Supervised Learning Capstone

U.S. college graduation rates

By: Megan Dibble

1

Introduction

Motivation, Explanation of Data

Motivation

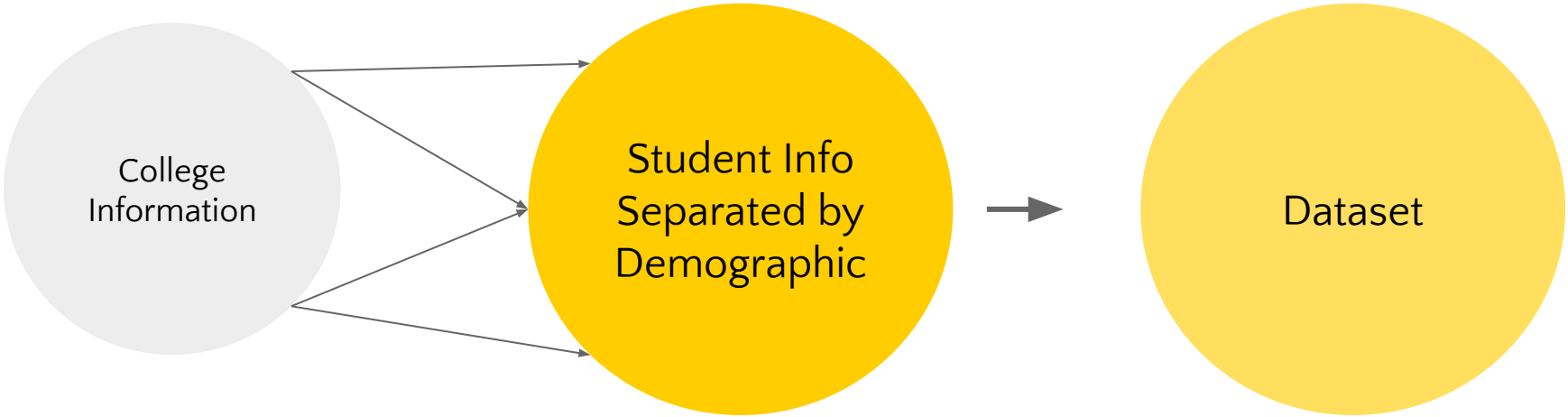
- Interest in college graduation rates/performance
  - Hot topic of increasing costs
- Data Source: Chronicle of Higher Education with support from the Bill & Melinda Gates Foundation
- What dimensions are most correlated to grad rates? What changes could be made? Any importance of student demographics?

Explanation of Data

- College completion data from 3,800 degree-granting institutions in the U.S.
- Full-time degree-seeking undergraduate cohort of at least 100 students at the undergraduate level in 2013
- Awarded undergraduate degrees **between 2011 and 2013**

