Determinants of Grad Rates in Public Higher Ed.

Not all spending is created equal for every school

Dennis Rhodes, Garren Wegner, Naveen Gubba

Determinants of Grad Rates

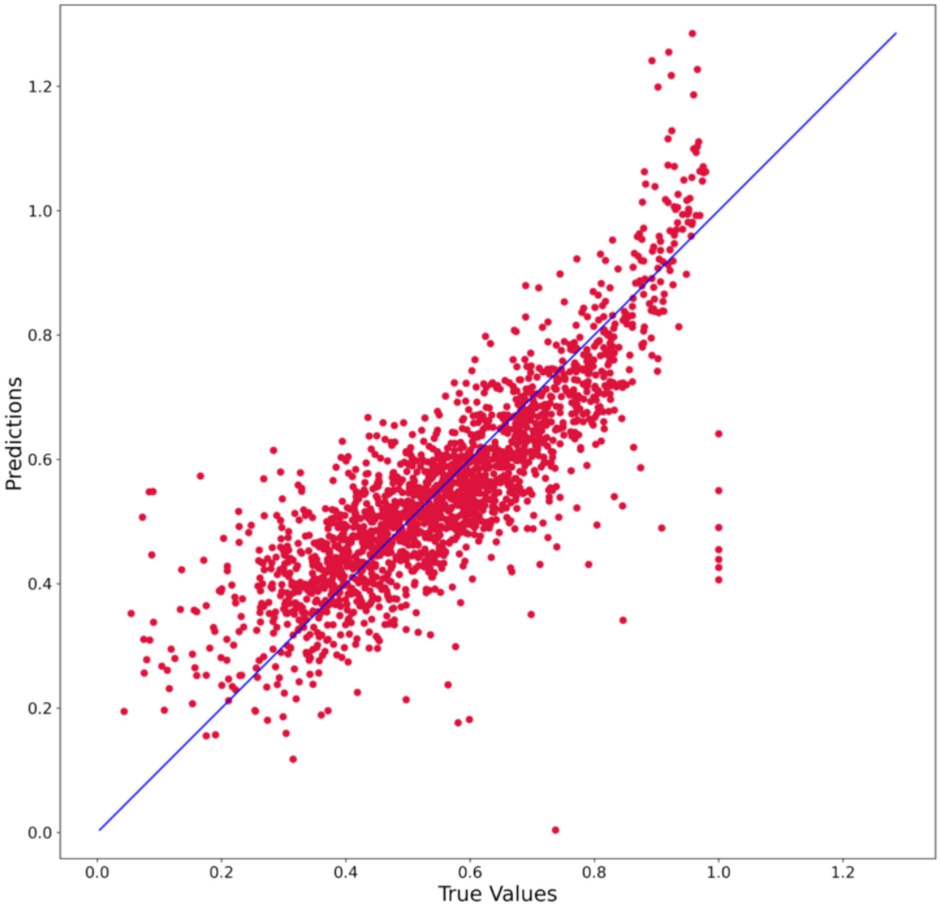
For our master's Capstone in Data Science & Analytics, we chose to explore the data surrounding higher education organizations, the spending patterns of public higher education organizations awarding at least baccalaureate degrees, and their graduation rates. We wondered what the determinants of graduation rates might be and how public dollars might best be spent to increase graduation rates. This question is particularly salient in this period of lower enrollment and a generation of students at least one grade level behind due to the pandemic. Some of our findings and conclusions include:

- Many of the best predictors of graduation rates are related to the academic preparation of students.
- Public funding processes that prioritize graduation rates without considering measures of accessibility create incentives for colleges and universities to restrict access to less academically prepared students.
- Reducing access to higher education would shut out disadvantaged populations whom public higher education organizations are supposed to serve.
- Models excluding selectivity measures perform poorly in the prediction of graduation rates, even considering a multitude of other factors.
- Not all spending is equally effective at raising graduation rates and certain Carnegie Class-Organization Structure combinations show different correlational relationships with different types of spending.
    - Instructional and Academic Support Spending at Research Universities is positively correlated with graduation rates, especially for schools in multi-organizational systems.

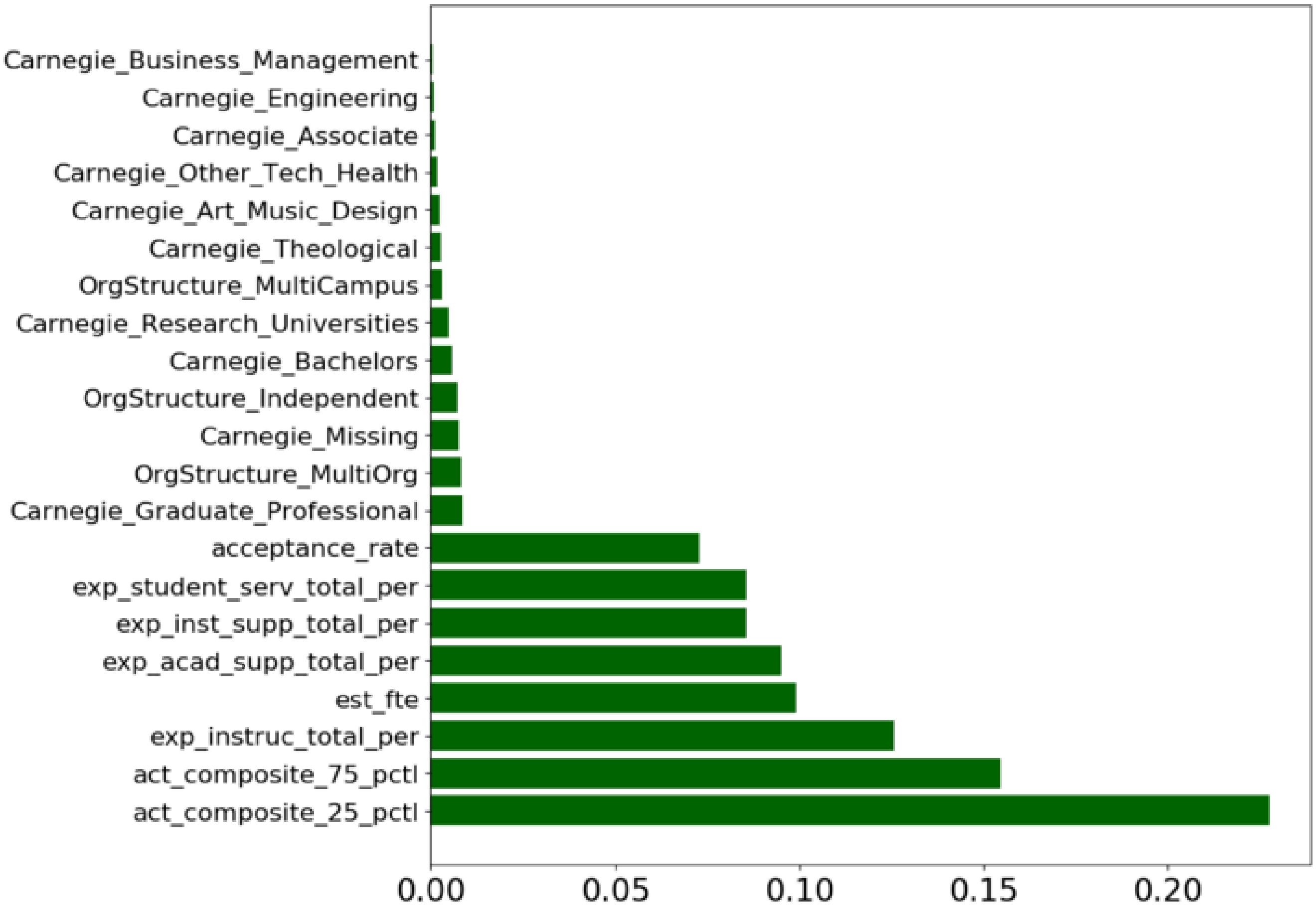
    - Student Services spending at Baccalaureate Arts & Sciences Colleges and Research Universities with High Research Activity in independent and multi-organizational structures is positively correlated with graduation rates.
    - Institutional Support spending at Doctoral and Master's (smaller program) Colleges in multi-organizational systems is also positively correlated with graduation rates.
- Further research exploring causal relationships between these Carnegie Class-structure combinations would be highly beneficial.

Predictors & Perverse Incentives

Predicting graduation in public higher education is a fraught topic with serious consequences. Many states have implemented performance-based funding mechanisms that penalize schools with low graduation rates and/or incentivize increases in graduation rates. Using selectively measures, Carnegie classification, institutional spending, and location of the school, we were able to build a linear regression model with a score of .724 (within the .7-.9 industry standards) for predicting graduation rates that explains 68% of variance and has an R^2 of .68 that can be seen below.



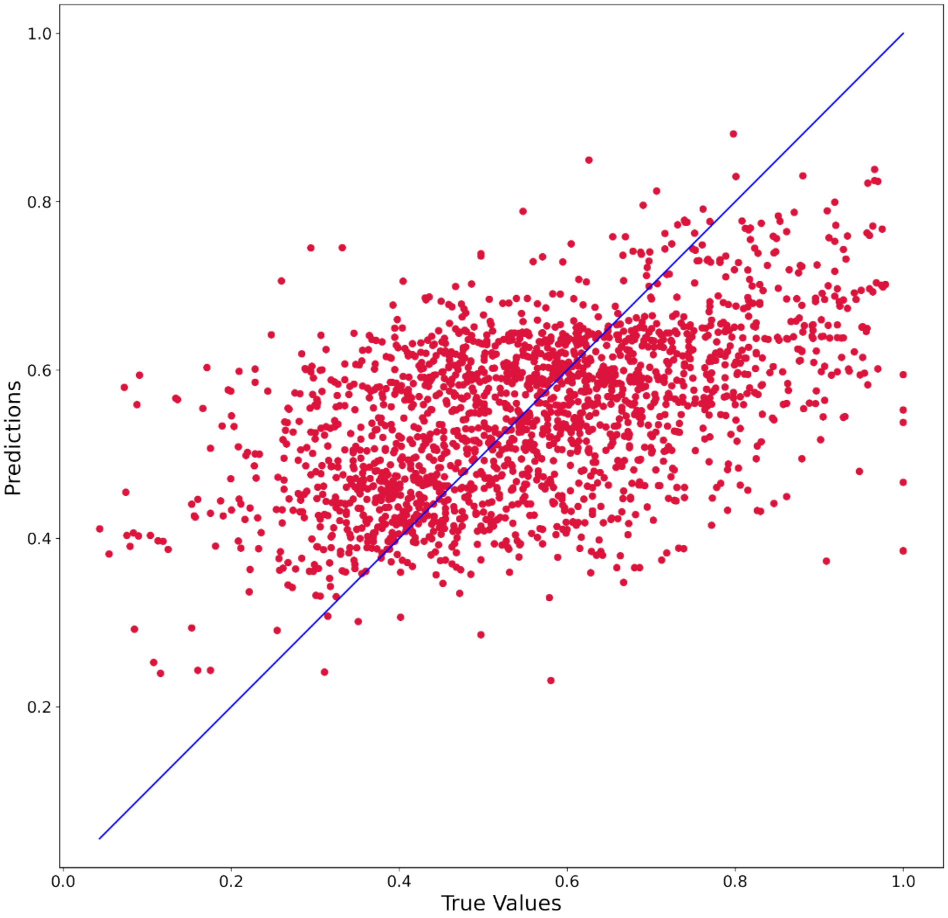
While the intentions are doubtlessly good, performance-based funding mechanisms can create some perverse incentives. As shown in the figure below, the top two predictors of graduation rates are the ACT scores at the 25th and 75th percentiles.



