Final Report: Postsecondary Education Data Analysis

Problem Statement

Institution leaders at the Chicago State University wants to explore to understand how their own admissions test requirements (ACT and/or SAT) and graduation rate for bachelor's degree earners compare to other 4-year institutions belonging to the American education system. In an effort to increase the success of student outcomes, they want to model different scenarios of admissions requirements to see its impact on bachelor degree graduation rate.

By merging Integrated Postsecondary Education Data System data and geographic/ demographic data from Wikipedia I created a tool that can reasonably explain 64% of the graduation rate variation of the institutions considered with reliable data.

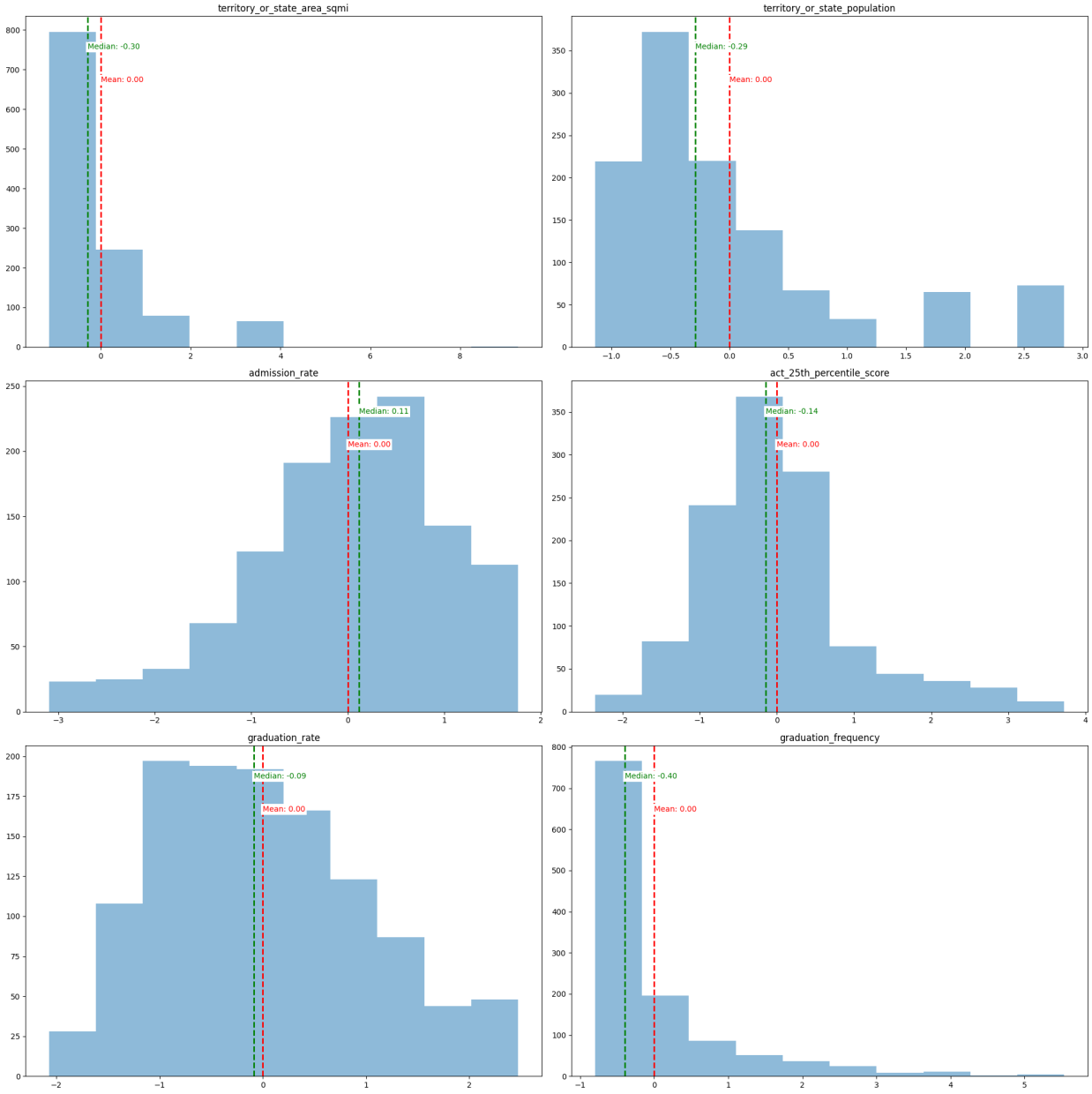
Data Wrangling

To begin the project, I merged two datasets from the Integrated Postsecondary Education Data System in order to follow a class of bachelor's degree destined students between the 2015/2016 academic year until their expected graduation in the 2018/2019 academic year. This resulted in a data frame with 6,354 records and 5,977 variables. Before investigating the reliability and importance of these variables, I merged additional geographic data of the states or territories each institution belongs to that includes population and total area in square miles. This addition of two features brings my column count to 