Stat 471 Final Project

Madeline Ellis

19 December 2021

github.com/madelinels012/NY-HS-grad_rates

Executive summary

-Problem

According to statistics from 2019, New York state measures 0.51 on the Gini index, which is a guide used to determine the distribution of wealth across a population. New York's index is only surpassed by Puerto Rice, making it the worst of all the states when it comes to income equality.[1] (footnote citation).

Of course, this fact shouldn't surprise anyone. New York houses the country's largest city -- which, in turn, houses Manhattan -- which, coincidentally, also encompasses some of the world most expensive real estate.

Since education has an enormous influence on income, I decided to study graduation rates from New York state public schools. In analyzing graduation rates, I took into consideration test scores, school size, funding per student, teacher qualifications, and a small amount of socio-economic data.

-Data

I collected the data from the New York State Education Department (NYSED) for the 2018-2019 school year. It appears this data is used mostly for self-evaluation purposes, as it had many defined metrics used for school, student or teacher performance. A few examples of these include graduation rates, high school equivalency exams, test scores, teacher certifications, CCCR (College Career Civic Readiness), N/RC (Need to Resource Capacity), which is an index involving both location (Urban, Suburban, Rural) as well as socio-economic data (no. of students enrolled in free or reduced lunch).

-Analysis

To analyze the data, I created a number of graphs and tables displaying the distribution of graduation rates, the relationship between a school's N/RC classification and its graduation rates, as well as a few other interesting findings. In addition, I performed three regression techniques to the dataset (ridge regression, and lasso regression), as well as a decision tree for predicting graduation rate outcomes. Towards the end, I evaluated the analyses (ridge, lasso, and decision tree) to determine which of the features seem to predict graduation rates most strongly

-Conclusions

I conclude the analysis with the observation that test scores are very indicative of high graduation rates. Another interesting feature in terms of graduation rate is expenditure per student. It appears that, in many of the models, the coefficient corresponding to expenditure per student was negative, meaning it was associated with lower rates of graduation. I believe this would be especially useful to the stakeholders because I think the data needs to be divided into more precise socio-economic classes in order to understand fully the problem of income disparity and take the appropriate steps.

[1] "Gini coefficient as a measure for household income distribution inequality for U.S. states in 2019." *Statista*, www.statista.com/statistics/227249/greatest-gap-between-rich-and-poor-by-us-state/

Introduction

-Background information

New York state has some of the worst wealth disparity in the United States. In order to understand one aspect of this larger phenomenon, I decided to analyze graduation rates of New York high schools in the academic year 2018-2019. My aim is to bring some clarity to this problem by highlighting some of the factors which may lead to higher graduation rates and those which may lead to lower graduation rates. Finishing high school is not only important for determining future income, but it is also very important for crime, homelessness, teen/early 20s pregnancy, and all the other difficult circumstances that often accompany poverty.

-Analysis goals

For my analysis, I am primarily interested in how a few key features seem related to graduation rates and thus how those features can be used to predict future graduation rates. To evaluate success in my analysis, I will be analyzing the RMSE (Root Mean Squared Error) of three predictive models as well as analyzing several graphs and tables for trends. The features I am particularly interested in include:

Feature	Description	Type
needs_index	An index used to measure socio-economic and location information (urban, suburban, rural) for each high school; Numbered 1 through 6, with 1 being schools in NYC; 2 being schools in Buffalo, Rochester, Syracuse, Yonkers; 3 being other urban/suburban schools with lower poverty; 4 being rural schools with lower poverty; 5 being schools with average poverty; 6 being schools with high poverty. Poverty is measure by the No. of students enrolled in the Free of Reduced Lunch Program.	Categorical
expenditures	The total amount of government money allocated per student.	Continuous
overall_status	Gives the school's current status. Categories include: Good Standing, Targeted Support and Improvement, Comprehensive Support and Improvement, Target District.	Categorical

cccr_level	An index which measure College, Career, Civic Readiness (CCCR). Numbered 1 through 4, with 4 being the best.	Categorical
ela	Average test score for ELA (English language arts) in the high school.	Continuous
social_studies	Average test score for Social Studies in the high school.	Continuous
science	Average test score for Science in the high school.	Continuous
math	Average test score for Math in the high school.	Continuous
combined_test	Composite test score including all subjects.	Continuous

-Significance

Education is very important to the structure of society. This is, in part, why education often is discussed by political candidates. Beyond politics, I think studying educational datasets like this one could be highly beneficial to government officials who work on education policy. New York is unique in that it has several large urban areas but also many small rural communities. Analyzing the graduation rates of New York could assist government officials of other states with similar demographics, such as Illinois, California, and Texas. Furthermore, studying the education levels of a state like New York could have important political implications since urban/suburban and rural areas often desire different values/policies in an elected official.

Data

-Data source(s)

I collected the data from the New York State Education Department (NYSED) for the 2018-2019 school year (https://data.nysed.gov/downloads.php). From research on the internet, it seemed the process to open and use Microsoft Access databases in R involves some SQL knowledge, which I lack. I eventually ended up opening the sheets I wanted to analyze in Microsoft Excel and saving them as .xlsx files. Then, I imported them into RStudio using the function “read_excel” from the package “readxl.” The columns relevant to my analysis goals didn’t experience any data loss, but there appeared to be some data loss from a few columns that seem to be for the administration’s record keeping such as a column marked “override,” which contained values “Y,” as well as other markers of success on a school/administrative basis.

To convert them into .xlsx files, I opened excel and chose “import dataset.” Then I selected the Access file and chose the sheet I wanted to open in Excel. Then I saved it.