While schools find it difficult to increase spending for things like instructional expense per student, it is far easier for them to tighten admissions standards and play the incentive system by blocking out students less likely to succeed. This risk adverse behavior is counter to the mission of most public organizations and to the public interest. Even adjusting for things like school quality, students of color score at least 1.5 points lower on average on the ACT. Students from rural communities, poorer communities, and communities with private primary and secondary school options also tend to have lower quality schools and less prepared graduates. If public education is intended to be an engine of economic and social mobility, then these sorts of policies are counterproductive.

Non-selectivity Determinants

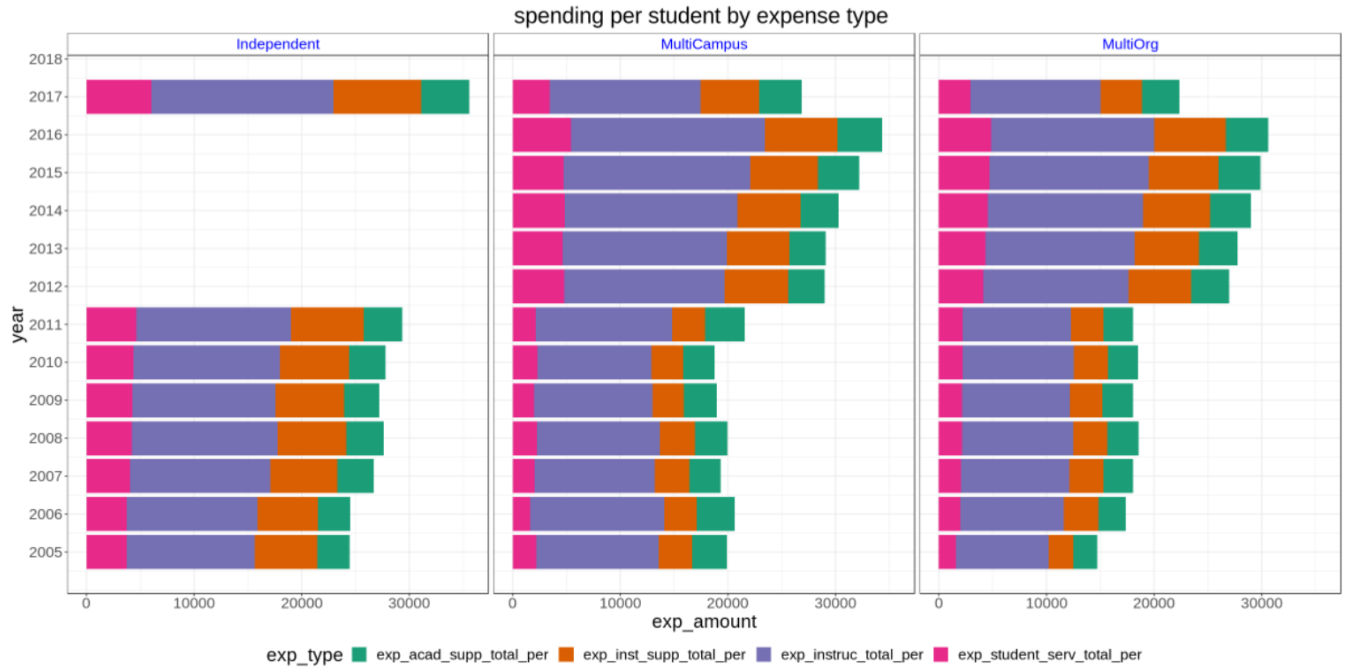
That being the case, we sought to develop a model that would predict graduation rates without considering selectivity measures like ACT scores and admissions rates. We developed a PCA linear regression pipeline that processed our entire dataset less the selectivity measures which produced a score of only .29 explaining 27% of variance with a .27 R^2 value which can be seen below.

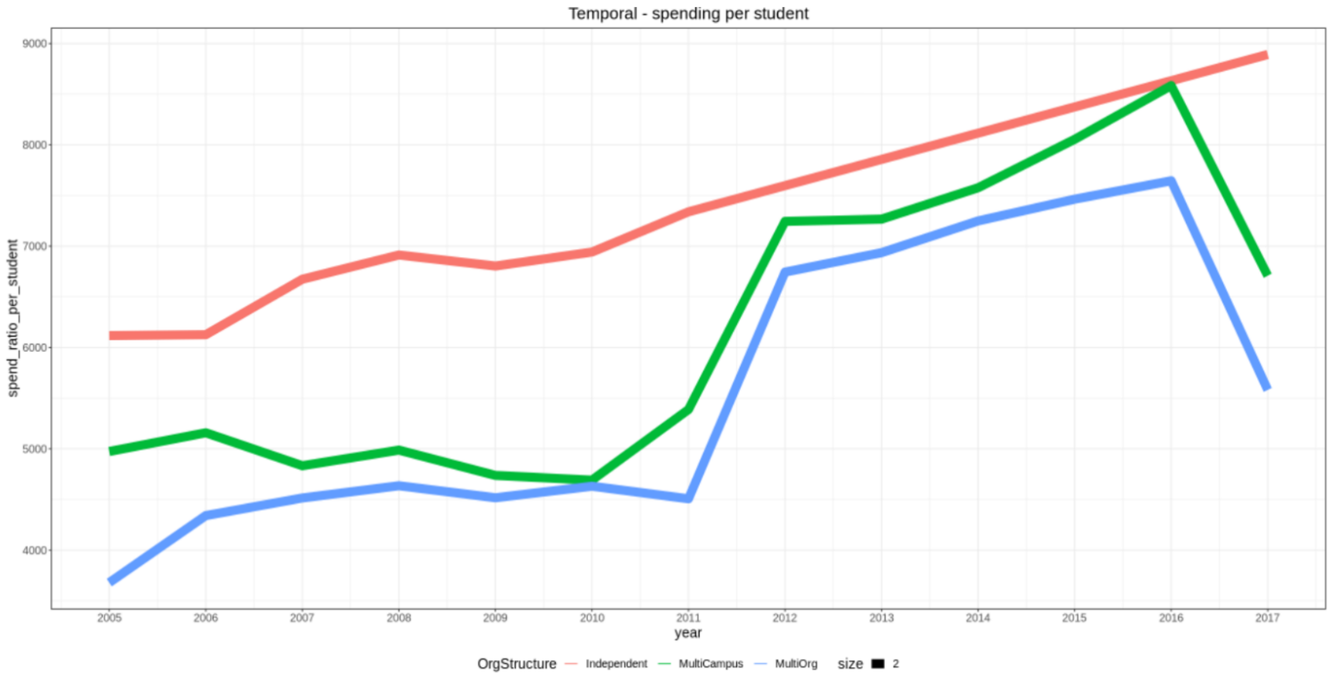


Since all spending variables were included in the PCA and the best model produced by the pipeline performed very poorly, it can be concluded that they are not particularly strong predictors of graduation rates across all institutions. By extension, the problem is unlikely to be resolved by simply throwing money at the sector without considering the interaction effects we begin to explore below.

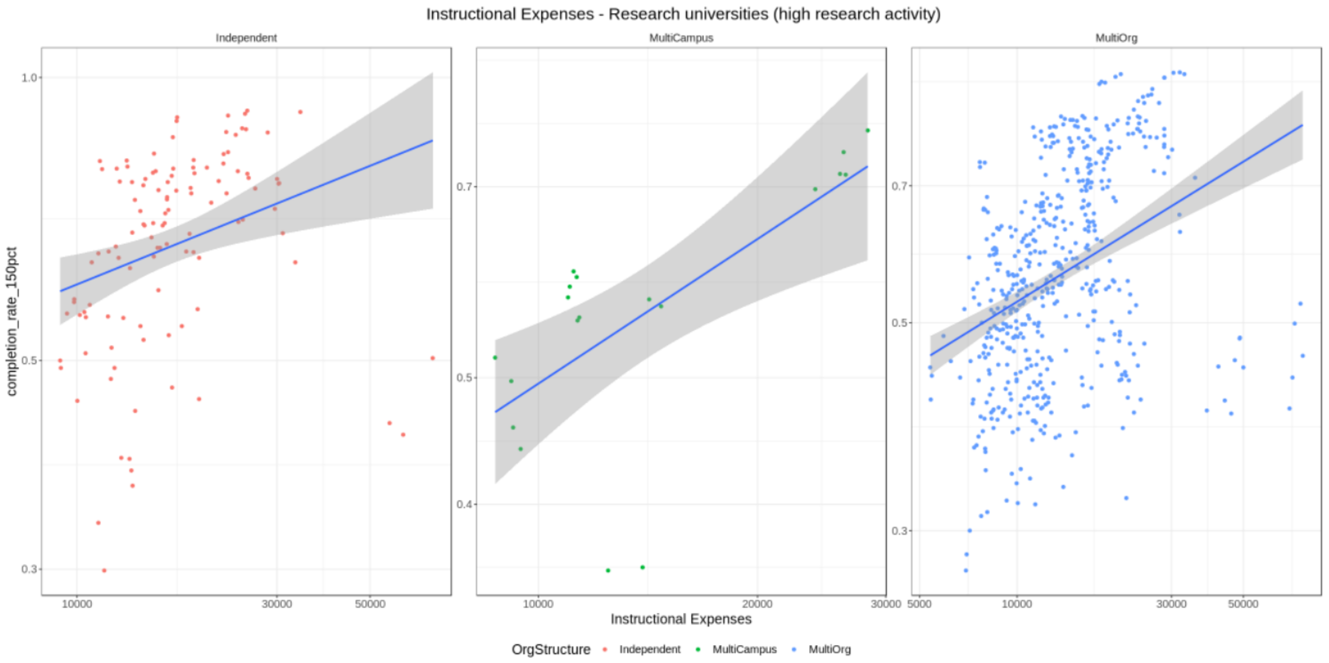
Not All Spending Is Created Equal for Every School

There are four major spending categories that the literature has connected to graduation rates: academic support, institutional support, instruction, and student services. A breakdown of CPI-indexed average spending per student per year per spending category for each of the three organization structures can be seen below along with a multi-line graph showing total spending per student for each organization structure over time. Variation across years can be partially attributed to the size of graduating high school classes and enrollment level variation across the sector.



