New York High School Graduation Rate

Stat 471 Final Project

Kennedy Manley

December 19, 2021

https://github.com/kennedymanley/grad-rate-project

Table of Contents

Executive Summary

High school graduates are proven to have better quality of life due to higher income, better living conditions, expanded access to health care, and overall more beneficial opportunities. It is important to encourage the population to graduate from high school because it benefits personal success, but also the economy. Hence, for my final project, I decided to look into the graduation rates from the New York high school system for the 2019-2020 school year. Although, this is just a singular state, the New York population is one of the largest in the United States. Additionally, this data only looks at one school year which could have some particular factor that caused that specific year to give unique results. However, analyzing New York High School Graduation Rates could increase knowledge about trends, find certain subsets of people or geographic regions that need to be supported more throughout high school to encourage high school graduation, and determine the alternate paths student choose instead of receiving their high school diploma.

The data for this research project was acquired from the New York State Education Department website. The New York State Education Department (NYSED) is committed to publicly reporting data so that people can be better informed in their work to improve student achievement. The NYSED separates data by school year and provided their graduation rate database for download. The graduation rate database is described as "This database contains annual graduation, and dropout data for the state as well as by county, Need to Resource Capacity group, district, public school and charter school. Annual graduation data is included for the current four-year cohort (June and August graduates), five-year June and August, and six-year June and August cohorts." The primary response variable of interest is the graduation rate, which had to be calculated by hand since the formatting of the downloaded data set did not allow for much manipulation of the given variable. The numerical explanatory variables were also manipulated to be in numeric form to allow for calculation and manipulation. There are both categorical and quantitative variables, but this research focuses primarily on the quantitative factors and the categorical factors that can be written as factors.

After downloading the data, but before exploring the data and running analysis, the data was split into a training dataset and test dataset. The test dataset was exclusively used for evaluating performance of the difference models. The training dataset was used to run many different tests and models to see how the different variables contributed to the graduation rates. Models were built and run to determine the best way to classify this data. The different models included linear regression, ridge regression, lasso regression, elastic net regression, random forest, and classification trees. The efficiency of the model was evaluated based on the computed root mean squared error (RMSE) and the misclassification error rates.

The overall data analysis pointed towards the lasso regression model being the best model for this data analysis. The lasso method worked well for this data because it shrinks the coefficients of some factors which can provide the best fit since there is no inclusion of unnecessary factors. The root mean squared error was used to assess model fit for lasso regression and it concluded an RMSE of 0.76 which is significantly lower than the error rates of the other models. The linear regression model provided the lowest error rate, but inaccuracies with the factors and type of test led to an inconclusive result that will be discussed later. Hopefully, this analysis can lead to the increased efforts to escalate high school graduation rates for New York State, and the entire US population.

Introduction

New York State consists of 731 different districts with 4, 412 public schools and 355 charter schools. Within these schools are 2, 598, 921 students obtaining an education[1]. The state of New York requires students complete a minimum of 22 credits, pass four Regents exams with one in each discipline (English Language Arts, Mathematics, Science, Social Studies), and pass one pathway. Pathways encourage the engagement of students in rigorous and relevant academic programs, and students may choose between the Arts, World Languages, Career and Technical Education, Career Development and Occupational Studies, Humanities, and/or STEM pathways. Students in the New York Education System have the opportunity to graduate with one of three unique diplomas: Local, Regents, or Regents with Advanced Designation. The difference between these degrees is determined by the number of Regents scores that were appealed and the number of Regents exams taken and passed[2].

Previous research has shown that "increased education attainment provides individuals with the opportunity to earn a higher income and gain access to better living conditions, healthier foods, and health care services." Additionally, lifetime wealth for male and female high school graduates are on average $219, 500 and $182, 000 higher, respectively, than high school dropouts[3]. The New York State Education Department is very aware of the positive benefits of a high school degree and have dedicated plenty of effort to collecting data to create programming that will hopefully increase the graduation rate. Specifically, in 2020, during the peak of the COVID-19 pandemic, the NYSED created special guidelines to ensure no student was negatively affected academically so students who were otherwise eligible to graduate could do so.[4]

However, there is still more to learn about how 2020 data led to the graduation rates of the New York State Education System's students. Given our knowledge about graduation rates and education, this project seeks to determine how graduation rates were affected by educational, demographic, and geographic factors. The data is curious to determine which specific factors had a large influence in determining the 2020 graduation rate in New York. Distinctively, this research hopes to determine whether geographic (county), demographic (subgroup type, need to resource capacity, aggregation), or educational path (GED rate, diploma type, still enrolled percentage) has the largest impact on graduation rates and within these categories which factor is

[1] NY State Education Department Overview https://data.nysed.gov/

[2] NYSED Diploma Types http://www.nysed.gov/curriculum-instruction/diploma-types

[3] Office of Disease Prevention and Health Promotion https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-health/interventions-resources/high-school-graduation

[4] NYSED 2021 Release of Stats from 2016 http://www.nysed.gov/news/2021/state-education-department-releases-2016-cohort-high-school-graduation-rates

most substantial in increasing the graduation rate. Hopefully, this analysis will provide critical informative statistics on which type of students are most likely to graduate from high school and which type of student is least likely to graduate from high school in New York. This information could lead to the formation of programs to support the least likely students and further analysis into what specifically about the high performing groups are allowing them to succeed in high school. The results of this project will give important details about the way certain groups are able to perform and how factors outside of the classroom may lead to detrimental effects than can lead to lower graduation rates, and thus a lower quality of life after age 18.