-Data cleaning

Once all the files were loaded into R, I began with the data set labelled "ACC HS Composite Performance." This is the dataset which contained test scores. First, I filtered for the year 2019 only. Then I filtered for "All Students" – this was important because the test score data was included in total but was also included in different subcategories (by race, ethnicity, disability, etc.). Next, I pivoted the data wide, so that each test score would have its own column – originally, they were all thrown together. In one column. Finally, I had to remove all schools which were not public high schools. In the documentation associated with the database, it explains that the school id (referred to as the entity_id) has certain digits indicating different school types. I did not want elementary schools, middle schools, or charter schools, so I had to separate the entity_id into three different columns and filter the dataset such that those columns did not contain certain numbers. For example, charter schools were associated with "86" so filtered the dataset such that the column did not contain 86. Afterwards, I pasted the entity_id back together, and moved on to tidying the other 7 datasets.

These were much easier to clean because I didn't have to divide the entity_id again. Rather, once I was done editing the columns, I used the function "inner_join" so that only the schools with entity_id's matching the first dataset would be added. For several of the other datasets, there was not subdivided information based on demographics, so I also did not have to pivot wide in those cases. These factors made cleaning the other datasets much easier. Throughout cleaning, I forced many of the columns to be numeric. This was 1) to make sure I wouldn't run into problems later and 2) because the original database wrote the letter "s" for missing data, so by forcing it to be numeric, I was also forcing these missing values to be "NA"s. Following these steps, I omitted all rows which contained a "NA." I also reorganized and renamed the columns, deleting a few columns I no longer thought were relevant. Finally, I wrote the tibble to a .cvs file.

-Data description

There are 1054 observations in the dataset; each observation represents a high school in New York. There are 18 features. The response variable is "grad_rate_combined", and it is continuous. In addition to the features below there are three non-feature parameters: "entity_cd", "district_cd", and "county_cd". These values correspond to unique identification numbers belonging to the school, school district, and county, respectively. They are categorical.

Feature	Description	Type
district_name	The name of the school district the high school is associated with	Categorical
county_name	The name of the county the high school is within	Categorical
num_teach	The number of teachers at the high school	Continuous

teach_oc	The number of teachers who are working out of certification	Continuous
teach_inexp	The number of teachers who less than four years' experience in their role	Continuous
needs_index	An index used to measure socio-economic and location information (urban, suburban, rural) for each high school; Numbered 1 through 6, with 1 being schools in NYC; 2 being schools in Buffalo, Rochester, Syracuse, Yonkers; 3 being other urban/suburban schools with lower poverty; 4 being rural schools with lower poverty; 5 being schools with average poverty; 6 being schools with high poverty. Poverty is measure by the No. of students enrolled in the Free of Reduced Lunch Program.	Categorical
pupil_count	The number of students enrolled in the high school for the 2018-2019 academic year	Continuous
expenditures	The total amount of government money allocated per student.	Continuous
overall_status	Gives the school's current status. Categories include: Good Standing, Targeted Support and Improvement, Comprehensive Support and Improvement, Target District.	Categorical
cccr_level	An index which measure College, Career, Civic Readiness (CCCR). Numbered 1 through 4, with 4 being the best.	Categorical
ela	Average test score for ELA (English language arts) in the high school.	Continuous
social_studies	Average test score for Social Studies in the high school.	Continuous

science	Average test score for Science in the high school.	Continuous
math	Average test score for Math in the high school.	Continuous
combined_test	Composite test score including all subjects.	Continuous
grad_rate_6	6-year graduation rate.	Continuous
grad_rate_5	5-year graduation rate.	Continuous
grad_rate_4	4-year graduation rate.	Continuous

-Data allocation

Before performing any data visualization or analysis on the dataset, I first split it into training and testing sets. The training sets included 80% of the total samples, while the testing sets included 20% of the total samples. I used the training set of data exploration, and used the test set solely to analyze the results of the predictive models I made.

-Data exploration

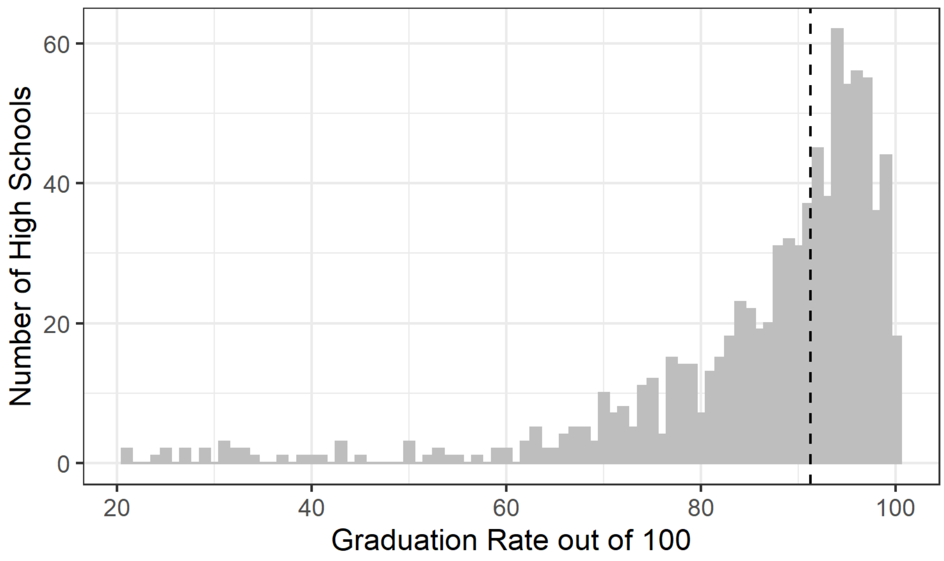
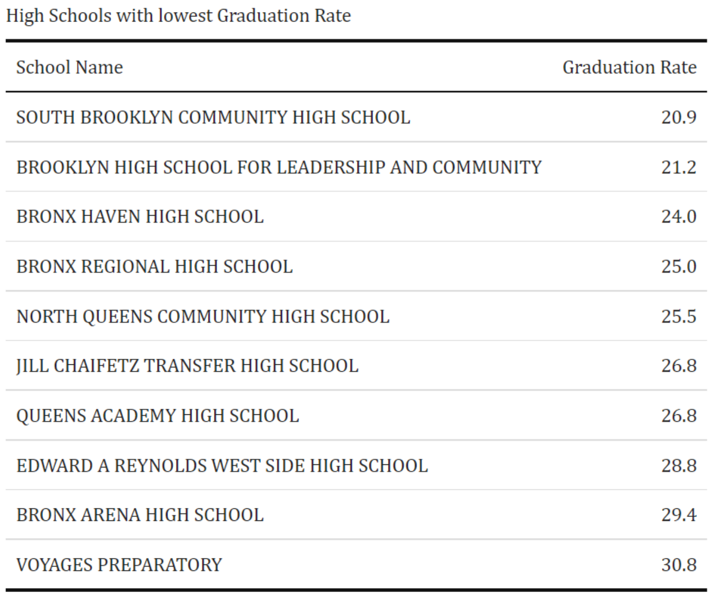