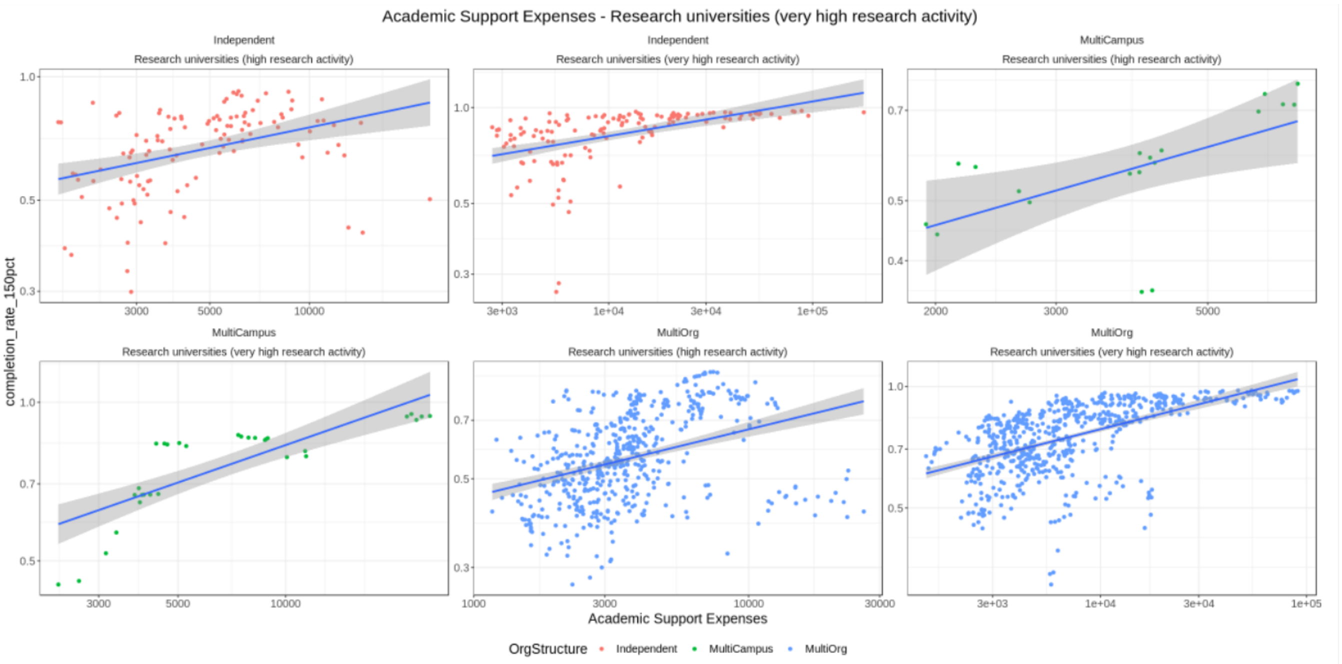


We eventually found intriguing relationships between spending at some school class / organization structure combinations and specific types of spending. The first of these was a strong positive relationship between instructional expenses at Research Universities with High Research Activity. While the trendline is similar across all three structures, it is particularly compelling in the case of multi-organizational systems, as seen in a logarithmic transformation of spending below.



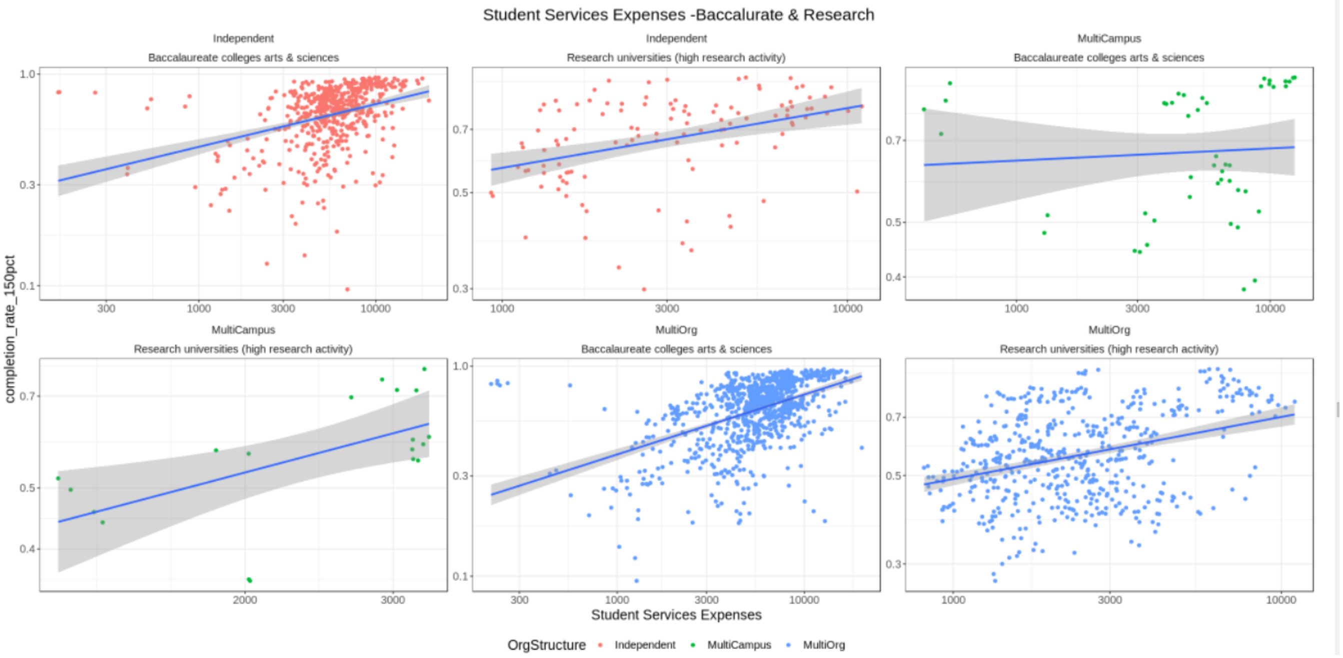
This relationship makes a lot of sense – Research Universities tend to have high student-faculty ratios and faculty incentive (tenure) systems that prioritize research over teaching. This relationship seems to hold only until around the $30,000 per student per year mark, indicating that there may be diminishing returns to this type of investment. Nevertheless, these results do appear to establish a strong correlational if not causal relationship between instruction spending and graduation rates at public Research Universities with High Research Activity that policymakers and administrators should seek to explore further. Perhaps surprisingly, this relationship was not apparent at any other class/structure combination.

We observed similar relationships between Academic Support Expenses (which include things like librarian and academic technologist salaries) and graduation rates at Research Universities with High and Very High Research Activity, as seen below.



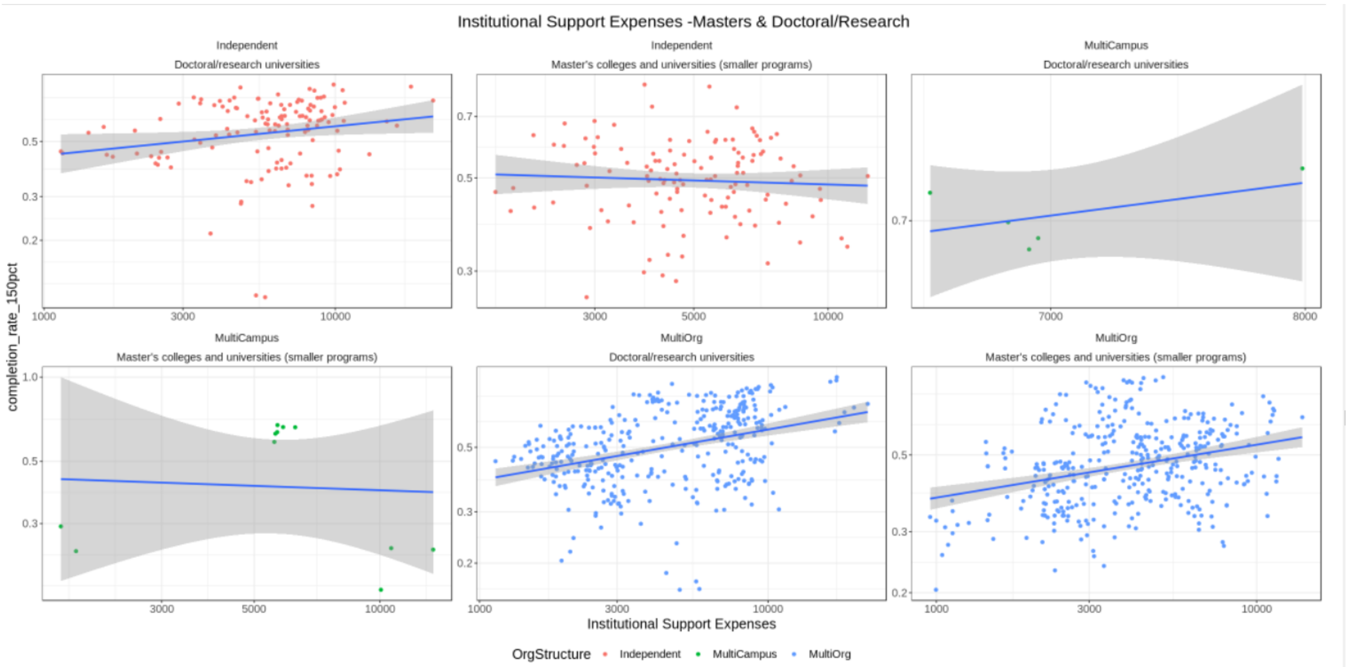
These relationships also make sense: larger class sizes require organizations to offer additional support outside the classroom for students to feel supported and remain engaged. The research-intensive focus of these organizations also typically result in faculty adopting researcher-practitioner program philosophies instead of more practice oriented philosophies at organizations with faculty more focused on instruction. As a result, librarians, technologists, and writing tutors become increasingly valuable to students. Notably, the cutoff in spending level impact appears far lower here than in instruction expenses at less than $10,000 per student per year, making further exploration of these relationships important for further exploration for policymakers and administrators.