5979. Dimensionality reduction is necessary to move forward with finding insights. I used the FullDataDocumentation pdf and CollegeScorecardDataDictionary spreadsheet to narrow down my extensive list of features down to 11. After dropping records of institutions who do not grant at a bachelor of arts or higher, as well as and those missing ACT test, graduation rate and admissions entries, the final shape of my dataset was 1187 rows and 11 columns.

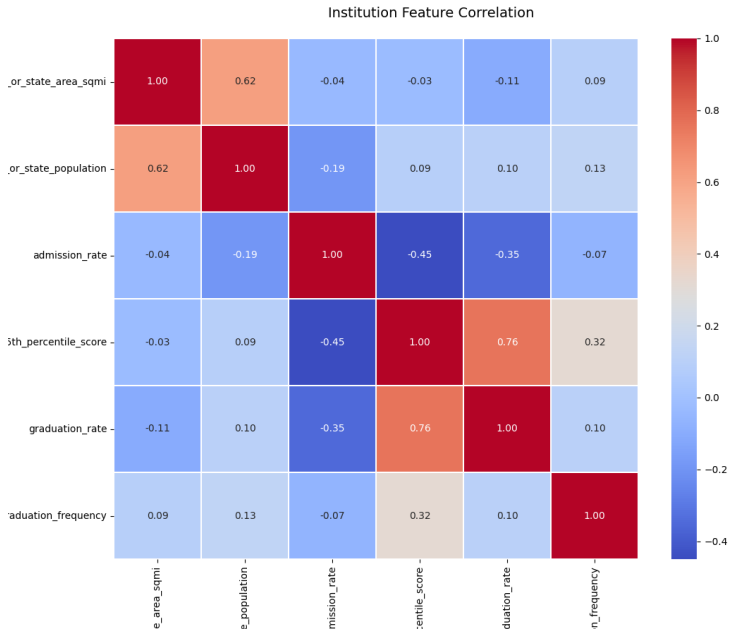
Exploratory Data Analysis

In order to learn more about how Chicago State University compares to the other institutions in the data, I began with the state it resides in. Illinois is the 25th largest territory and shares 5th place with Ohio in terms of institutions with a total of 48. Admission selectivity of Chicago State University places it 39th, but its ACT 25th percentile score is extremely low placing at 1135th during the 2015/2016 academic year. In terms of graduation rate it is also extremely low placing at 1176th, only 11 positions from the absolute lowest belonging to the Southern California Institute of Architecture in the 2018/2019 academic year. Although Chicago State University's graduation rate seems unusually low for a reputable institution, during the same period it ranks at 884th in terms of total students graduating.

Following these observations, I created visualizations to describe the shape of the features with continuous data:

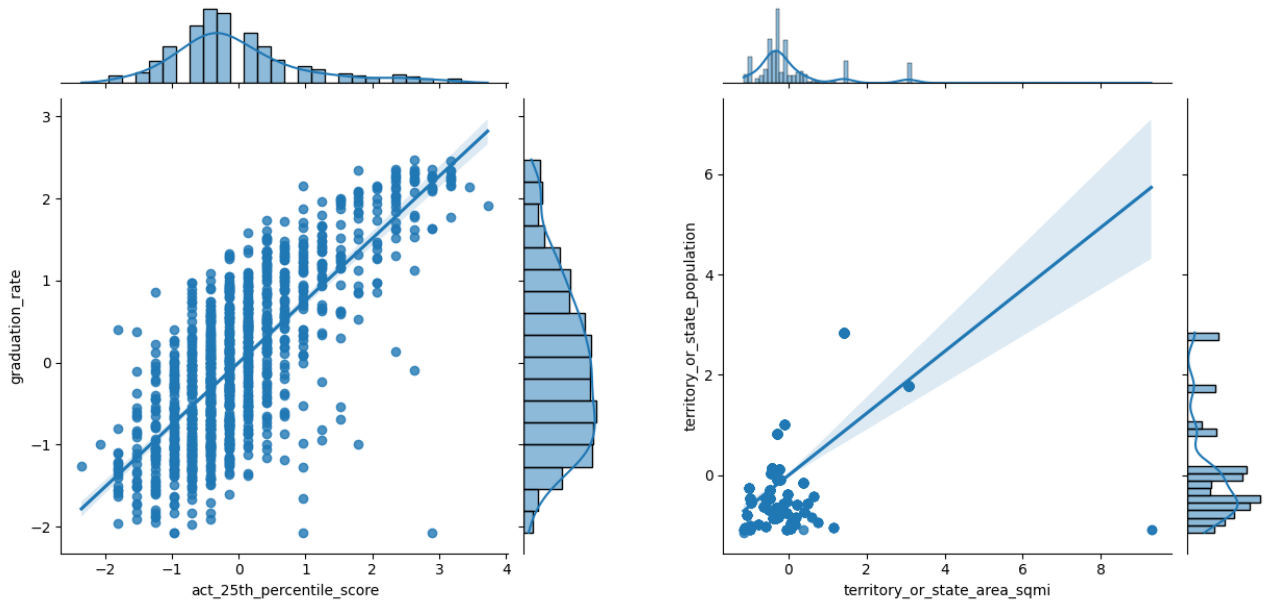


I found it interesting that only the admission_rate distribution is skew-left while the remaining 5 are skew-right. Next, I built a correlation a heat map to investigate possible correlations among all variables.

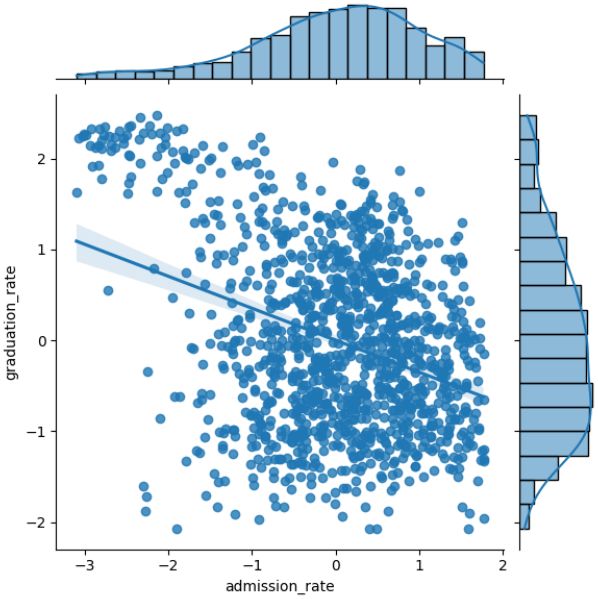


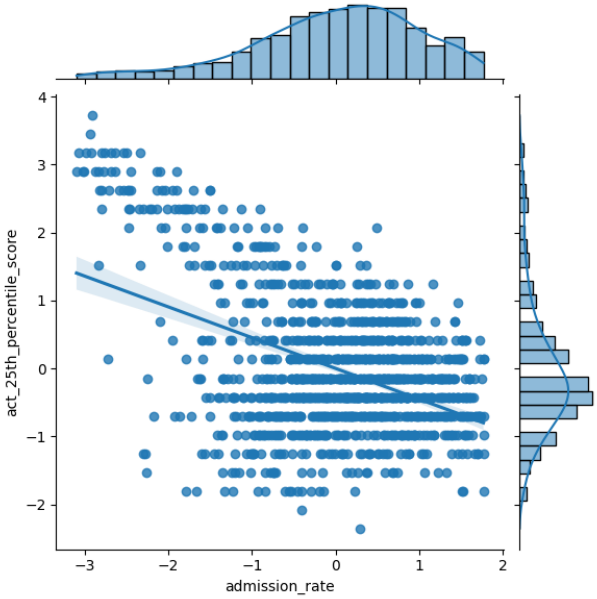
Some notable positive strong and medium/strong correlations that appear include the act_25th_percentile_score and graduation_rate features (0.76), and territory_or_state_area_sqmi and territory_or_state_population features (0.62). The feature act_25th_percentile_score belongs to both positive correlations mentioned. Some notable negative medium/strong correlations include admission_rate and act_25th_percentile_score features (-0.45), and admission_rate and graduation_rate features (-0.35). The feature admission_rate belongs to both negative correlations mentioned.