Figure 1: Histogram of Graduation Rates

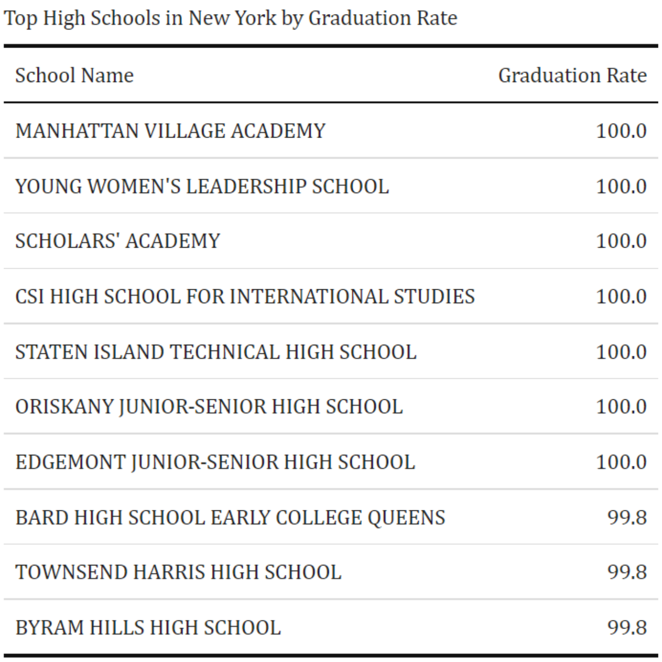
There appears to be quite a bit of spread in the distribution of graduation rates per high school (see Figure 1). The histogram skews to the right, with there being more outliers on the low end than on the upper end. The average graduation rate is high (86%), with the median being 91%. It is not surprising that the average graduation rate is higher than 50% -- in fact, we would expect hope that the average graduation rate is much higher than 50%. However, it is a bit surprising that the median is over 90% and that it is larger than the mean. This must indicate that there are quite a few schools in New York which have near perfect graduation rates. Looking at Figures 2 and 3, we can see the high schools with the best graduation rates and the worst. The top 10 high school graduation rates range from 100% to 99.8%, while the bottom 10 range from 30.8% to 20.9%. When it comes to variation within the features, it seems to depend entirely on the feature. For example, test scores has relatively typical variation, as does the number of teachers and pupil size since some high schools are very large and others are very small. The standard deviation of test scores (the variable is “combined_test”) is 33, while the minimum value is 49 and the maximum value is 247. But there are some features which are categorical, such as “needs_index.”



High Schools with lowest Graduation Rate

School Name	Graduation Rate
SOUTH BROOKLYN COMMUNITY HIGH SCHOOL	20.9
BROOKLYN HIGH SCHOOL FOR LEADERSHIP AND COMMUNITY	21.2
BRONX HAVEN HIGH SCHOOL	24.0
BRONX REGIONAL HIGH SCHOOL	25.0
NORTH QUEENS COMMUNITY HIGH SCHOOL	25.5
JILL CHAIFETZ TRANSFER HIGH SCHOOL	26.8
QUEENS ACADEMY HIGH SCHOOL	26.8
EDWARD A REYNOLDS WEST SIDE HIGH SCHOOL	28.8
BRONX ARENA HIGH SCHOOL	29.4
VOYAGES PREPARATORY	30.8

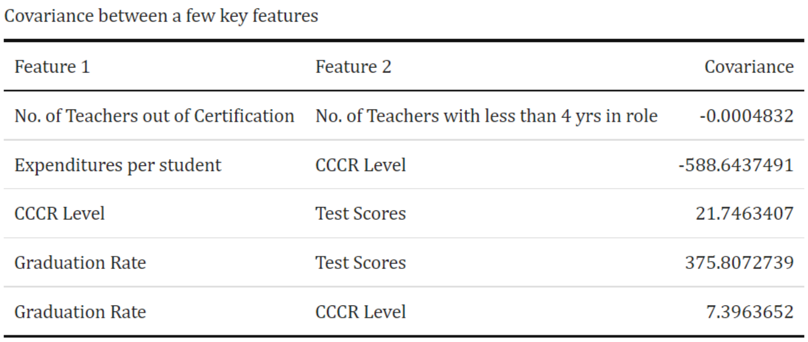
Figure 2: High Schools with Lowest Grad Rates



Top High Schools in New York by Graduation Rate

School Name	Graduation Rate
MANHATTAN VILLAGE ACADEMY	100.0
YOUNG WOMEN'S LEADERSHIP SCHOOL	100.0
SCHOLARS' ACADEMY	100.0
CSI HIGH SCHOOL FOR INTERNATIONAL STUDIES	100.0
STATEN ISLAND TECHNICAL HIGH SCHOOL	100.0
ORISKANY JUNIOR-SENIOR HIGH SCHOOL	100.0
EDGEMONT JUNIOR-SENIOR HIGH SCHOOL	100.0
BARD HIGH SCHOOL EARLY COLLEGE QUEENS	99.8
TOWNSEND HARRIS HIGH SCHOOL	99.8
BYRAM HILLS HIGH SCHOOL	99.8