Student Services spending (for things like career advisor and student affairs professional salaries) also appeared more relevant to some structure/class combinations than others. Liberal Arts Baccalaureate Colleges and Research Universities with High Research Activity in independent and multi-organizational structures appear strongly related, as seen below.



The explanation for these relationships is less apparent for these types of organizations. It might be attributable to the role of these professionals in connecting theory-based courses to the real world; unlike more practice focused Masters Colleges and Universities, these classes of schools are often seen as focusing more on theoretical foundations better suited to preparation for advanced study than immediate professional application. Similar to Academic Support expenses, the relatively low impact cap, below $10,000 per student per year, makes this area ripe for additional research.

Institutional Support, the final category of expense, appears to only have strong apparent relationships for Doctoral and Master's (smaller program) Colleges in multi-organizational systems, as can be seen below.



These relationships may reflect the cost and benefits of coordinating the activities of multiple organizations with independent faculties that can specialize and attract students interested in those specialties. Rather than having mediocre engineering programs at multiple organizations, for instance, students interested in engineering may be more successful in a specialized organization dedicated to engineering. These relationships are particularly intriguing as excesses in spending on university administration is regularly lambasted in popular media, yet it appears that the spending may not always be as wasteful as popularly perceived.

A Methods section detailing our process is included below.

Machine Learning Methods

The Data

Source

We leveraged an R package developed by the Urban Institute's Education Data Portal to access data collected by the National Center for Education Statistics and stored in the Integrated Postsecondary Education Data System (IPEDS) and the College Scorecard. All higher education organizations receiving Title IV federal funding are required to complete a 12-part survey each year which populates the data in IPEDS. We harvested the data from five different IPEDS tables including Finance, Graduation Rates, Enrollment, Institutional Characteristics, and Institutional Directory. We selected the 150% completion rate as our measure of graduation because it was the most complete metric and few students who have not finished an undergraduate degree within six years are likely to at all. From College Scorecard we gathered measures of selectivity including ACT aggregate scores and application, admissions, and enrollment figures for each school. Our initial dataset consisted of 2,813,686 observations and 368 variables. We reduced the columns to fields relevant to the question and removed redundant columns found in both tables. We also reduced the rows by only including institutions that offered a bachelor's degree. This brought the dimensions down to 98,407 observations and 64 columns.

Data Obstacles

Initially, we were also interested in exploring the dispersion of institutional support (largely administrative) spending across constituent units of complex organization structure. This analysis would have compared the three different school classes, Multi-Campus Organizations such as Penn. State, Multi-Organizational Systems, such as University of Missouri, and Independent Schools such as University of Central Missouri. However, we soon realized that NCES reporting standards do not require complex structured organizations to report financial data with their constituent units, so a cross-structure comparison of constituent units is not currently feasible. Another issue that became readily apparent was the absence of data for certain fields. This was particularly true for graduation rates among independent schools which had null value reports for several years, as can be seen in Figure 1.

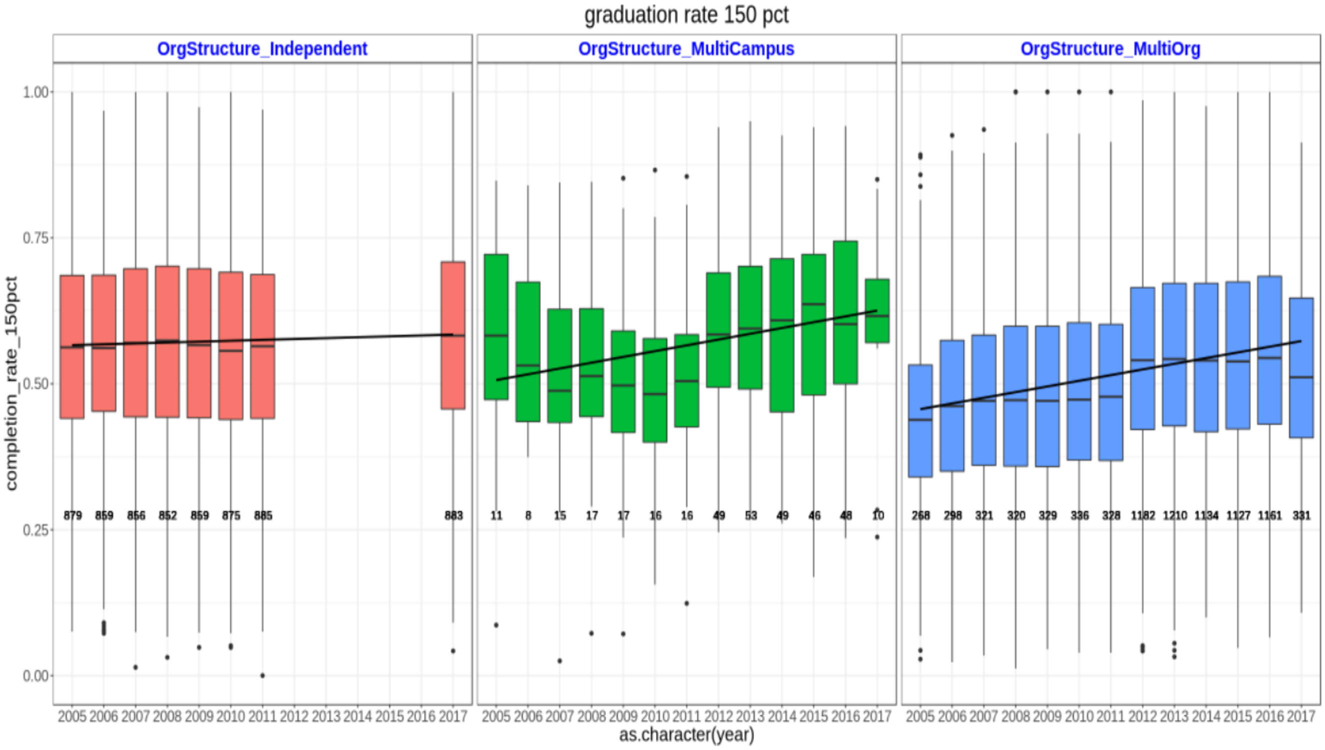


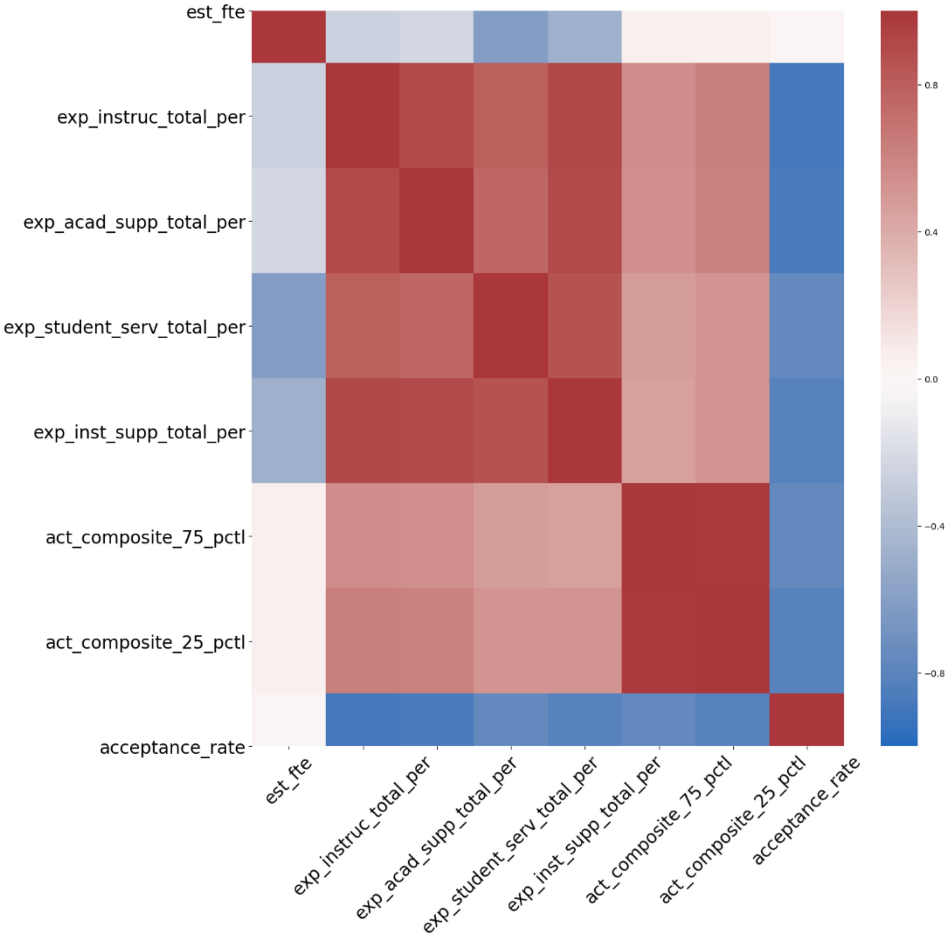
Figure 1

Regression and Classification

Multicollinearity

Our first step was to binarize various categorical factors. Once the variables were encoded, we Min-Max-Scaled the numeric values for machine-learning. As part of our initial data exploration, we considered the correlation between variables to ensure our methods accounted for multicollinearity. Although the majority of our variables did not have strong correlation, a truncated correlation matrix showing those with high levels of correlation can be seen in Figure 2. The high levels of correlation between variables which we anticipate having discriminate power in the prediction of graduation rates prompted us to consider Principal Component Analysis (PCA) for several of our models. PCA is a statistical method suited to cases when multicollinearity is a concern.

