

Chicago State University

Student Success

Maxwell Kreitman

The Problem:

Institution leaders want to increase the graduation rate of bachelor degree earning students.

The Solution:

Changing the ACT score requirement for matriculating freshman students will reliably effect their graduation potential.

Data:

- College Scorecard records from 2015/2016 academic year

Source: <https://collegescorecard.ed.gov/data/>

- College Scorecard records from 2018/2019 academic year

Source: <https://collegescorecard.ed.gov/data/>

- Institution geographic and demographic data

Source: https://simple.wikipedia.org/w/index.php?title=List_of_U.S._states&oldid=7168473

Data Wrangling:

- All three merged datasets revealed 6,354 records and 5,979 variables.
- College Scorecard documentation `FullDataDocumentation` and `CollegeScorecardDataDictionary` used to narrow variable total to 11.
- Dropped records of institutions who do not grant at a bachelor of arts or higher, missing ACT test entries, and graduation rate and admissions rate entries.
- Final shape of my data frame was 1187 rows and 11 columns.
- Target variable of the data is titled "graduation_rate".

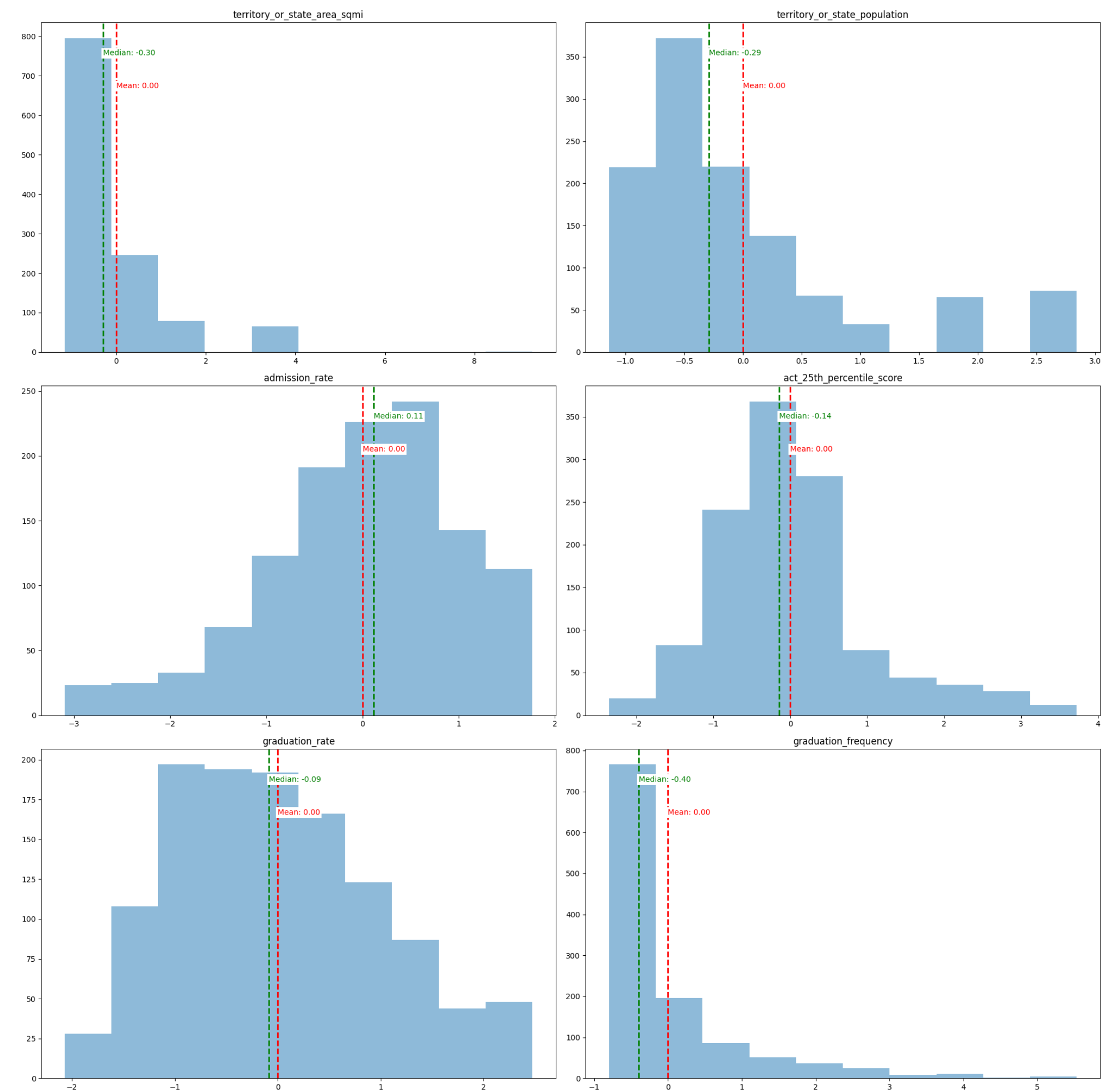
Exploratory Data Analysis:

Chicago State University

- 39th in admissions selectivity of matriculating students for the 2015/2016 academic year.
- 1135th in average ACT 25th percentile score for the 2015/2016 academic year.
- 1176th in graduation rate of bachelor degree earning students during the 2018/2019 academic year.
- 884th in terms of total bachelor degree earning students graduating during the 2018/2019 academic year.

Exploratory Data Analysis:

Distributions of continuous data



Exploratory Data Analysis:

Heatmap



Notable correlations:

- Positive strong correlation between act_25th_percentile_score and graduation_rate features (0.76)
- Positive medium/strong correlation between territory_or_state_area_sqmi and territory_or_state_population features (0.62).
- Negative medium/strong correlation between admission_rate and act_25th_percentile_score features (-0.45),
- Negative medium correlation between admission_rate and graduation_rate features (-0.35).

Machine Learning Modeling:

Preprocessing

- Dummy variables created from state/territory abbreviation series. The data features now total 57.
- Subset the data into training/test sets using 70:30 split.

Machine Learning Modeling: Selection

Since the data is labeled and continuous, I decided on the following supervised regression models:

- Simple linear
- Multiple linear
- Support vector machine
- Decision tree
- Random forest

Machine Learning Modeling: Performance

To assess the performance of each model, I compared the mean squared error and coefficient of determination results of each.

I determined that multiple linear and random forest were the two best out of the 5 I tried:

Multiple linear achieved a mean squared error of 0.38, and a coefficient of determination of 64%.

Random forest achieved a mean squared error of 0.39, and a coefficient of determination of 62%.

Machine Learning Modeling:

Hyper parameter tuning

Scikit-Learn's Grid Search Cross Validation function revealed new parameters to try on the random forest model.

Results from hyper parameter tuning include:

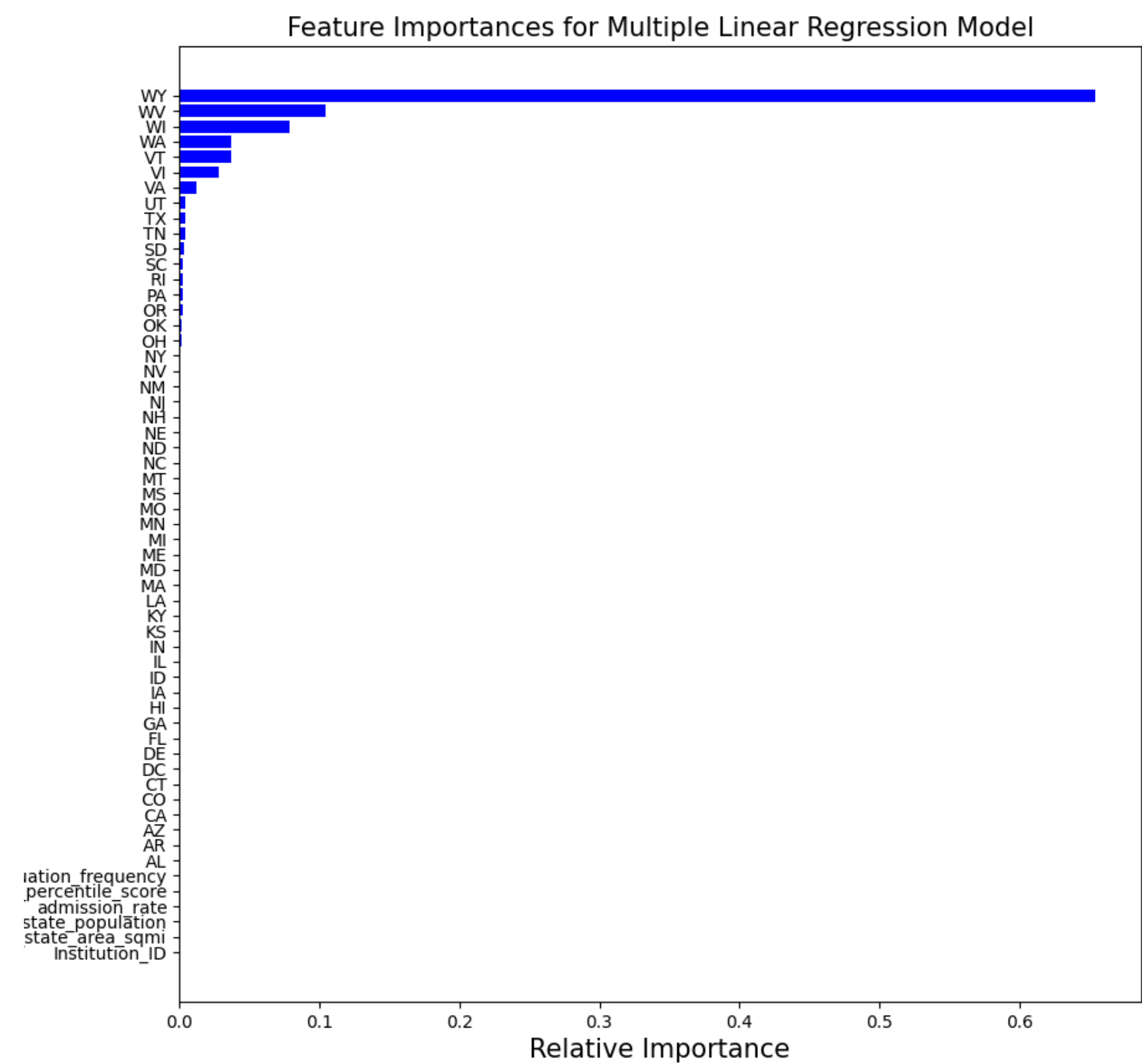
- Mean squared error increase from 0.39 to 0.399
- Coefficient of determination increase from 62% to 62.27%

Conclusion:

While the variance explained by the model got slightly better, the magnitude of errors got slightly worse.

The multiple linear regression model continues to be the best performing out of the 5.

Machine Learning Modeling: Feature Importances



Dummy variables of territories, especially WY, have an effect on graduation rate.

Further Research

- Institutions belonging to WY need to be investigated.
- Denoting institutions as private or public may contribute to understanding how each type effect graduation rate. Public and private institutions have different ways of dealing with students who are underperforming.
- Current data set does not track volume of transfer students that matriculate or leave an institution each year. Why this is problematic:
 1. A student who transfers in as a sophomore/junior/senior and graduates on time is counted in the data's graduation rate and graduation volume, but he/she does not contribute data to the ACT score or admissions rate values.
 2. A student who transfers out of an institution that he or she matriculated to as a freshman contributes ACT score and admission data but no graduation data.
 - This type of student must be in good standing, and therefore is likely to complete his or her bachelor's degree on time. However, this student's freshman matriculated institution counts the transfer negatively towards its graduation rate.

Thank you!

Ricardo Alanis for guidance during this project