I used Seaborn's Jointplot function to visualize these correlations in the following plots:

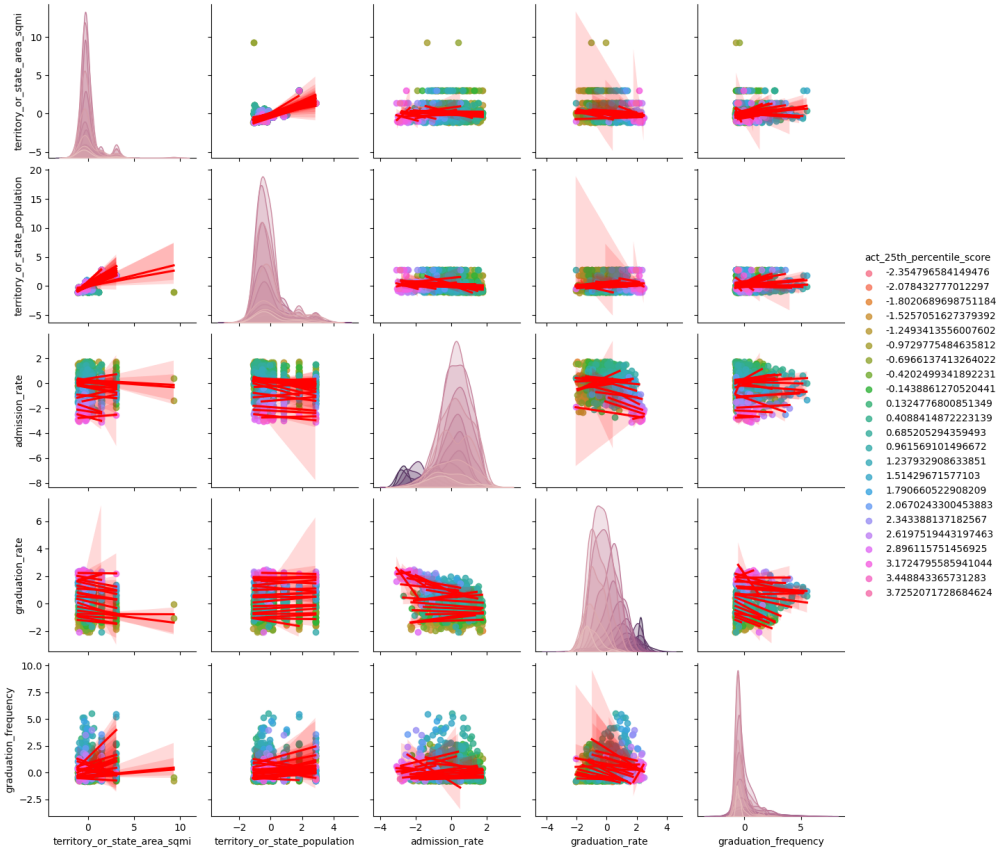


The results from these graphs lead me to some reflection and questions about the data. The strongest positive correlation I found (act_25th_percentile_score and graduation_rate) seems intuitive to me, but I wonder if it's a result from bias during the wrangling process by only including institutions that are predominately bachelor's and graduate degree granting. The second strongest positive correlation (territory_or_state_area_sqmi and territory_or_state_population) is quite unrelated to the problem statement and I don't think it holds any value. The first negative correlation (admission_rate and act_25th_percentile_score) seems reasonable given that ACT scores aren't the sole determining factor of a student's admission to an institution. In recent years, less weight has been put on high-stakes testing for applying students and some institutions are considering or have already dropped the requirement all together. I thought there would be a stronger negative correlation between admission_rate and graduation_rate, but now I can understand that the struggles students encounter during their 4 years at an institution non necessarily related to finishing curriculum (physical / emotional health, monetary issues, family-related challenges) may impact graduation a lot more than the institution's initial selectivity.