Figure 2



Feature Exploration

Next, we began exploring feature importance using a basic random forest model. Figure 3 shows that the two most important factors in predicting graduation rates are the 25th and 75th percentile ACT scores. The next seven factors included estimated full-time equivalent enrollment, three expense categories, acceptance rate. We observed a considerable decline in feature importance after the eighth variable, so we created a trimmed dataset consisting of just those features (Trimmed Set, 8 features) in addition to the Full Set (70 features). Next, we decided to convert our prediction problem to a series of increasingly challenging classification problems.

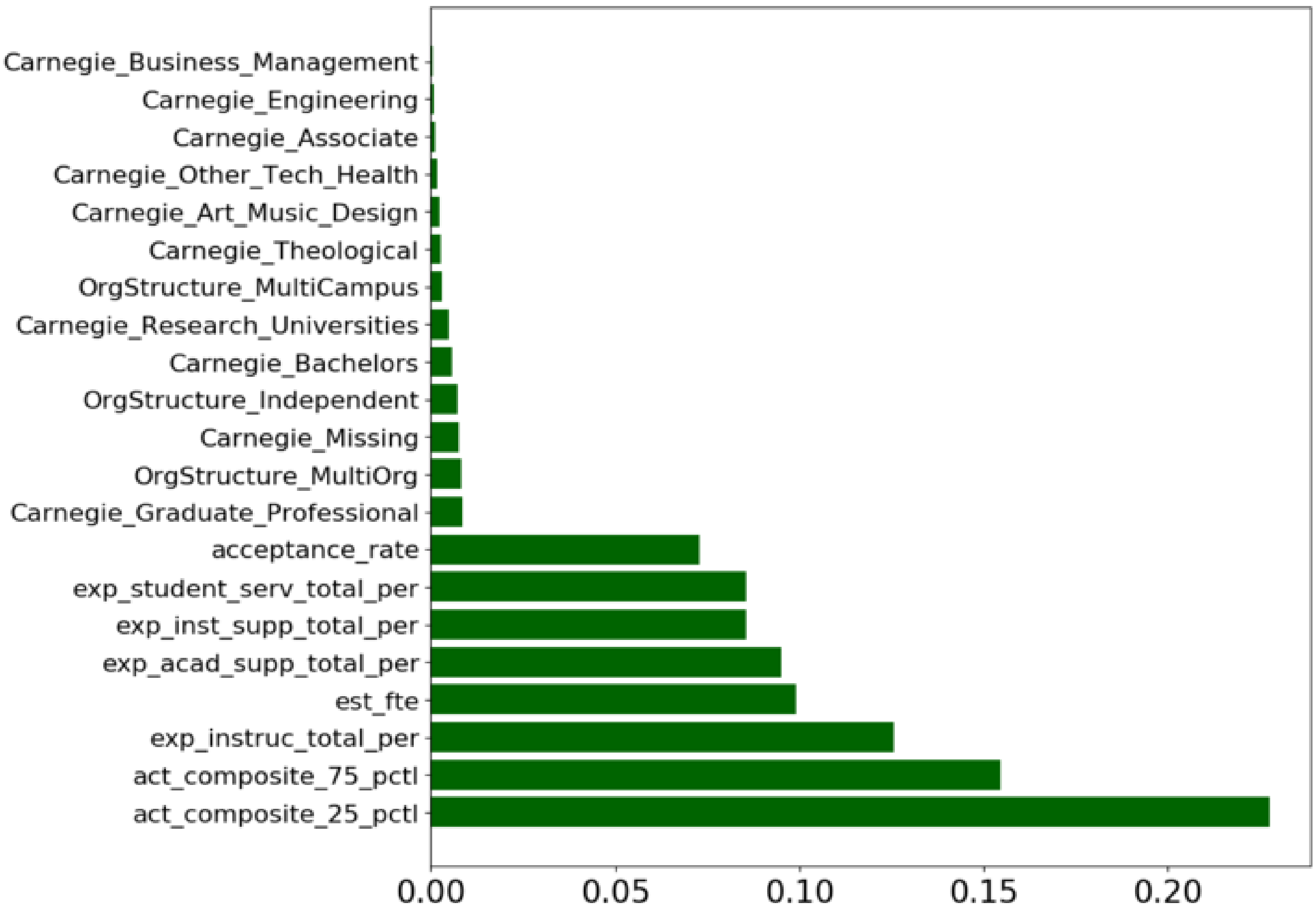
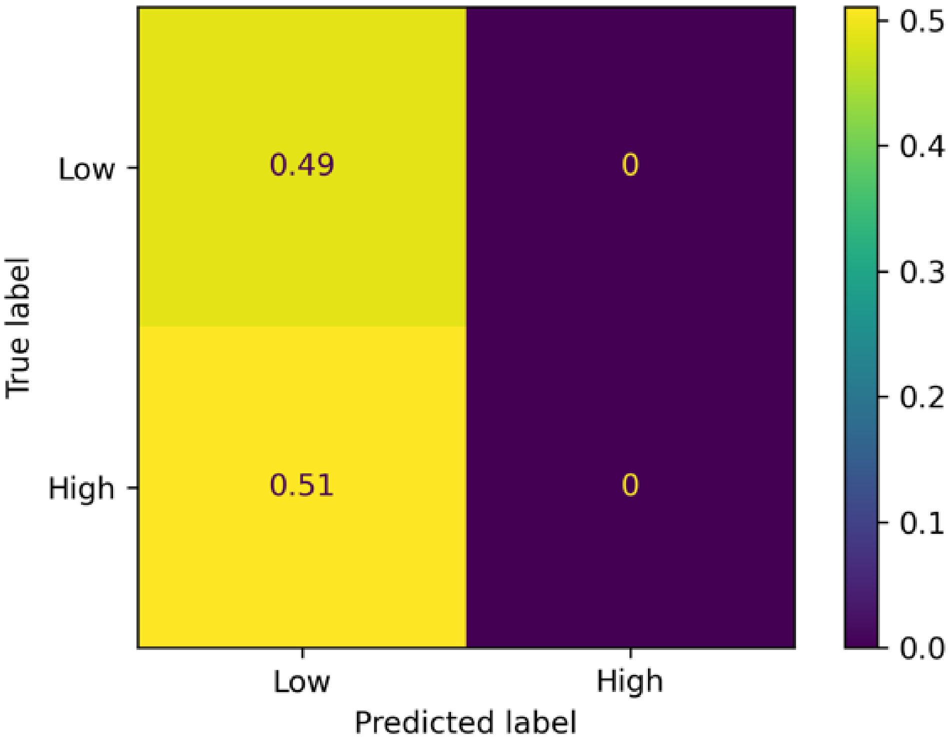


Figure 3

Classification Using Selectivity Measures

To create classification categories, we used the normalized graduation rates to first create a Low/High classification target feature with a graduation rate cut-off of 0.5. Then we created a Low/Medium/High classification target feature with a graduation rate cut-off of 0.33 and 0.66. Lastly, we created a Quartile classification target feature with a graduation rate cut-off of 0.25, 0.5, and 0.75. For each classification exercise, we ran a pipeline selecting the best of four dummy classifier models to compare our machine learning models against as a baseline. We then compared the baseline against the output of a Random Forest Classifier-PCA pipeline run over the Full Set, a Random Forest Classifier pipeline run over the Trimmed Set, and a Decision Tree Classifier pipeline run over the Trimmed Set. Brief analyses, data tables, confusion matrices, and random forest diagrams can be seen under each classification heading below. Data tables include model scores and Area Under the Receiver Operating Characteristic Curve (AUC) scores.

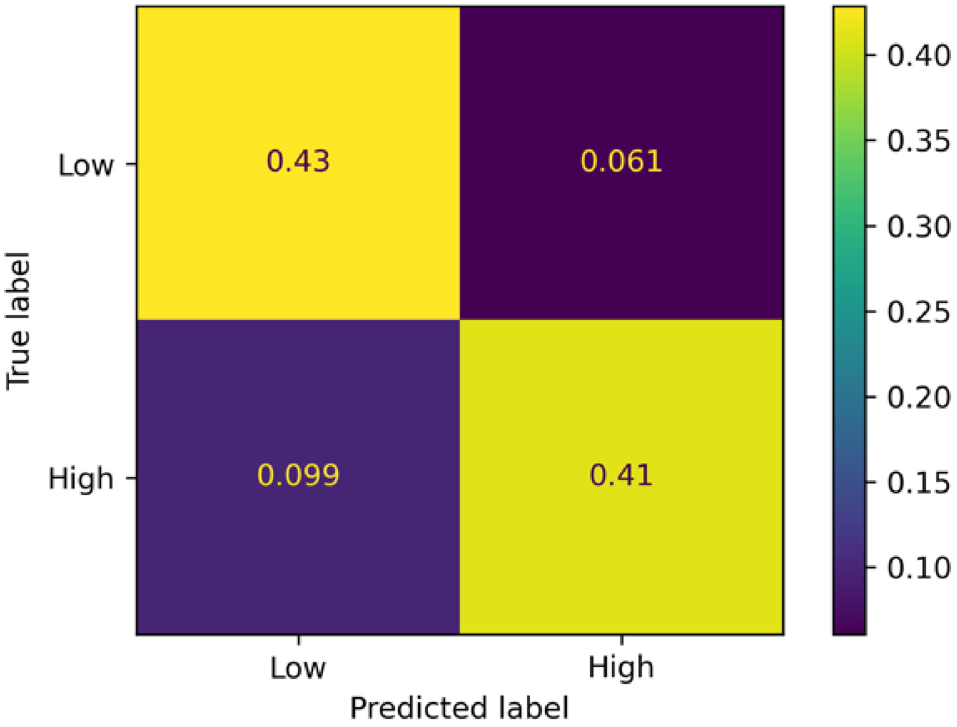


Two Classes

Unsurprisingly, our models all achieved over 80% accuracy with strong AUC scores and our best model, Random Forest Classifier + PCA, had roughly equivalent levels of correct prediction in both categories and equivalent levels of error between both categories.

