMATH', -0.0, -0.0),
 ('PCHEALTH', 664.250595766419, 3813.2692177281147),
 ('PCBUS', 143.45406936889694, 1130.022205437012),
 ('PCPSYCH', 107.3193049371959, 2165.972344007941),
 ('ADM_RATE_ALL', 124.51686544443928, 661.7532093875591),
 ('PCLEGAL', -66.96121513479353, -8187.288841723839),
 ('ENDOWSTUD', -1246.138931145982, -0.008089076969563786),
 ('PCSCI', -164.05719158128636, -50643.65458848483),
 ('top_50', 106.90696986758702, 952.0900693611467),
 ('PCEDU', -9.820676595169752, -164.4452037468024),
 ('PCENG_TECH', 60.198632134644434, 2345.795692466291),
 ('TWO_YEAR_INST', -348.8811540097601, -2327.9427844291345),
 ('AGE_ENTRY', -0.0, -0.0),
 ('PCTPELL', 305.9785139095461, 2236.334561020208),
 ('PUB_PRIV', 345.91139566758625, 705.4210714555124),
 ('PAR_ED_PCT_1STGEN', -1424.0957032765707, -15490.873507502936),
 ('TUITIONFEE_IN', 1496.0443320647587, 0.11036335547902568),
 ('TUITIONFEE_OUT', 1161.7130167515, 0.10949260789601761),
 ('FOUR_YEAR_INST', 0.0, 0.0)]
```