Structure of Data

- Joined 2 tables (one-to-many)
- Resulted in 790560 observations, 70 columns

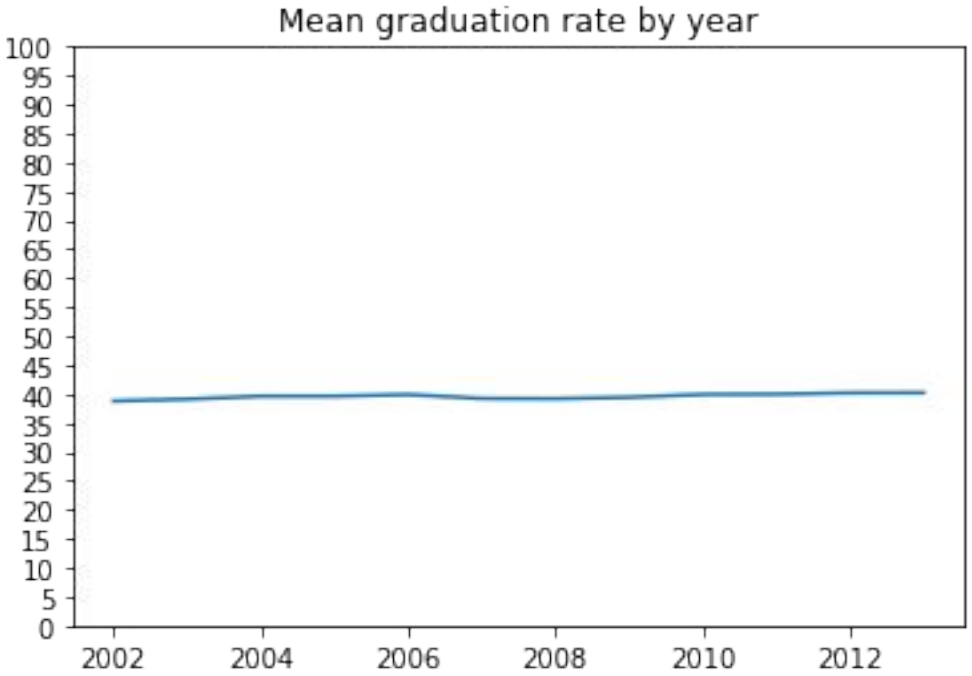


2

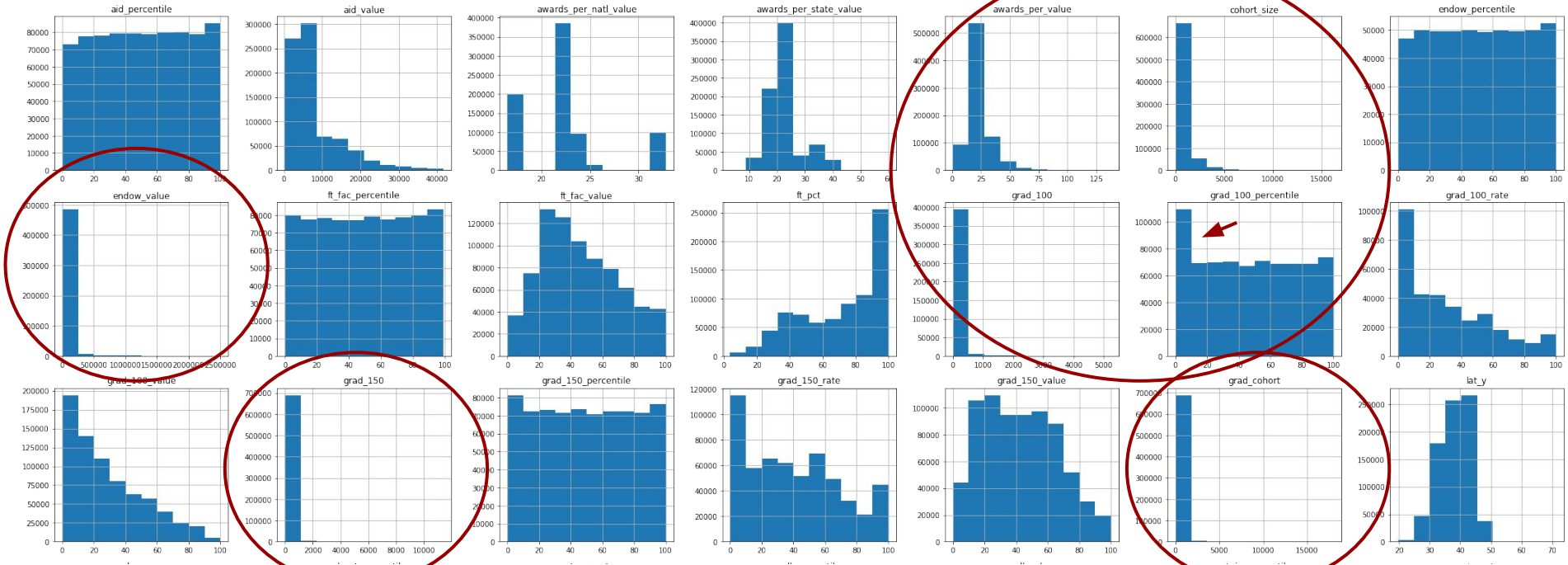
Data Exploration & Cleaning

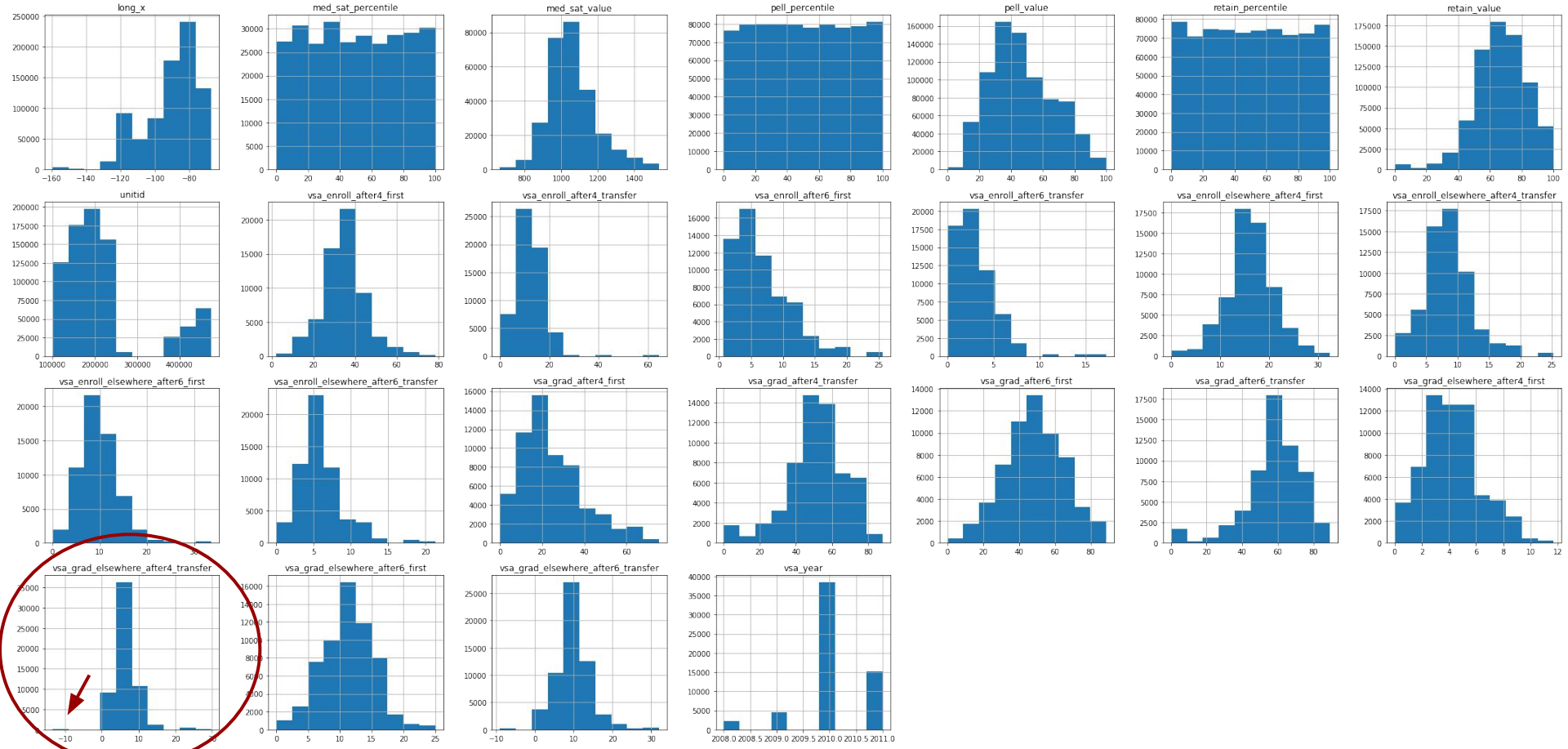
For continuous and discrete variables

Target Variable: 6-year Grad Rate



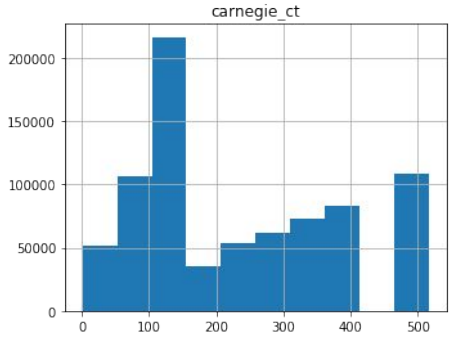
Continuous Variables

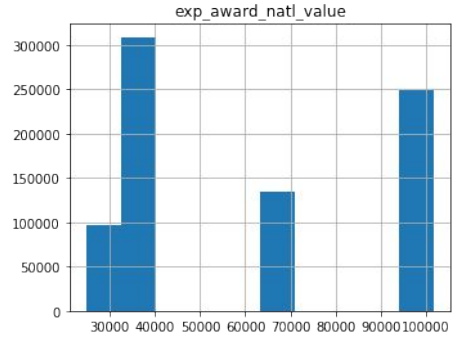


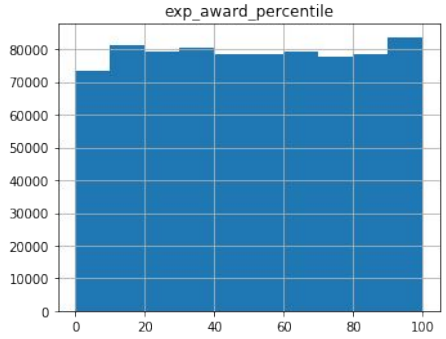

long_x
med_sat_percentile
med_sat_value
pell_percentile
pell_value
retain_percentile
retain_value
unitid
vsa_enroll_after4_first
vsa_enroll_after4_transfer
vsa_enroll_after6_first
vsa_enroll_after6_transfer
vsa_enroll_elsewhere_after4_first
vsa_enroll_elsewhere_after4_transfer
vsa_enroll_elsewhere_after6_first
vsa_enroll_elsewhere_after6_transfer
vsa_grad_after4_first
vsa_grad_after4_transfer
vsa_grad_after6_first
vsa_grad_after6_transfer
vsa_grad_elsewhere_after4_first
vsa_grad_elsewhere_after4_transfer
vsa_grad_elsewhere_after6_first
vsa_grad_elsewhere_after6_transfer
vsa_year