Model	*Model Score*	*AUC Score*
Dummy	.513	.500
RFC-PCA	.842	.841
RFC	.838	.842
DTC	.817	.820

Figures 4 (dummy classifier confusion matrix, above) & 5 (Random Forest with PCA confusion matrix, below)



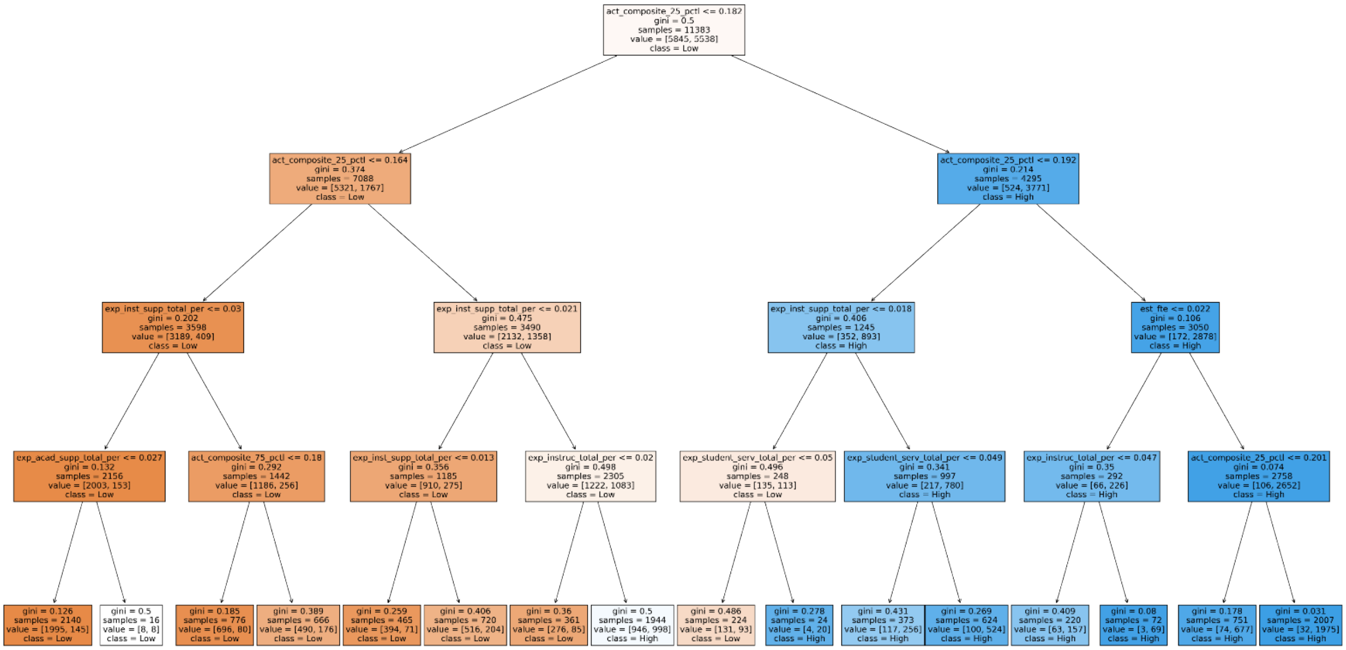
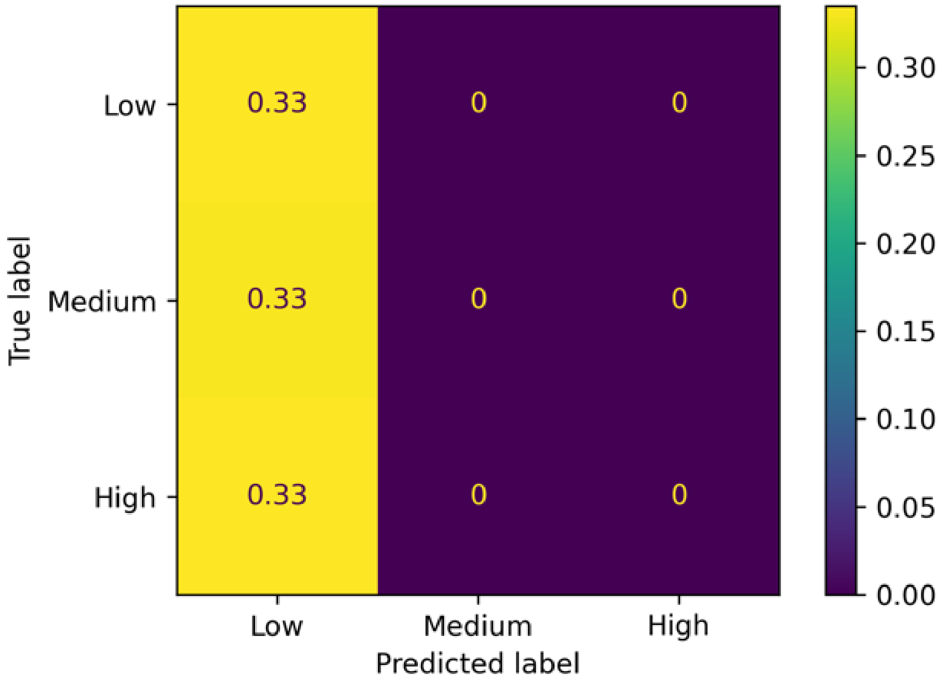


Figure 6, Random Forest Diagram

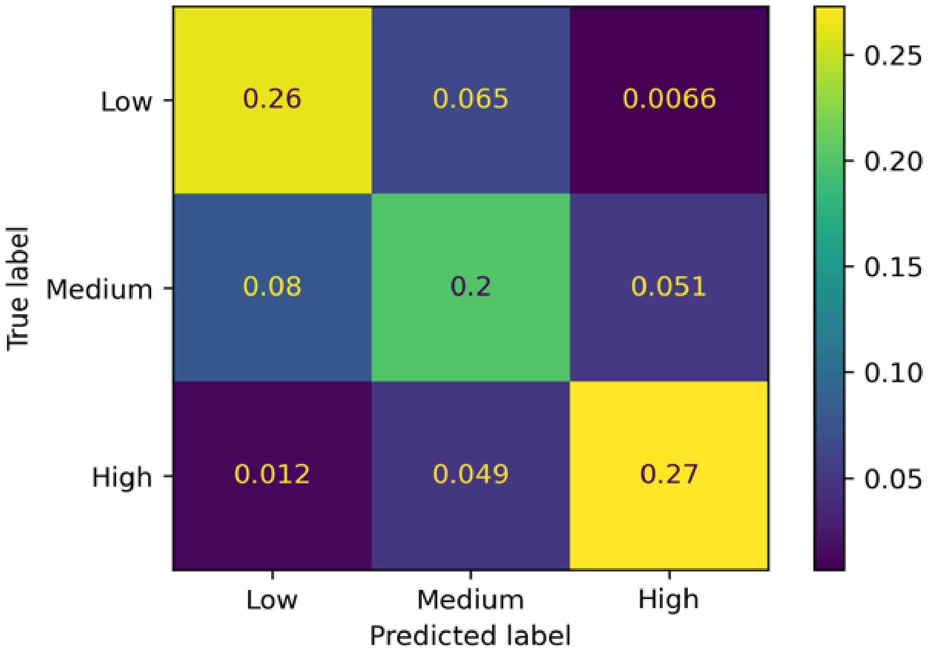
Three Classes

Each of our models performed close to the 70% mark, the lower bound of the acceptable model standards, though the Decision Tree Classifier performed below that standard. Our best model, RFC-PCA, had roughly equivalent levels of correct prediction in each category and equivalent levels of error between each error category. All three models produced scores roughly doubling that of the dummy classifier.

Model	*Model Score*	*AUC Score*
Dummy	.348	.500
RFC-PCA	.732	.889
RFC	.719	.884
DTC	.680	.853



Figures 7 (dummy classifier confusion matrix, above) & 8 (Random Forest with PCA confusion matrix, below)