To investigate pairwise relationships between act_25th_percentile_score and the remaining 5 features with continuous data I used Seaborn's Pairplot function with a kernel density estimation.



Modeling

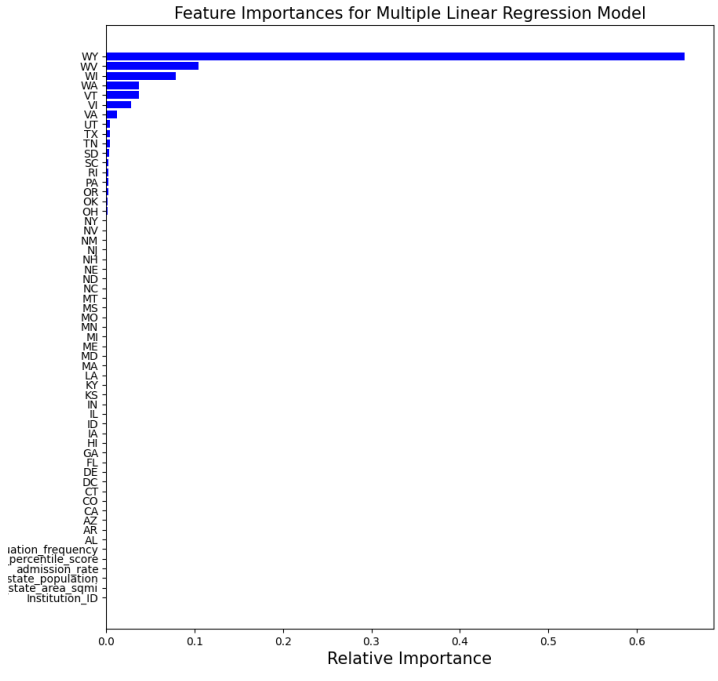
In preparation for modeling, I created dummy variables from the state abbreviation series. This step increased my features to a total of 57. Then, I subset my data into training and tests sets using a 70:30 split. Since my data is labeled and my target variable is a continuous data type, I opted to try 5 supervised regression models including simple linear, multiple linear, support vector machine, decision tree, and random forest. The models that performed best were multiple linear and random forest. In the multiple linear regression model, I achieved at a mean squared error of 0.38 and 64% of the variation in graduation rates explained by the remaining features. In the random forest model, I achieved a mean squared error of 0.39 and 62% of the variation in graduation rates explained by the remaining features.

Since both models are close in performance, I looked into tuning the hyper parameters of the random forest model to see if I could achieve more accuracy. After using SKLearn's GridsearchCV I arrived at the following results:
Best Hyperparameters: {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 200}
Mean cross-validated score of the best_estimator: 0.6626639341411187

After retrying the random forest model with these parameters, I arrived at a mean squared error of 0.399 and 62.27% of the variation in graduation rates explained by the remaining features. While the mean squared error got a tiny bit worse, the coefficient of determination improved slightly.

Conclusions

The multiple linear regression model continues to perform the best. To understand what features had the most impact on the prediction of graduation rates within the multiple linear regression model I used Scikit-Learn's feature_importances_ function and plotted it using a bar chart.



What this has revealed is the dummy variable WY has an extreme effect on graduation rate.

Further Research

Many questions arise from this project and more data is necessary to investigate them further. In the most recent observation, the institutions belonging to WY need to be analyzed. An additional feature designating each institution as private or public may contribute to understanding how each type effects the data distribution. Public and private institutions have different ways of dealing with students who are underperforming and the outcomes of student success may be due to their designation.

Another important issue with the data is the is no tracking of transfer students that join or leave an institution per year. A student who transfers in as a sophomore and graduates on time is counted in the data's graduation rate and graduation volume, but he/she does not contribute a data point to the ACT score or admissions rate. Alternatively, a student who transfers from 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 very likely to complete his or her bachelor's degree on time. However, this student's institution change counts negatively toward the original institution's graduation rate even in the event the student is successful in graduating on time. Tracking these students would help validate the graduation rate.