Data

Data Sources

The data from this project sources from one location, the New York State Education Department (NYSED) Data website. The NYSED has committed to releasing important data and statistics to the public about key features of their school system. In the graduation rate database, data about performance and outcomes of designated subgroups are reported by total public school, which are aggregated by all districts and charter schools in the state. The NYSED worked to ensure confidentiality of all students so data for groups with fewer than five students or data that would allow for easy identification of the performance of a group with less than five students were not published. The website to find this data is: https://data.nysed.gov/downloads.php

Data Cleaning

After downloading the dataset from the internet, the main portion of the data cleaning came from assigning variable types to each of the variables. This was especially important because when the data was initially imported, all the variables were assigned to a character variable which is especially not helpful when running statistical methods. The variables in the dataset were reclassified as either a factor or a double. Primarily, the social and geographic variables were factors while the educational variables were doubles.

Additionally, the rate data needed to be converted to decimal form. When the data was originally downloaded, the rate data was listed as percentages. However, this was not acceptable form for the data as it is easier to work with if we want everything as a number, meaning we did not want to include the "%" character. Therefore, it was necessary to divide the count of that variable by the enrollment count and then multiply by 100. This allowed the rates to be easier to compute in future data analysis.

Finally, some unneeded variables were removed from the dataset. The data did not need to include information about LEA (categorial way to identify county, city, and district number) and inactivity dates, so they were removed to focus primarily on the more influential factors.

Data Description

Observations

The dataset consists of 227, 450 observations. Each observation corresponds to a certain subgroup at one of the schools in New York. The unique subgroups are All Students, Female, Male, American Indian/Alaska Native, Black, Hispanic, Asian/Pacific Islander, White, Multiracial, General Education Students, Students with Disabilities, Not English Language Learner, English Language Learner, Formerly English Language Learner, Economically Disadvantaged, Not Economically Disadvantaged, Homeless, and Not Homeless. The subgroup code associated with the subgroup name are detailed in the Appendix.

Response Variable

The response variable for this project is the graduation rate. The graduation rate was calculated by the number of students who graduated divided by the number of students enrolled in the cohort based on data first date of entry in 9th grade (2016-2017 school year). The graduation rate is a continuous variable and contains information about all three unique types of diplomas awarded to graduates of the New York State Education System.

Explanatory Variables

The data provided 27 explanatory variables which fall into three main categories: geography, demographics, and educational statistics. For a detailed specification of the variables, refer to the Appendix. This allows the research to show whether certain locations in the state, certain subgroups of the population, and certain other educational factors of the school contribute to the graduation rate.

Data Allocation

First, the NA values of the data were removed. This allowed for all the data to be complete and every observation to be used in the analysis. Next, the data were split into training and testing sets. The training dataset is used for the predictive modeling and the testing dataset is used to evaluate the performance of the models. The training data was selected randomly; however, a seed was used to allow for reproducibility of the data. An 80-20 split was used for the distribution of the data into the training and testing datasets. This means that 80% of the full data set was allocated to training while the remaining 20% was allocated to the testing data set.

Data Exploration

The project sought to determine the overall distribution of the graduation rate. Figure 1 displays the distribution and it appears to be left skewed. There are many observations with 100% graduation rate and, unfortunately, some with a 0% graduation rate as well. The median graduation rate is 90.69%. To verify this data, the dropout rate was also observed as seen in Figure 2. As expected, the dropout rate distribution looks like the mirror image of the graduation rate distribution. The dropout rate is right skewed and has a median of 4.54%.

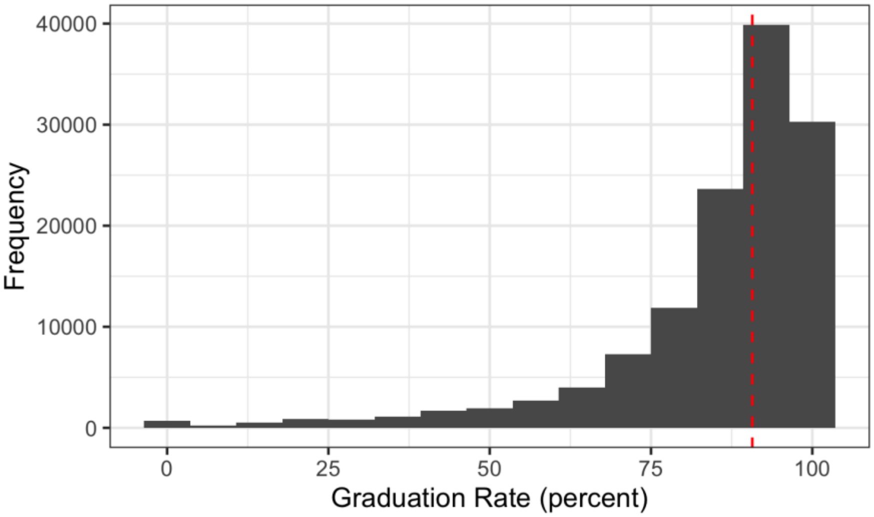


Figure 1: Distribution of Graduation Rate; vertical dashed line indicates the median