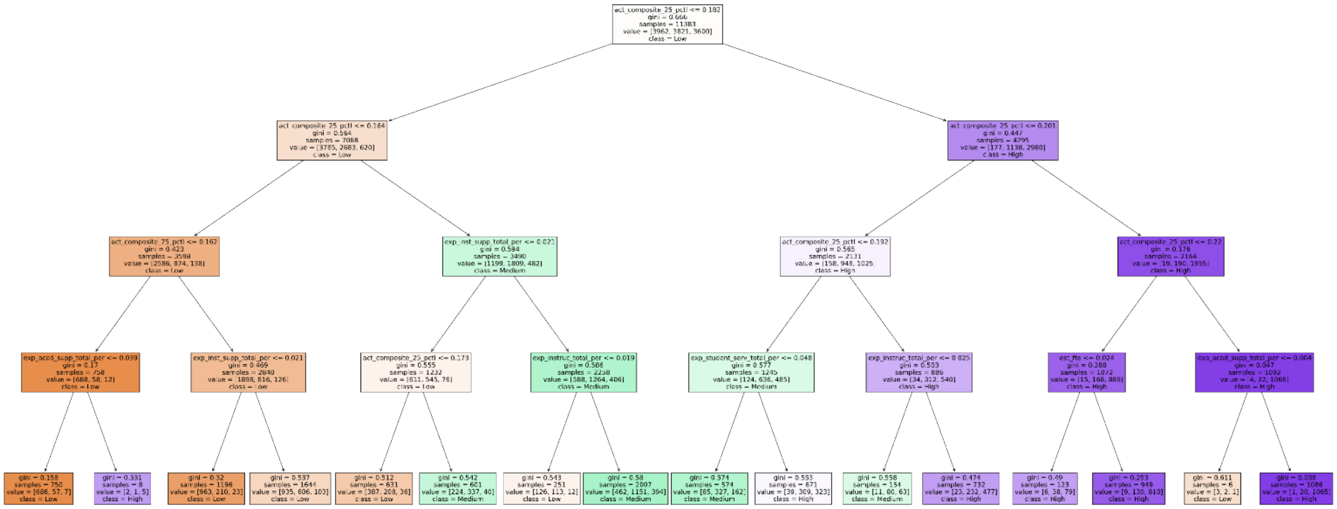


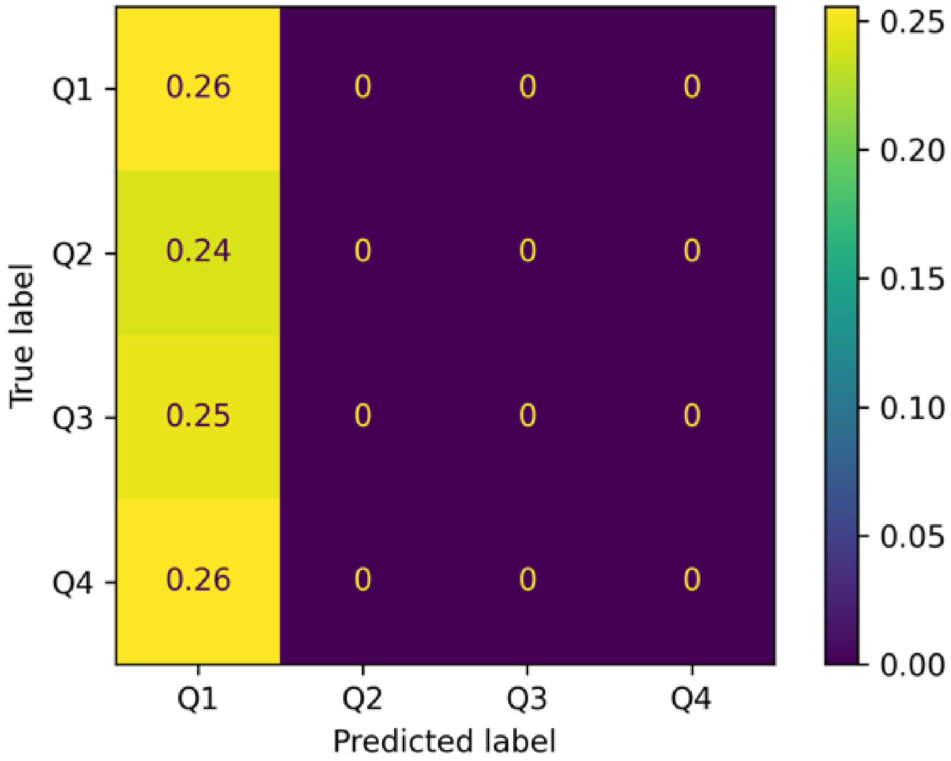
Figure 9, Random Forest Diagram

Four Classes

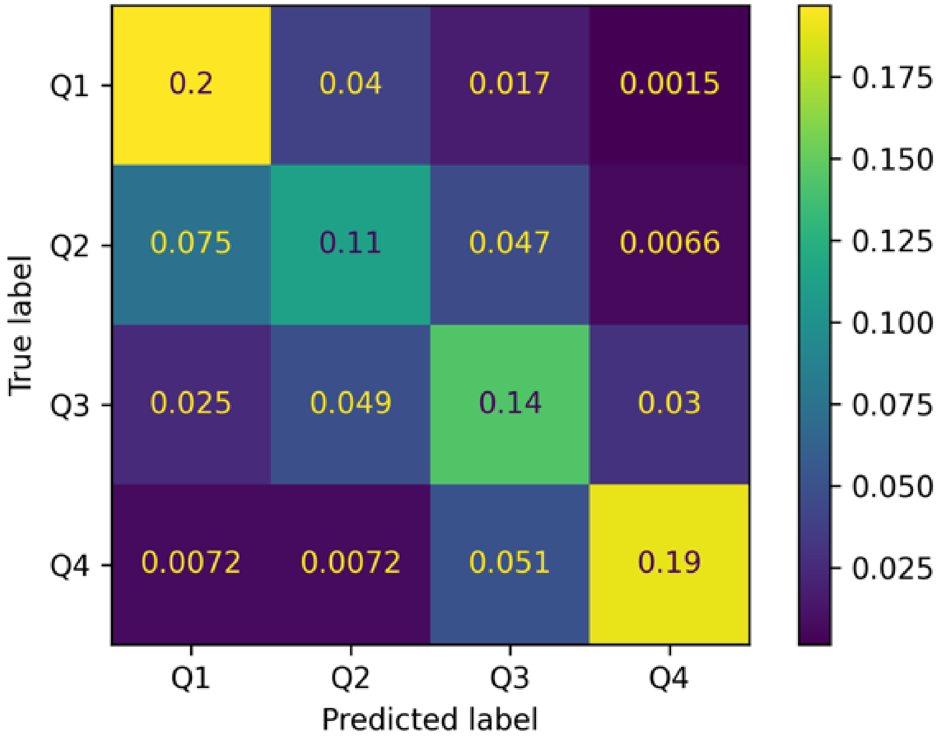
Each of the four-class classification models fell considerably below the acceptability threshold, though they each more than doubled the dummy classifier. The RFC-PCA, the best of the three prediction models, struggled to differentiate organizations in the inner quartiles but was relatively accurate in predicting the extremes.

Model	*Model Score*	*AUC Score*
Dummy	.268	.500
RFC-PCA	.655	.874

RFC	.628	.868
DTC	.587	.839



Figures 9 (dummy classifier confusion matrix, above) & 10 (Random Forest with PCA confusion matrix, below)



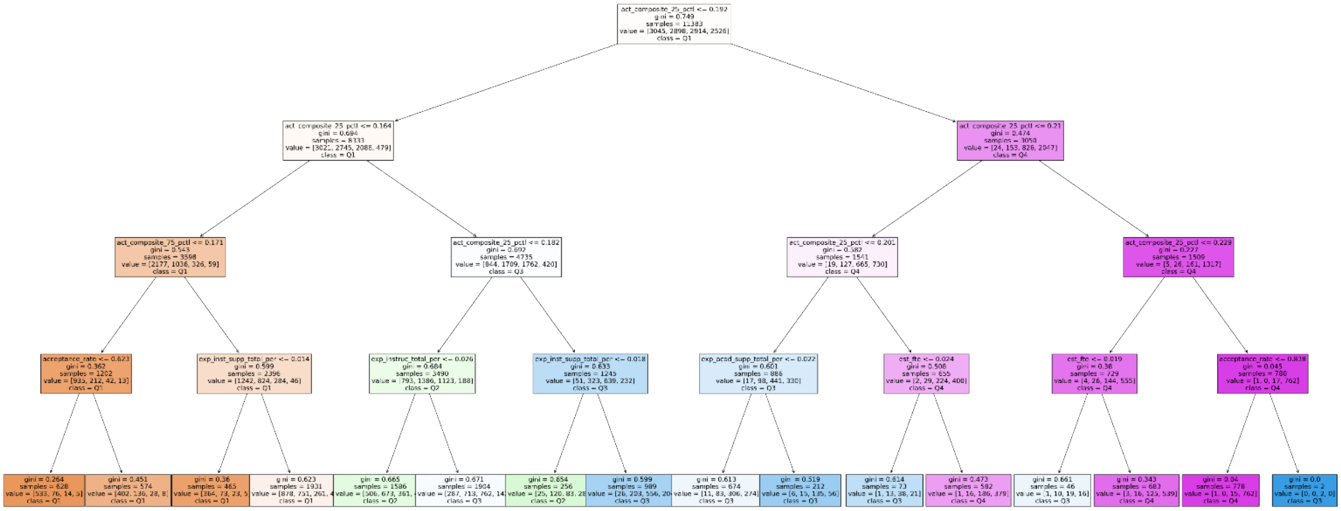


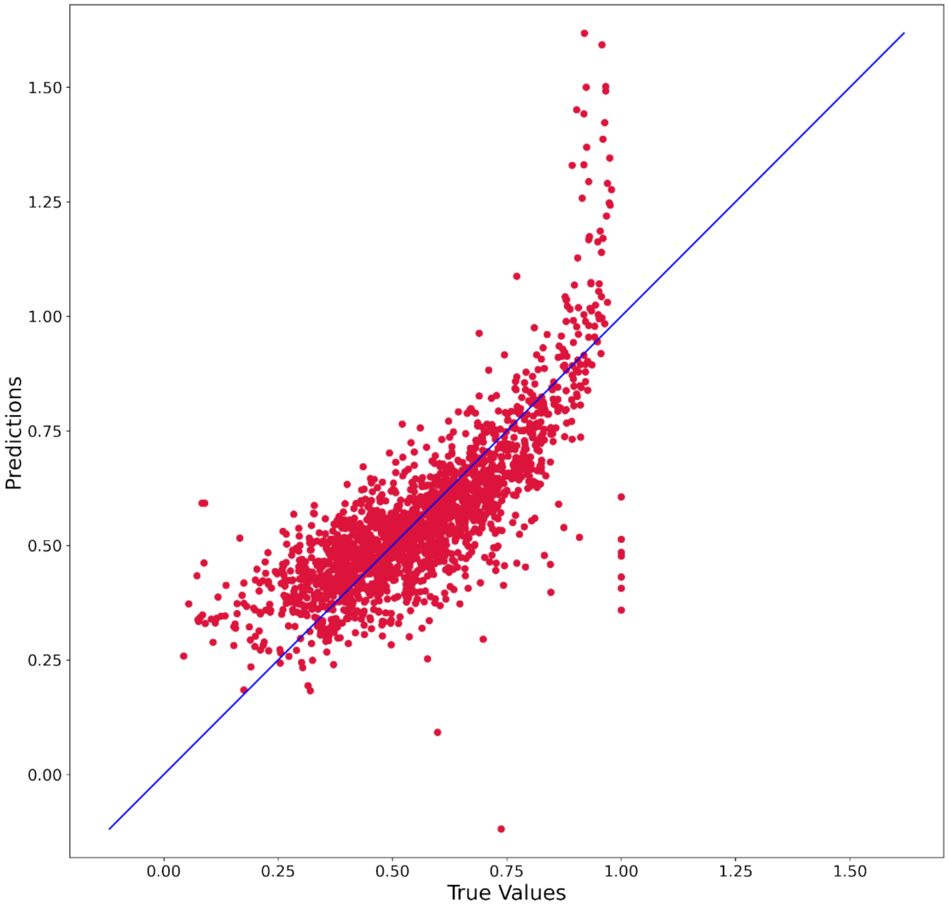
Figure 11, Random Forest Diagram

Prediction Using Selectivity Measures

Moving on from classification to prediction, we ran a linear regression on the Trimmed Set and a PCA-linear regression on the Full Set. A data table and regression plots can be seen below. Our PCA model achieved higher accuracy and explained variance, as can be seen in Figure 13. We further explored the contributions of each variable to the prediction in a Sequential Forward Selection Plot, visible as Figure 14.