Figure 3: High Schools with Highest Grad Rates



Covariance between a few key features

Feature 1	Feature 2	Covariance
No. of Teachers out of Certification	No. of Teachers with less than 4 yrs in role	-0.0004832
Expenditures per student	CCCR Level	-588.6437491
CCCR Level	Test Scores	21.7463407
Graduation Rate	Test Scores	375.8072739
Graduation Rate	CCCR Level	7.3963652

Figure 4: Covariance between some features

Another area of interest involves the features themselves. Figure 4 plots a few examples. Overall, the features were not too correlated, although there are a few feature pairs that are correlated in predictable way. For example, there is a very high positive covariance between graduation rate and test scores. At the same, time there is a very high negative covariance between expenditures per student and CCCR level. This seems strange because we would expect the CCCR level to increase as expenditures increase, but this weird relationship is also seen when studying graduation rate and expenditures. See Figure 5 for more detail. There is a large cluster of high schools who appear to receive an average amount of money per student and also tend to perform close to the average of 86% or above it. However, looking beyond this cluster, there is also a clear negative trend. Certainly, there are some high schools who receive lots of money and perform very well. However, it is clear that there is a large number of high schools who do receive more than the average expenditure per pupil and do not perform above average when it comes to graduation rates.

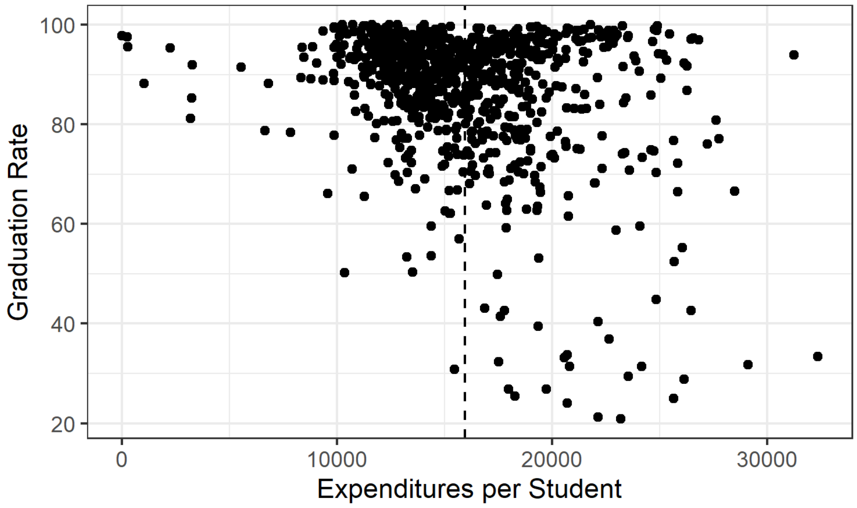


Figure 5: Expenditures and Graduation Rates; dotted line is mean expenditure

In addition to studying the relationship between the features, I also analyzed a few interesting correlations between the response and two features. Mainly, I plotted a box plot of the graduation rates in terms of the Needs Index (N/RC), which analyzes socioeconomic factors (no. of kid with free or reduced lunch), as well as location parameters (Urban, Rural, Suburban). Figure 6 features the box plot with the Needs Index. There is a clear correlation between the Needs of the students in a high school and the graduation rate of the high school itself. NYSED classifies N/RC 1 schools as being all schools in New York City, where there traditionally higher rates of poverty but also higher rates of affluence. This makes this box plot especially interesting. The average graduation rate in this group is still very high, reasonably on par with the other groups, although not as high as the high schools classified as being lower need (N/RC level 6). However, there appears to be a great deal of outliers in this group, which are present as data points at the bottom of the box plot. It’s clear from analyzing this that some of the poorest performing high schools are within New York City, but also some of the most privileged high schools.