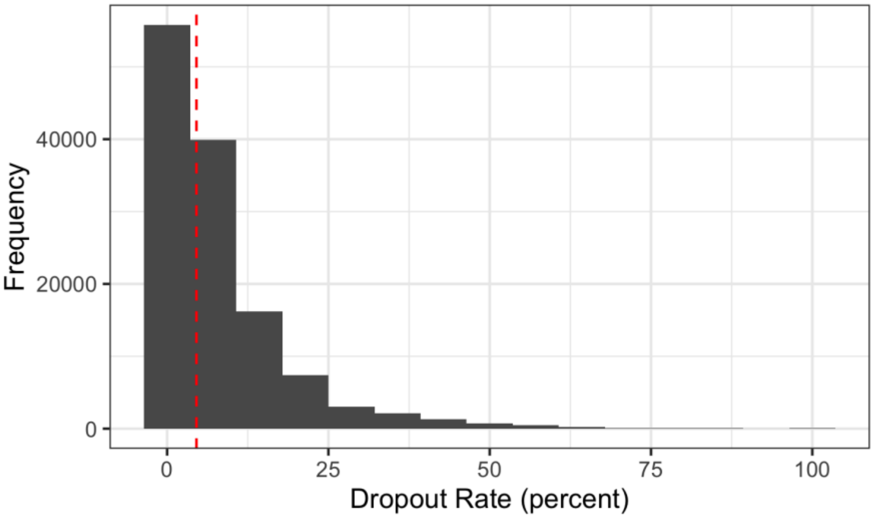


Figure 2: Distribution of Dropout Rate; vertical dashed line indicates the median

Additionally, the project sought to collect graduation rates for certain factors that can be grouped such as county, Need to Resource Capacity (NRC) and subgroup. The data was grouped together for all observations that were contained in the categories then the mean graduation rate was calculated for each subcategory within that group. Figure 3 shows the graduation rate for each county in New York. However, the number of counties requires the x-axis labels to be quite small, so the county names associated with the county codes are labeled in the Appendix. Figure 4 displays that County 32 (Bronx, NY) has the lowest graduation rate, while Counties 55 and 67 (Schuyler, NY and Wyoming, NY) have the highest. Figure 4 displays the graduation rate for each NRC and the meaning of the NRC Codes are detailed in the Appendix. It shows that NRC = 6 (Low NRC Districts) has the highest graduation rate, while NRC = 2 (Large City High NRC Districts) has the lowest graduation rate. This is very expected as areas with a greater ability to meet its students needs with local resources would be more likely to have a majority of their students graduate. Figure 6 displays the graduation rates for different subgroups. Subgroups 7 (Asian/Pacific Islander) and 14 (Formerly English Language Learner) have the highest graduation rate while subgroup 17 (Migrant) has the lowest graduation rate.