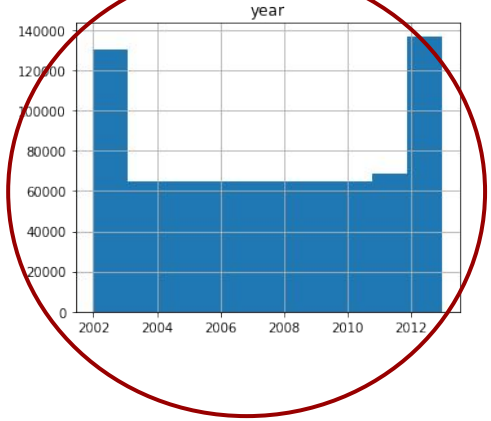
Discrete Variables

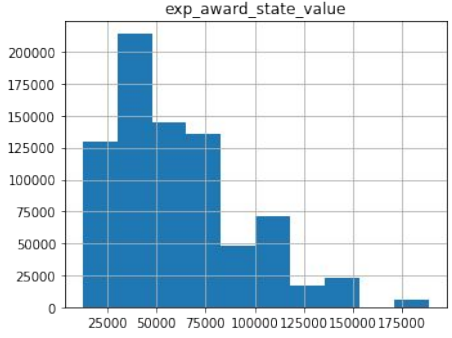
	chronname	city	state	level	control	basic	hbcu	flagship	site	counted_pct	nicknames	gender	race	cohort
count	790560	790560	790560	790560	790560	790560	19818	10800	787266	713286	66960	790560	790560	790560
unique	3793	1787	51	2	3	33	1	1	3333	1344	274	3	6	2
top	Metro Business College	Chicago	California	4-year	Public	Associates--Private For-profit	X	X	www.itt-tech.edu	100.0|07	NSU	M	H	4y bach
freq	648	8316	71550	479466	334260	108756	19818	10800	18090	19980	1080	263520	131760	468936

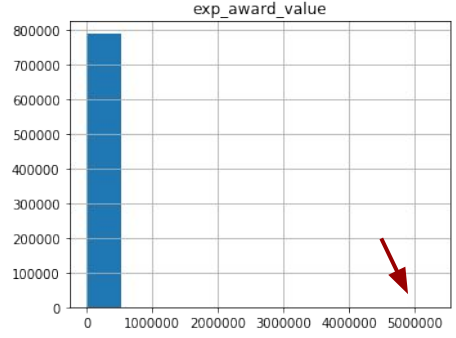

carnegie_ct
0
50000
100000
150000
200000
0
100
200
300
400
500

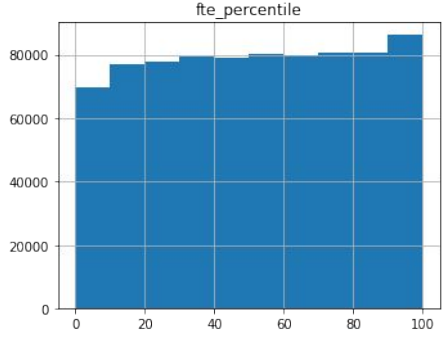

exp_award_natl_value
0
50000
100000
150000
200000
250000
300000
30000
40000
50000
60000
70000
80000
90000
100000

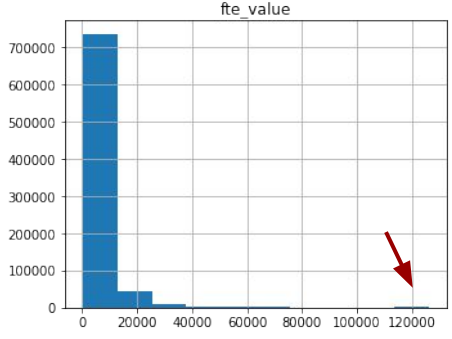

exp_award_percentile
0
10000
20000
30000
40000
50000
60000
70000
80000
0
20
40
60
80
100