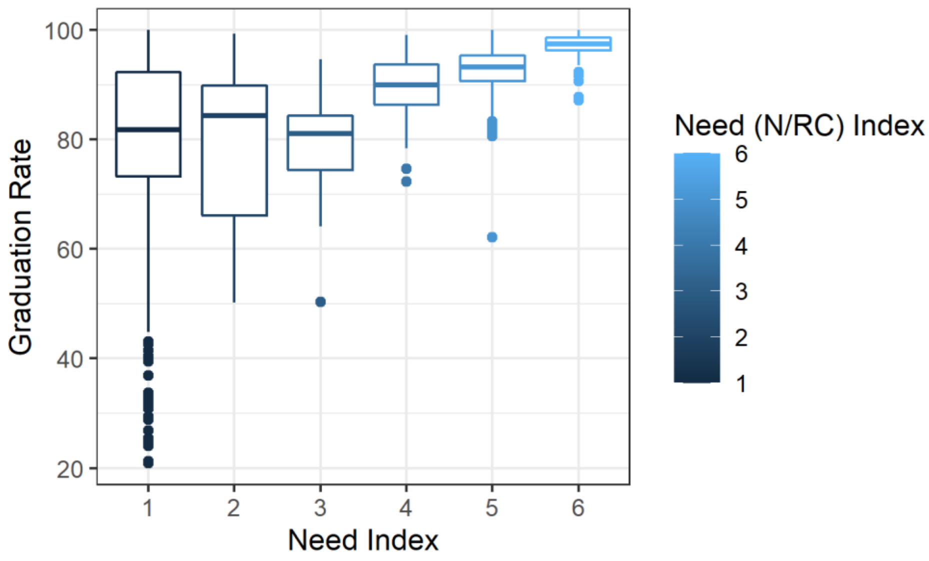


Figure 6: Graduation Rate Box Plots with Needs Index

Modeling

-Model class 1

-Ridge Regression

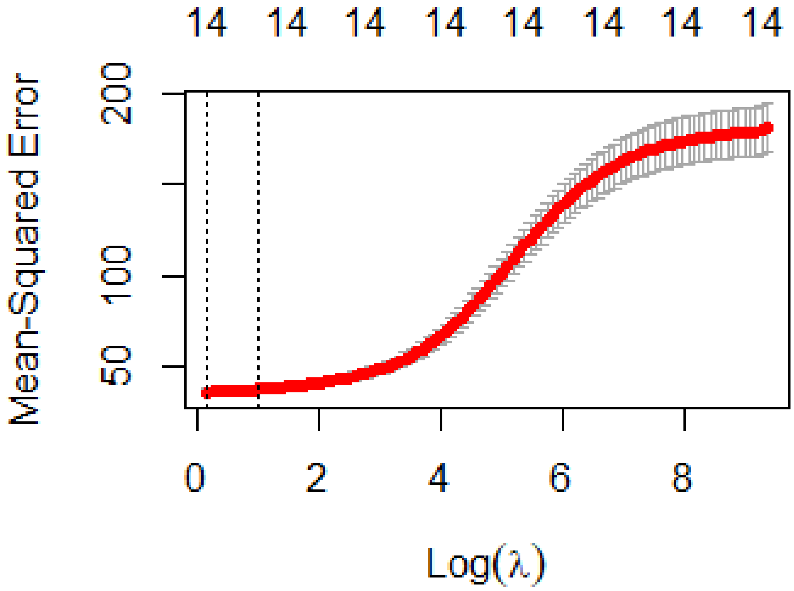


Figure 7: Ridge Regression CV Plot

After fitting a ridge regression on the data, we cross validate over the penalty variable λ. This can be seen in Figure 7. The minimum λ is about 1.2 and the λ associated with the one standard error rule is 3.3. Since these values are so close to zero, it suggests that the data doesn't require much penalization. More interesting is the trace plot for ridge regression, which will show which features seem to be most prominent under this model. See Figure 8 for the trace plot.

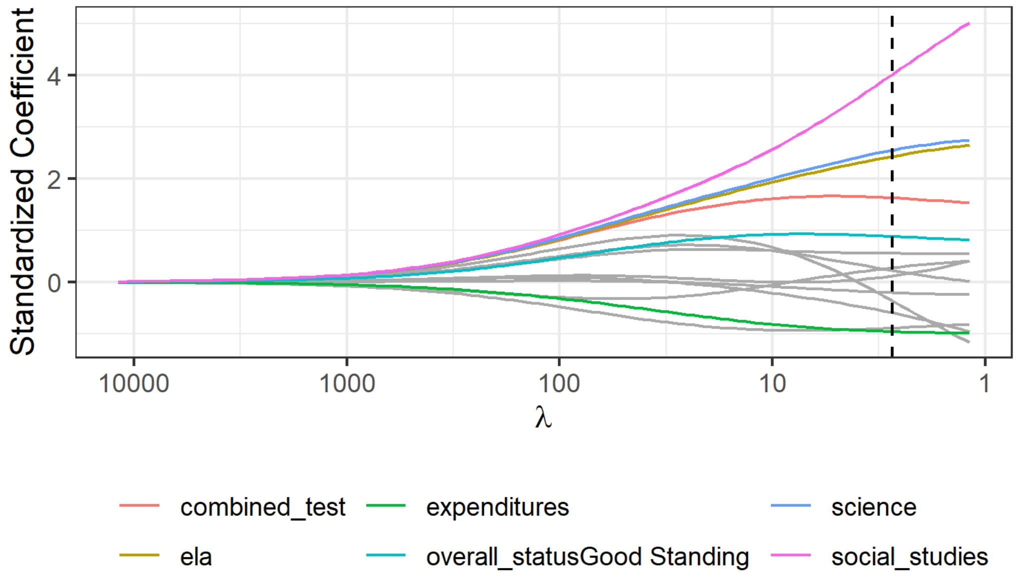


Figure 8: Ridge Regression Trace Plot

From Figure 8, it is clear that, as I noted in the data exploration section, expenditures has a negative effect on the graduation rate, meaning increased expenditure per student tends to cause the graduation rate to decrease. The other features have a positive effect on graduation rate in the model. Unsurprisingly, test scores seem to be very indicative of high graduation rates. If the school is in good standing, it also will lead to slightly higher graduation rates according to this model. To see the specific values of these coefficients, please see the index.

-Lasso Regression

In theory, we would expect lasso to work better, since graduation rates seem to be strongly influenced by a few features rather than many features. After fitting a lasso regression to the dataset, we cross validate over the penalty variable λ. This can be seen in Figure 9. The minimum λ is about .009 and the λ associated with the one standard error rule is .5.

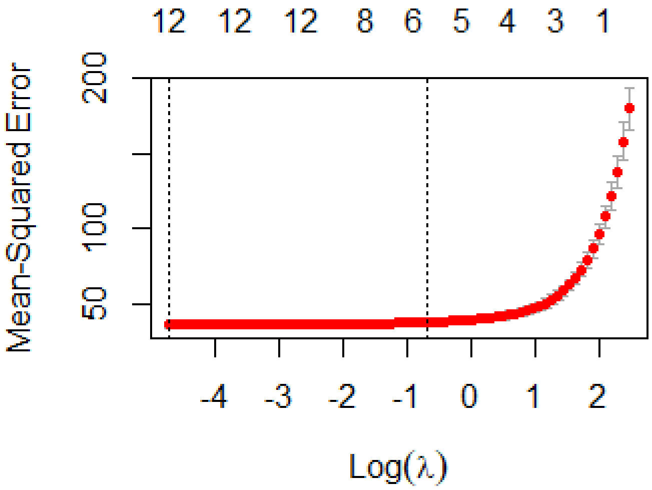


Figure 9: Lasso Regression CV Plot

Moreover, the λ chosen by the one standard error rule would suggest that this model performs better with less features – six, rather than twelve. To see which features have the most effect on the response variable, see Figure 10 for the trace plot for lasso regression. It is clear from this figure that there are a few other features with negative coefficients, mainly if a school has a status of "Comprehensive Support and Improvement." The feature pupil_count is also negative, suggesting this model associates very large schools with worsening the graduation rate. As in ridge regression, expenditures is also a negative coefficient, but it does not appear to be as influential as in ridge regression. To see the specific values of these coefficients as well as the coefficients of other features, please see the index.