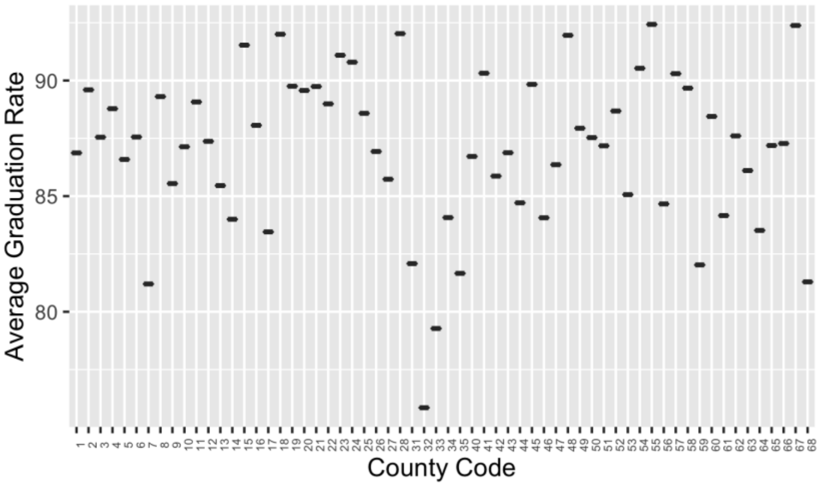


Figure 3: Graduation Rate by County

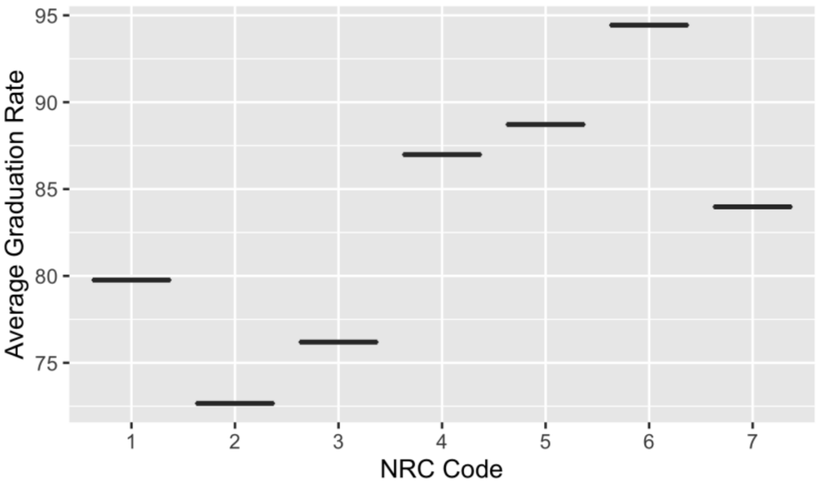


Figure 4: Graduation Rate by NRC

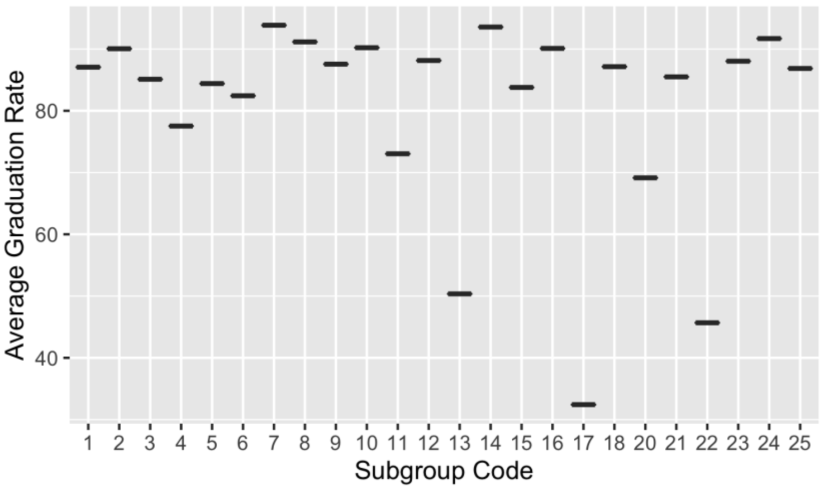


Figure 5: Graduation Rate by Subgroup

We see a strong correlation between a few features based on the correlation plot shown in Figure 6. Besides, the features on the diagonal that represent a correlation with themselves, a strong correlation between enrollment count and regents diploma count is observed. As well as membership code and membership key, which is highly expected since they both correspond to the same information: the cohort membership and year the data was reported. On the other hand, some features have very little to no correlation. An example of these are membership key and GED rate where there seems to be no connection between these features. Finally, we see a negative correlation between graduation rate and dropout rate which is expected since they should always sum to 1.

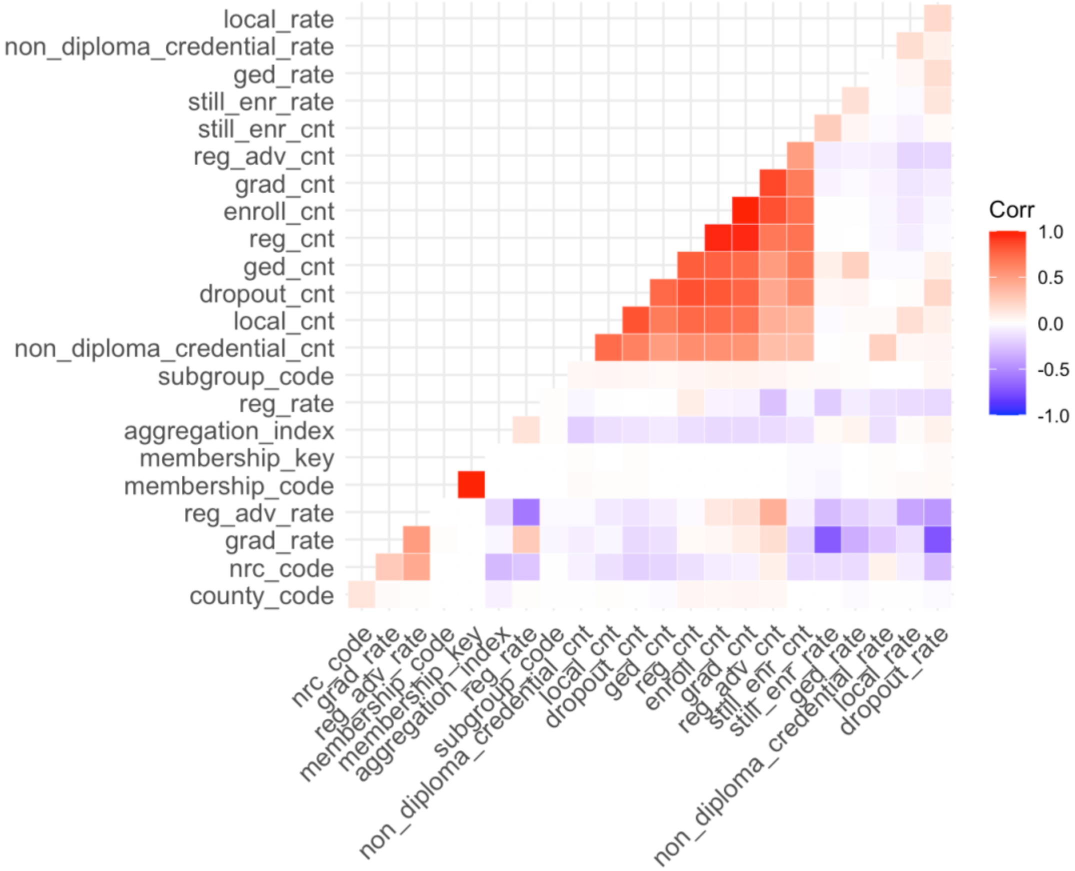


Figure 6: Correlation Between Features

Modeling

Model Class 1 (Regression Methods)

Linear Regression

To begin with, a linear regression model was run on a few of the variables. This model was just for brief knowledge and information before getting information on more complex models that are better fitted for large data with multiple features and observations. The summary of the linear regression model gave coefficients for each explanatory variable and an intercept term. However, the most notable result might be the small p-value (p-value: < 2.2e-16) which tells us that the data is statistically significant.