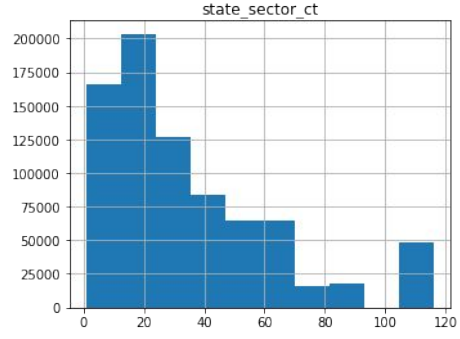

year
0
20000
40000
60000
80000
100000
120000
140000
2002
2004
2006
2008
2010
2012

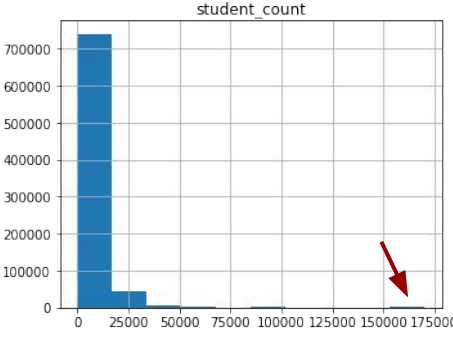

exp_award_state_value
0
25000
50000
75000
100000
125000
150000
175000
200000
25000
50000
75000
100000
125000
150000
175000


exp_award_value
0
100000
200000
300000
400000
500000
600000
700000
800000
0
1000000
2000000
3000000
4000000
5000000


fte_percentile
0
20000
40000
60000
80000
0
20
40
60
80
100


fte_value
0
100000
200000
300000
400000
500000
600000
700000
0
20000
40000
60000
80000
100000
120000


state_sector_ct
0
25000
50000
75000
100000
125000
150000
175000
200000
0
20
40
60
80
100
120


student_count
0
100000
200000
300000
400000
500000
600000
700000
0
25000
50000
75000
100000
125000
150000
175000

3 Feature Engineering

New Features Examined

- Student/Faculty Ratio
- Interaction Variables:
  - Public School * Endowment
  - Public School * Aid
  - Public School * Percentage of students who transfer out and graduate elsewhere
    - Had to be an interaction variable because of limited data
    - Only significant new variable

4

Modeling & Tuning Process

Modeling Process

Decision Tree Regressor

1. Randomized Grid Search
2. Fit Best Model
3. Prune Features
4. Randomized Grid Search on Smaller Dataset
5. Fit Best Model

Gradient Boosted Random Forest Regressor

1. Randomized Grid Search
2. Fit Best Model

Linear Regression

1. PCA
2. Fit Model with PCA Components
3. Fit Model with top 7 Variables (from Random Forest)
4. Chose to use PCA moving forward

Regularized Linear Regression Models

1. Randomized Grid Search
2. Fit Best Model
3. Repeat for Each Model

5 Results

Metric Comparison

	Train Score	Test Score	Train - Test
Decision Tree	.371	.344	.027
Decision Tree (Pruned)	.256	.107	.148
Gradient Boosted Random Forest	**.420**	**.398**	**.022**
Linear Regression (with PCA)	.322	.322	0
Best Regularized Linear Regression (Ridge)	.322	.323	-.001

6

Conclusion

Key Takeaways and Lessons Learned

Top 3 most important variables:

Freshman retention rate
Estimated educational spending per degree
Median SAT value (for incoming students)

Best model:

Gradient Boosted Random Forest

42% | 40%

Train accuracy | Test Accuracy

Takeaways

- Freshmen retention, educational spending per student, and student achievement prior to college (measured by SAT scores) are the most related to higher 6 year graduation rates.
- While correlation does not indicate causation, this is still meaningful insight
- Can provide data to back decisions made by universities and 2 year colleges.

Missing Variables

- Student GPA (first year)
- Transfer information for non-public 4-year institutions
- Percentage of faculty that are full time
- Faculty Salary

Further study: Look at 4 year grad rates or collect data above for colleges in the existing dataset

Lessons Learned

- Data cleaning can be an iterative process
  - Spending a lot of time with the variable descriptions helps
- Regularization (of linear regression models) provides most benefit when looking to increase prediction scores
- Ensure all engineered features have real meaning and the units make sense

Thanks!

Any questions ?

Sources

- Data Accessed On: https://data.world/databeats/college-completion
- Secondary Sources: https://www.chronicle.com/

  https://www.gatesfoundation.org/

  https://www.clarion.edu/about-clarion/offices-and-administration/university-support-and-business/office-of-institutional-research/retention-and-graduation-rate-analysis-clarion-university.pdf

  https://pdfs.semanticscholar.org/64f7/064c0d0d5b3417166e6e7ca9e4e157673eb2.pdf