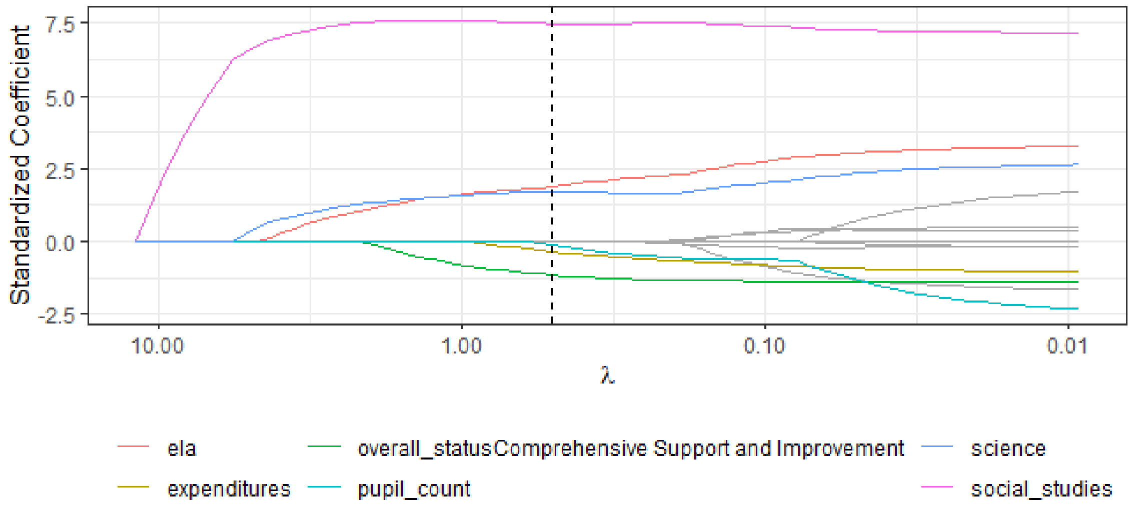


Figure 10: Trace Plot for Lasso Regression

-Model class 2
-Decision Tree

After fitting the dataset to a decision tree model, I received the following tree as output (Figure 11).

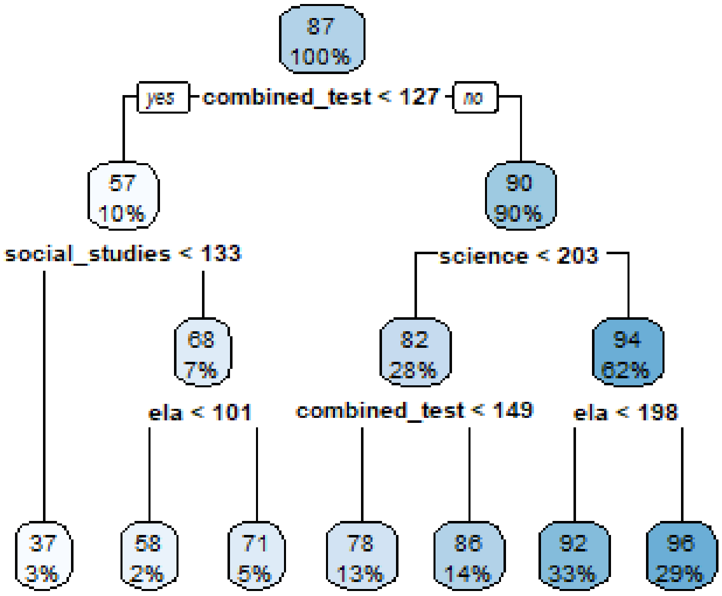


Figure 11: Decision Tree

Before pruning the tree, I run a cross validation over the number of terminal nodes and receive the following in Figure 12. It appears that the minimum number of nodes is 6, but through the one standard error rule, we choose 4 terminal nodes for our model. Thus, I will prune the tree so that it has 4 terminal nodes (Figure 13). The final pruned decision tree has four terminal nodes and seems to focus on the test scores of the students entirely. First, it uses the “combined_test” as an initial threshold, then it splits between the subjects “social_studies” and “science.” To see the complexity parameter table and look at these values more closely, please see the index.