Lasso Regression

A 10-fold lasso regression was run on the training data. Lasso regression is a variable selection and regularization method that is used to enhance prediction accuracy and interpretation. According to Figure 6, the one standard error rule tells us the ideal value of lambda is 0.11 which is denoted by the rightmost vertical dashed line on the plot. As the log of lambda increases, the more coefficients for factors are shrunken to be 0. When the log of lambda is chosen using the one standard error rule, the ideal number of features is 4. Additionally, Figure 7 lets us see how the highlighted features change as lambda decreases (left to right). All of the features in the plot have coefficients equal to or less than 0. The top 6 features were chosen to be highlighted and they are displayed in color. However, the 4 features that are non-zero and are highlighted are the still enrolled rate, the dropout rate, the GED rate, and the non-diploma credential rate.

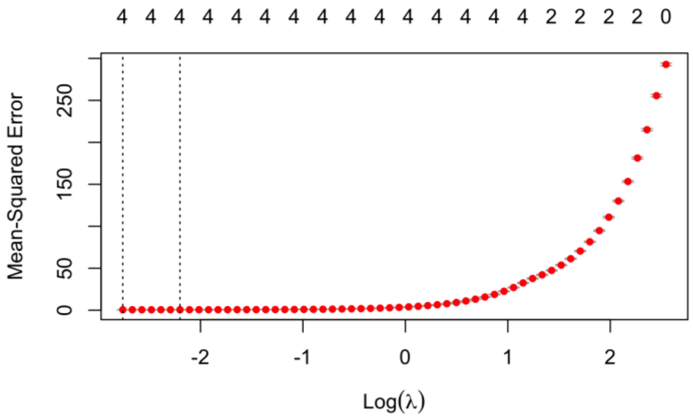


Figure 6: Lasso Regression CV Plot

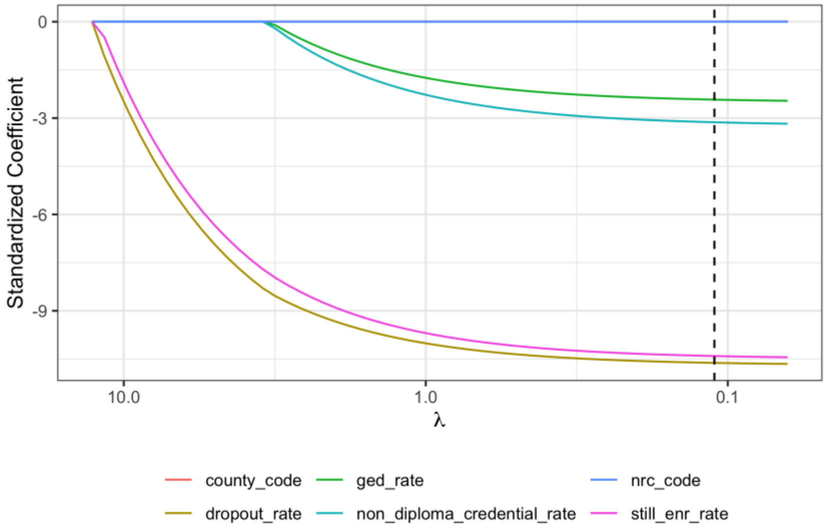


Figure 7: Lasso Regression Trace Plot

Ridge Regression

A 10-fold ridge regression was run on the training data. Ridge regression is used to estimate the coefficients of multiple regression models when independent variables are highly correlated. According to Figure 8, the one standard error rule tells us the ideal value of log of lambda is 1.27 which is denoted by the vertical dashed line on the plot. When the log of lambda is chosen using the one standard error rule, the ideal number of features is 20. Additionally, Figure 9 lets us see how the highlighted features change as lambda decreases (from left to right). All of the features in the plot are monotonic and do not change sign. The top 10 features were chosen to be highlighted and they are displayed in color.