Model	*Model Score*	*R^2*
Linear	.640	.561
PCA-Linear	.724	.684

Figures 12 (Linear Regression, Above) & 13 (PCA-Linear Regression, Below)



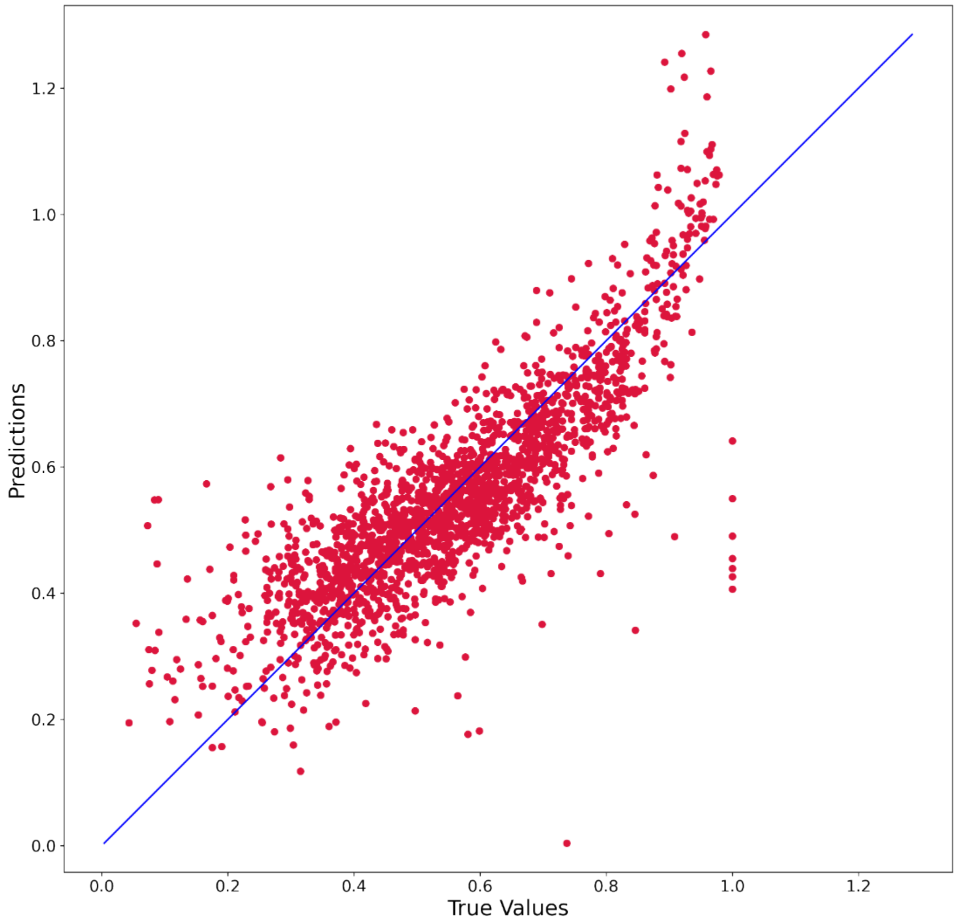
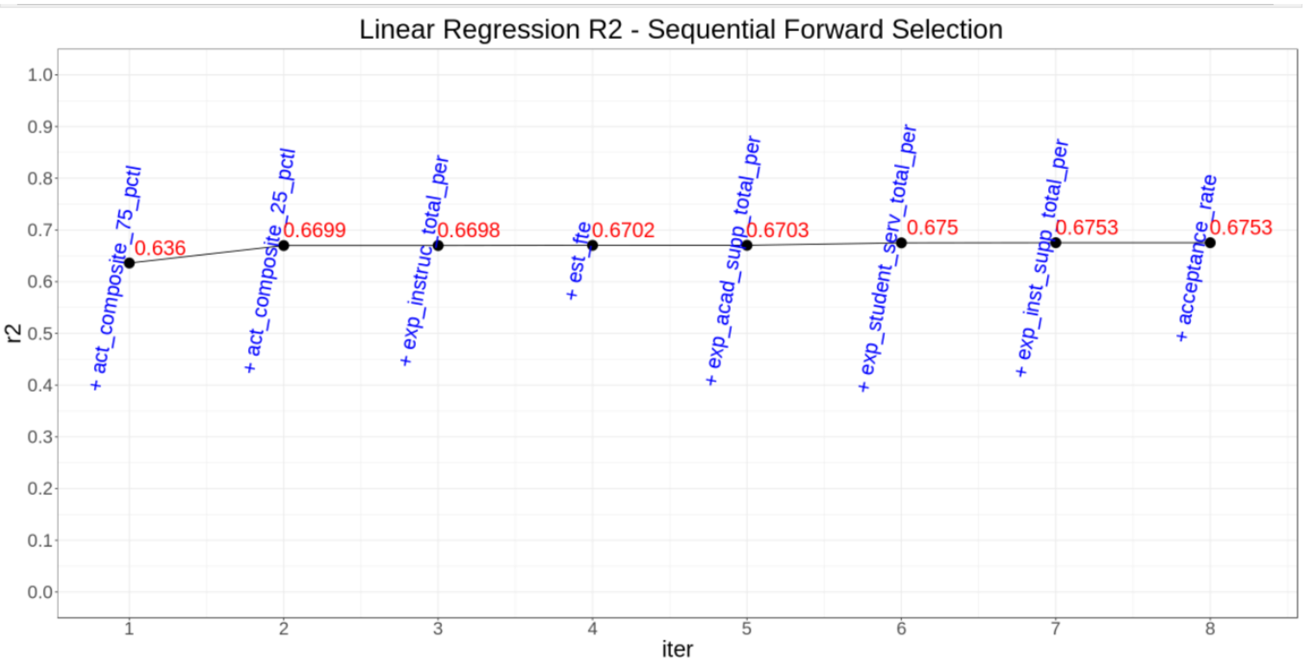
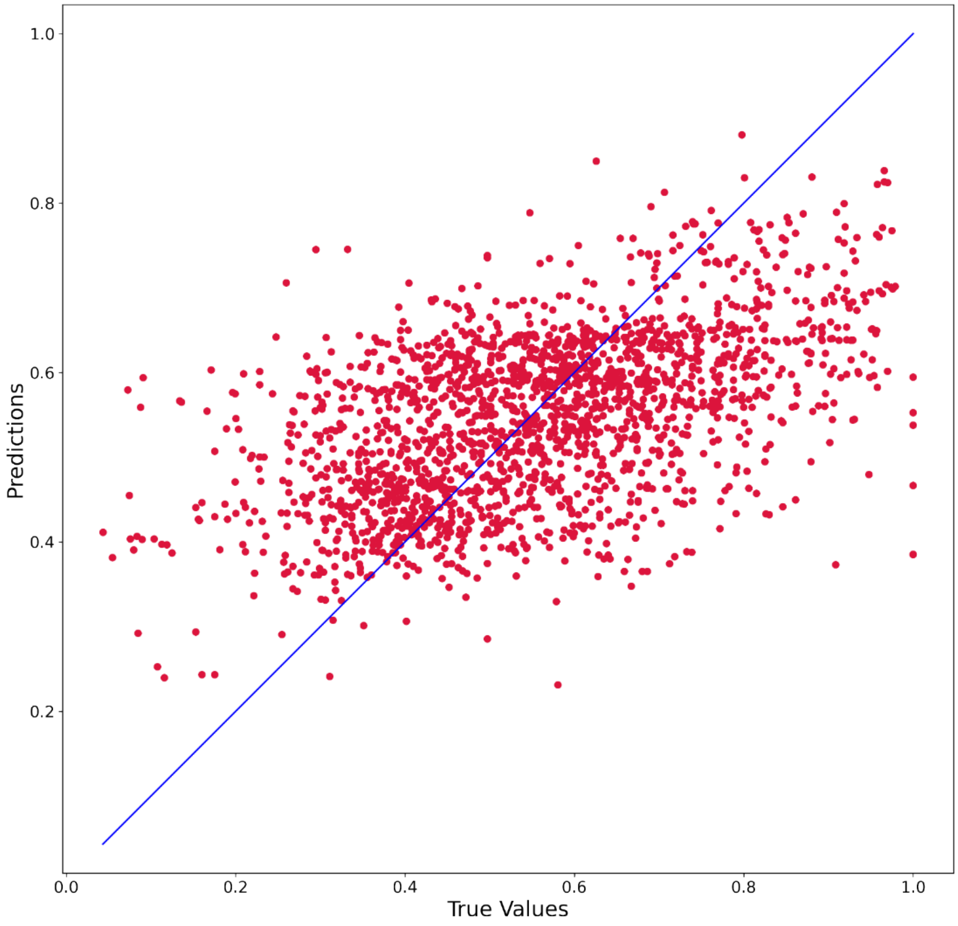

1.2
1.0
0.8
0.6
0.4
0.2
0.0
Predictions
0.0
0.2
0.4
0.6
0.8
1.0
1.2
True Values

Figure 14


Linear Regression R2 - Sequential Forward Selection
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
r2
+ act_composite_75_pctl
0.636
+ act_composite_25_pctl
0.6699
+ exp_instruc_total_per
0.6698
+ est_fte
0.6702
+ exp_acad_supp_total_per
0.6703
+ exp_student_serv_total_per
0.675
+ exp_inst_supp_total_per
0.6753
+ acceptance_rate
0.6753
1
2
3
4
5
6
7
8
iter

Prediction Without Using Selectivity Measures

In light of the equity concerns associated with “gaming” of boosting graduation rates through denying access to higher education to those less likely to succeed, we chose to explore the possibility of predicting graduation rates without considering measures of selectivity. Compared to the previous models, the regression model excluding selectivity measures performed less than half as well with a model score of .286 and a R^2 value of .265. This is not surprising as the ACT scores contributed 67% of the explanatory power of the linear regression over the Trimmed Set, as visible in Figure 14 above. Figure 15 below depicts the low predictive power of the model excluding selectivity variables.


Predictions
True Values