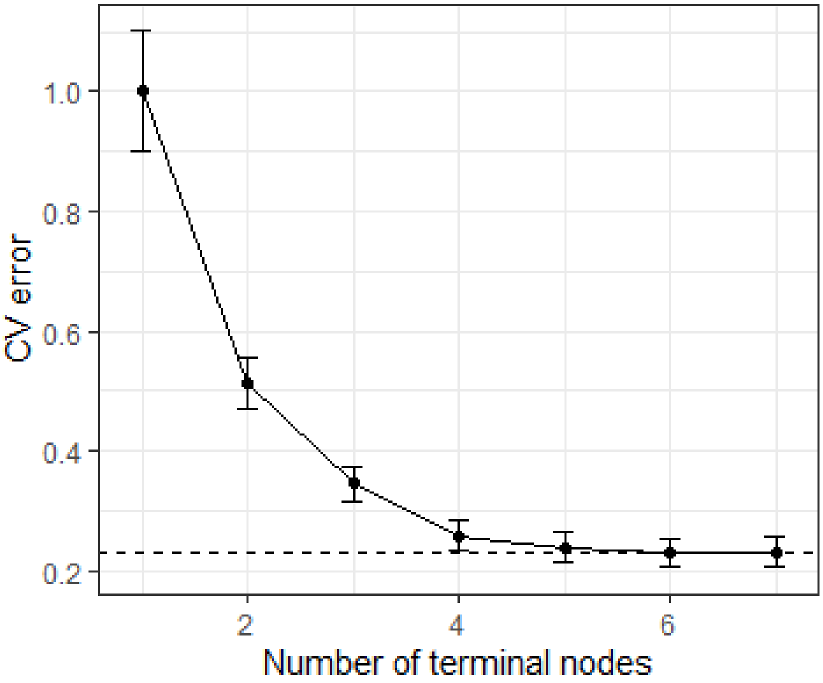


Figure 12: CV plot for Decision Tree

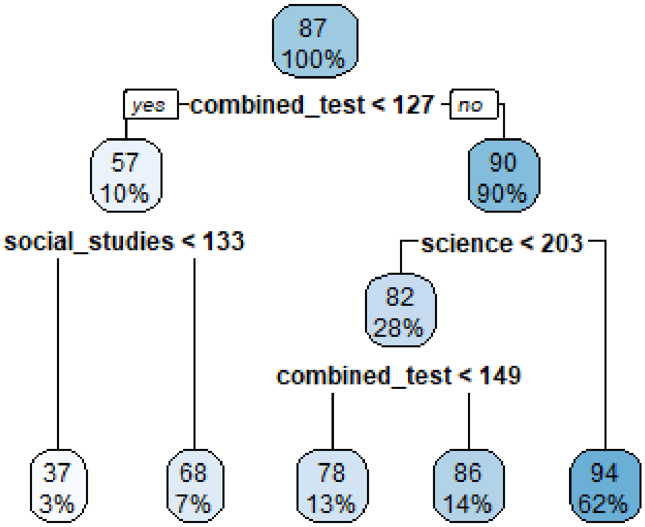
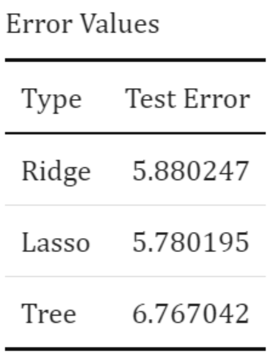


Figure 13: Pruned Decision Tree

Conclusions

-Method comparison

To evaluate the performance of the models, I made predictions on my test set of the response variable (graduation rates) and then compared those predictions to the actual graduation rates through the root mean squared error (RMSE) test. Unfortunately, my models have high error as can be found in Figure 14. I think this is because my dataset needs more features. With this in mind, lasso regression performed slightly better than the others, and I think this likely due to the fact that it isolated only the few features that actually made a big difference on the outcome (graduation rate).



Error Values

Type	Test Error
Ridge	5.880247
Lasso	5.780195
Tree	6.767042

Figure 14: RMSE values for models

-Takeaways

I would explain to stakeholders that there are a few inefficiencies when it comes to the expenditure, mainly that there seems to be a trend between increased expenditures to N/RC level 1 (New York City) while still seeing consistently very low outliers in this area in terms of graduation rate. This should be concerning, in my opinion, and I think they should divide the N/RC levels such that more affluent NYC high schools are in a different category than struggling NYC high schools. This way, it would be much faster to see what the issue is from a data science perspective. But additionally, I believe it would improve efficiency from an administrative angle as well, in that it would make communication more precise, and the trends would be immediately noticeable to people who don't have a background in data science. Furthermore, I would recommend that they seek to do more analysis on teacher quality. A very strong trend in my analysis was that higher test scores often corresponded with higher graduation rates. Therefore, I think it would be wiser to spend more money on hiring and retaining excellent teaching staff.

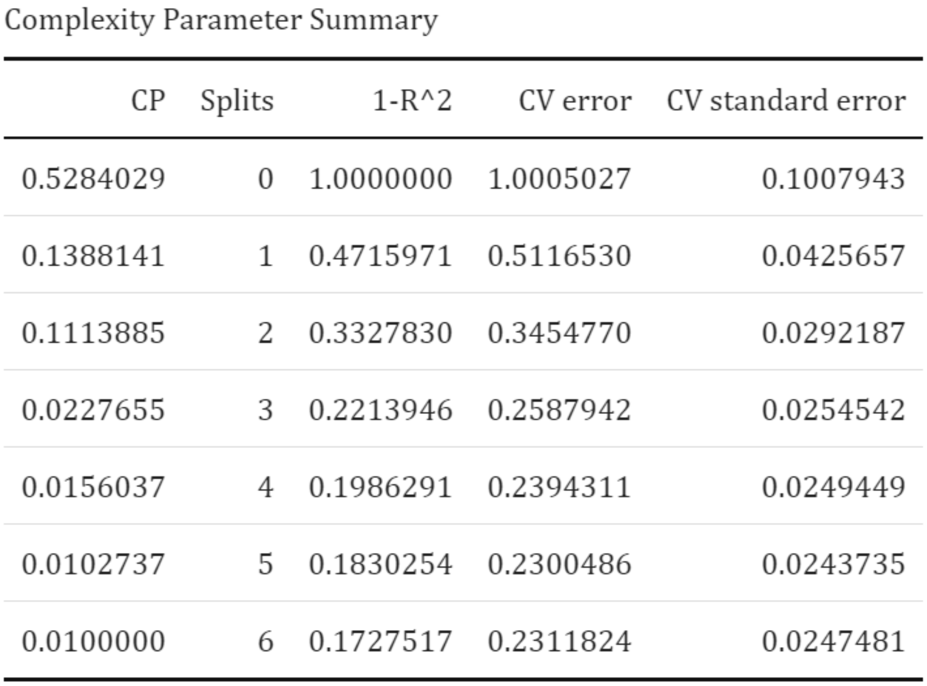
-Limitations

There are a few limitations in the dataset, mostly that there should be more features. For example, there could be more data collection on the following things: No. of teachers with Bachelor's degrees in the subject they teach in; No. of teachers with Master's degree or higher; teacher salary; No. of students who matriculate to college; No. of students who graduate with a degree (2-year, 4-year, trade school). However, I also think another issue is that the N/RC index should be separated into its parts rather than taken together. There should be separate feature variables for the following categories: No. of students with free or reduced lunch; No. of students whose parents have college educations; location (Urban, Rural, Suburban). These obstacles mainly made it very difficult to get at the underlying issue behind some of the problems. On the other hand, keeping the N/RC index as it was given was valuable in that it allowed me to see clearly why the index is so poor, and I think because of that if I were able to talk to the stakeholders, I would be able to give them a very clear picture of the issues and recommend next steps.