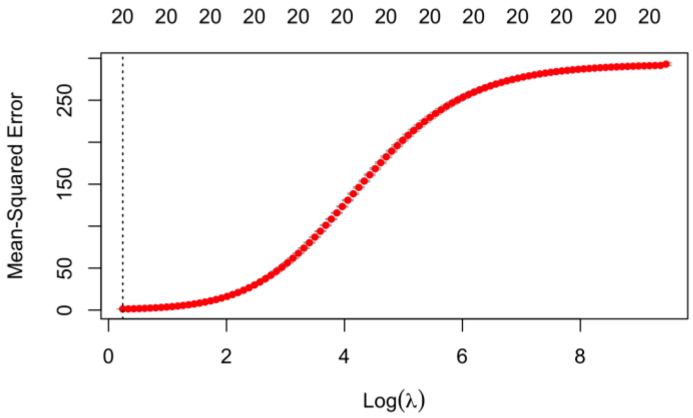


Figure 8: Ridge Regression CV Plot

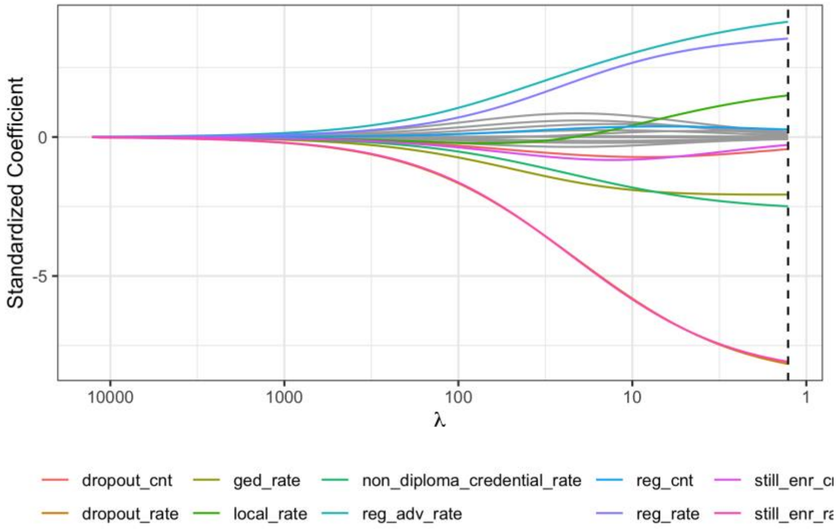


Figure 9: Ridge Regression Trace Plot

Elastic Net

A 10-fold elastic net regression was run on the training data. Elastic Net regression is used to regularize regression method and combines the penalties of ridge and lasso regression. The best elastic net fit was utilized to create the plots. According to Figure 10, the one standard error rule tells us the ideal value of the log of lambda is 0.02. When the log of lambda is chosen using the one standard error rule, the ideal number of features is 9. Additionally, Figure 11 lets us see how the highlighted features change as lambda decreases (from left to right). All of the features in the plot have a coefficient of 0 or less. The top 6 features were chosen to be highlighted and they are displayed in color. Figure 11 highly resembles the trace plot for the Lasso regression, however since the alpha value is 0.216, we know it is not exactly a lasso regression since the alpha for lasso regression is equal to 1.

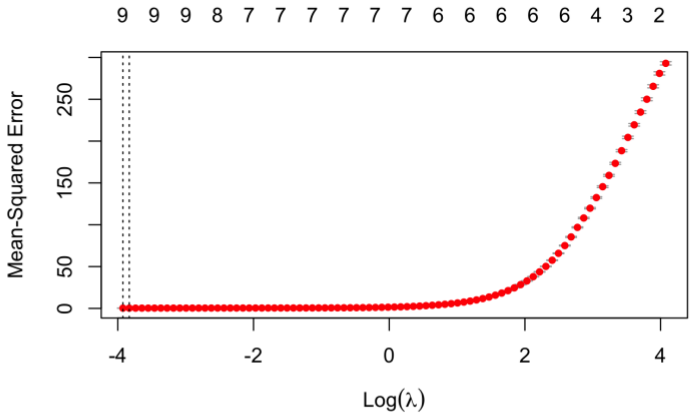


Figure 10: Elastic Net CV Plot

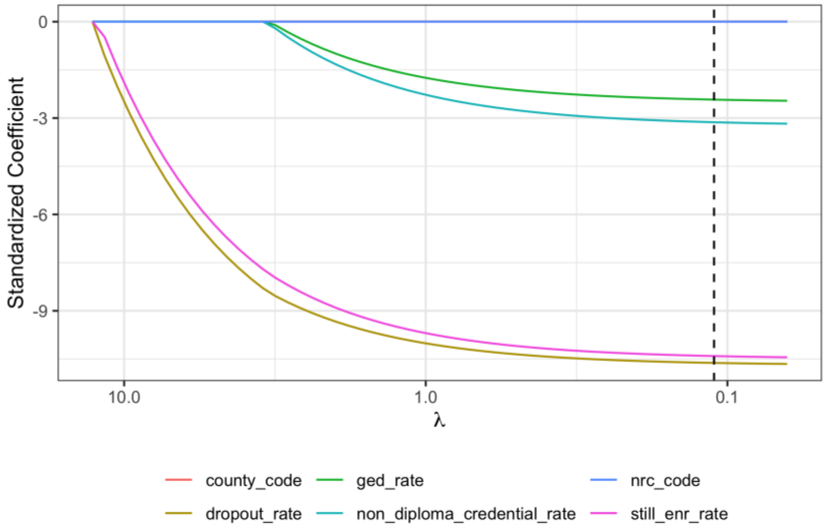


Figure 11: Elastic Net Trace Plot

Model Class 2 (Tree Based Methods)

For tree-based methods, the entire training data set was too large to run analysis on and computer systems would crash. Therefore, a random sampling method chose 1000 observations in the training data to use when running tree-based methods.

Classification Tree

The deepest tree was found as a way to maximize the classification tree and have the most splits. The deepest tree is the classification tree is the tree with the most possible splits, so that every terminal node only has one observation. Figure 12 is the CV plot where only trees with at least two splits are displayed and the x-axis is on a log scale. Using the one-standard-error rule, the optimal tree has 735 terminal nodes and that gives a CV error of 0.906 which can be seen as the dashed, horizontal line in Figure 12.

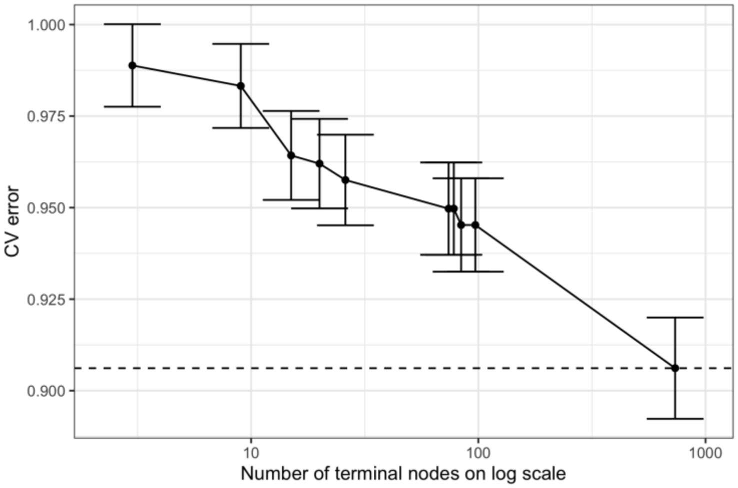


Figure 12: CV Plot for Deepest Tree

Random Forest

Random Forest was run on the training data. Bagging is achieved when all variables are considered at every split. However, that method is not ideal since it leads to high variation, so it is important to find the optimal level of features to train at each split. This number of features is typically labeled as m and in our data, it turns out to be equal to 4. The number of features is determined by running a model that attempts splitting at every number of features from 1 to 18 (the number of features in the model).

Additionally, the out of bag (OOB) error is minimized when number of trees is approximately 125. Figure 13 displays the OOB error as a function of number of trees. This shows how as number of trees increases, the OOB error generally decreases. Additionally, it is promising to see that the OOB error has stabilized by 500 trees and that is seems to actually stabilizes by 200 trees if we wanted to consider a lower number of trees in the model.

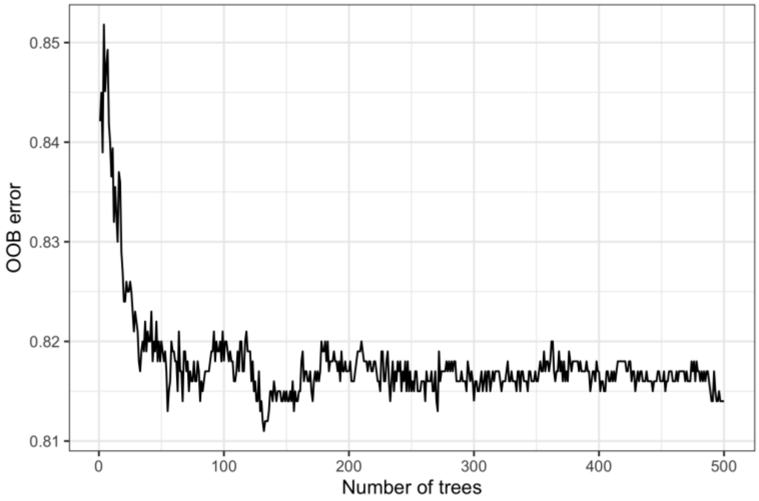


Figure 13: OOB Error

Mean Decrease Accuracy and Mean Decrease Gini were methods used to determine the top features and it can be visualized in Figure 14. For the left plot in Figure 14, the top features were grad count, enrollment count, and dropout rate. For the right plot in Figure 14, the top features were dropout rate, grad count, and enrollment count. The top three features are the same for each plot, however the order differs. We can expect these features to be very critical in determining the graduation rate because, for example, the grad count is a number that is used in the calculation of the graduation rate. When the graduation count increases, keeping all other factors the same, graduation rate will of course increase.