-Follow-ups

I would want to follow this up by analyzing the demographic information of the students, such as race, ethnicity, first generation (meaning, their parents did not graduate from a four-year institution), disability status, immigrant status, etc. This was present to some extent in the original database, but I was unable to do much with it because it was only present for a few features, such as test score, but not the others. This might uncover additional instances of inefficiency which could then be analyzed. Another interesting analysis for further study would be to study the test scores in terms of student outcomes. For example, we could study if there is a correlation between a certain score on ELA (English language arts) and attending college at higher rates.

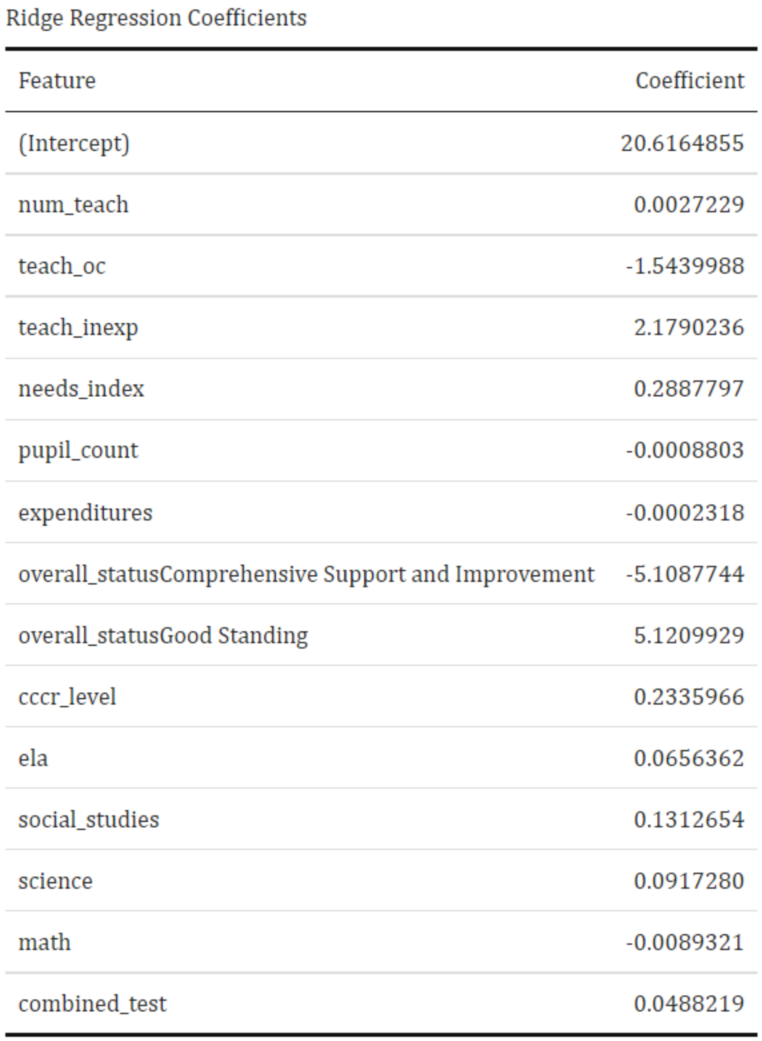
Index



Complexity Parameter Summary

CP	Splits	1-R^2	CV error	CV standard error
0.5284029	0	1.0000000	1.0005027	0.1007943
0.1388141	1	0.4715971	0.5116530	0.0425657
0.1113885	2	0.3327830	0.3454770	0.0292187
0.0227655	3	0.2213946	0.2587942	0.0254542
0.0156037	4	0.1986291	0.2394311	0.0249449
0.0102737	5	0.1830254	0.2300486	0.0243735
0.0100000	6	0.1727517	0.2311824	0.0247481

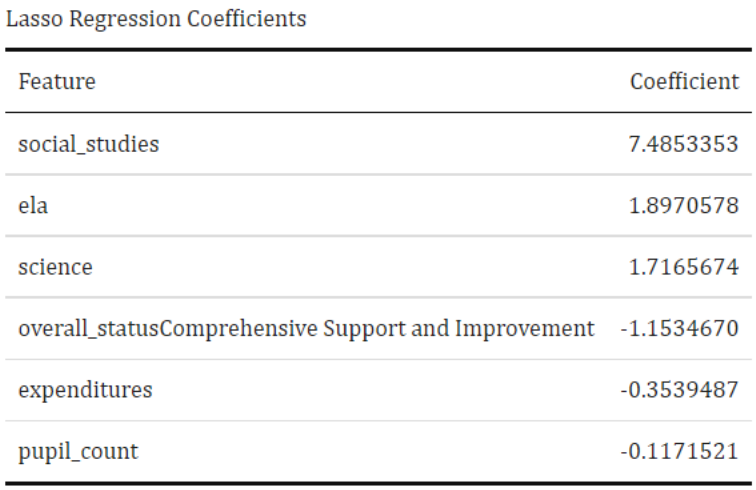
Figure 15: Complexity Parameter CV from Decison Tree



Ridge Regression Coefficients

Feature	Coefficient
(Intercept)	20.6164855
num_teach	0.0027229
teach_oc	-1.5439988
teach_inexp	2.1790236
needs_index	0.2887797
pupil_count	-0.0008803
expenditures	-0.0002318
overall_statusComprehensive Support and Improvement	-5.1087744
overall_statusGood Standing	5.1209929
cccr_level	0.2335966
ela	0.0656362
social_studies	0.1312654
science	0.0917280
math	-0.0089321
combined_test	0.0488219

Figure 16: Ridge Regression Coefficients



Lasso Regression Coefficients

Feature	Coefficient
social_studies	7.4853353
ela	1.8970578
science	1.7165674
overall_statusComprehensive Support and Improvement	-1.1534670
expenditures	-0.3539487
pupil_count	-0.1171521

Figure 17: Lasso Regression Coefficients