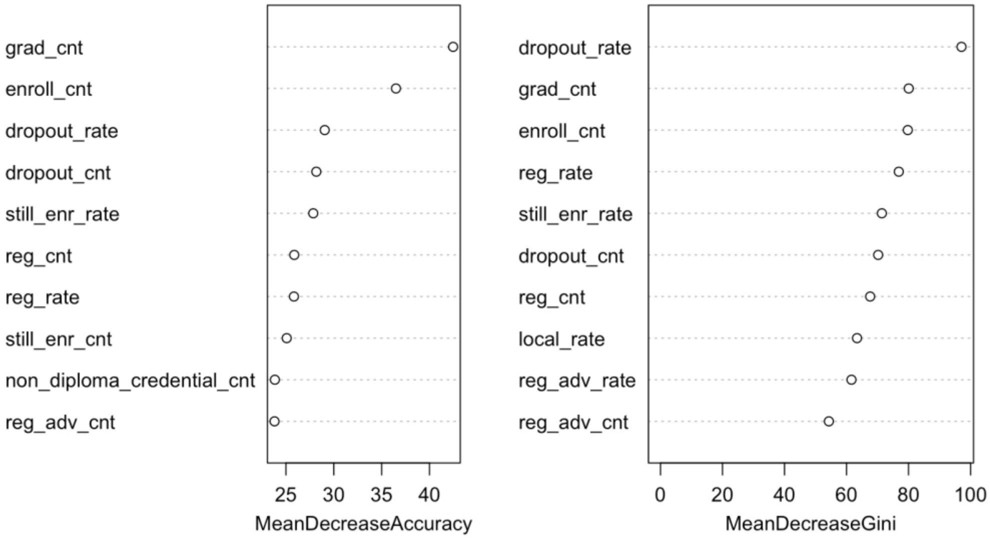


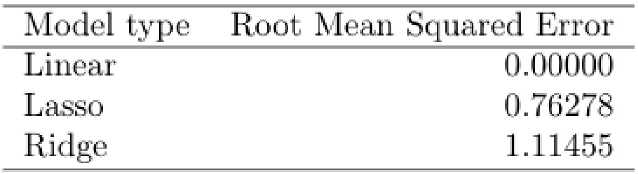
Figure 14: Variable Importance Plot

Conclusions

Method Comparison

Multiple methods and analysis were run in this report to analyze the graduation rate in New York State schools using the variables and factors provided by the New York State Education Department. From the data collected with the multiple methods used to analyze the data, Lasso regression appears to be the best method since it provides the lowest error. For the regression methods, the root mean squared error (RMSE) was used to assess accuracy. According to Table 1, the linear regression model has the smallest RMSE since it is equal to 0; however, the analysis on the linear regression showed that this model was inaccurate to use on this large set of data. Additionally, the linear regression model should not be used because of the correlation between variables in this data. Therefore, we can choose between the Lasso and Ridge regressions which shows us that Lasso gives an RMSE of 0.76 which means that average distance between the predicted model (using the training data) and the actual values (the training data) is 0.76. Broadly, lower RMSE values are ideal because it means the model is a better fit of the dataset.

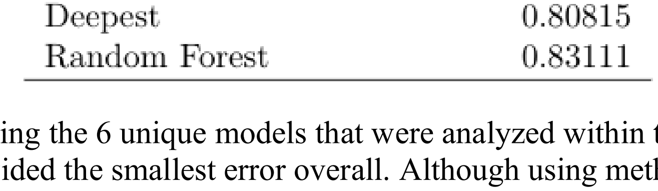
Table 1: These are the test RMSE of the Linear Regression, Lasso Regression, and Ridge Regression



Model type	Root Mean Squared Error
Linear	0.00000
Lasso	0.76278
Ridge	1.11455

With respect to the tree-based methods, the deepest tree possible is the best method for representing the graduation rate data. For the tree-based methods, the misclassification error was used to determine which method is the best predictor. The misclassification error looks at all the observations in the predictive model (training data) and compares them to the actual data (test data) and observes the percentage that are not classified correctly. Therefore, similarly to the RMSE, the lowest misclassification error is ideal. As observed in Table 2, the deepest tree method provided the smallest misclassification error. The deepest tree method is a tree-based method that looks to find the classification tree with the maximum number of splits. In our analysis, the deepest tree method resulted in a misclassification error rate of 0.808.

Table 2: These are the test misclassification errors of the classification tree, deepest tree, and random forest classifiers.



Model type	Misclassification error
Classification	0.89291
Deepest	0.80815
Random Forest	0.83111

Overall, when comparing the 6 unique models that were analyzed within this research, the lasso regression model provided the smallest error overall. Although using methods with both a regression and a tree-based method are useful for comparison, the Lasso method provided the lowest error rate. Likely, the Lasso method performed the best because it is a method that selects variables for its analysis. The Lasso method likely had the lowest error rate since it also includes the shrinkage of coefficients. Therefore, unnecessary factors were not included, and factors' coefficients were shrunken based on influence. This is beneficial when calculating error rates because it minimizes bias and increases variation. This is especially important because Lasso regression runs best when the number of features is high and, in this analysis, the number of features is considerably high.

Takeaways

The results in this report point to a few important factors that contribute to the graduation rate in the New York Education System, that the education department should consider when adjusting policies on graduation rates and when determining how to improve performance. The Lasso regression model, which had the strongest predictive performance, suggests that county code, NRC code, GED rate, non-diploma credential rate, still enrolled rate, and dropout rate are the most important and influential factors when determining graduation rate. County code is a geographical code that corresponds to each specific county within the state. In New York, there are 57 counties each with their own school system. Similarly, the NRC code follows a geographic standpoint as it describes the school districts' ability to fund and distribute necessary resources to all their students. Certain counties, likely in wealthier areas of the state, are more likely to be able to provide better resources and opportunities to their students which in turn will increase the graduation rate. Educational factors such as GED rate, non-diploma credential rate, still enrolled rate, and dropout rate are all factors that describe why students might not graduate. This shows that the school district should analyze reasons students may opt for other degrees (non-diploma or GED) and why students may dropout or take longer than the traditional 4 years of high school (still enrolled rate) to graduate.

Given the lack of demographic factors considered in the LASSO method, a future investigation may look into how certain counties demographics vary and how that might lead to higher or lower graduation rates. Traditionally, counties with a greater number of minorities have lower graduation rates as these counties and school districts are less funded which leads to worse school conditions, less engaged teachers, and fewer resources provided to students. Another factor that would be critical to consider is school district budget. A school district with a large budget is more likely to see their students thrive.

As the world continues to value education, it is important to see high school graduation rates continue to rise. These results suggest that school systems must combat geographical differences and the factors that may come from those if we want to see increased graduation rates. Additionally, school systems should allocate future time into investigating why students might choose other ways to finish their schooling rather than by receiving a high school degree. School administrations might observe a common trend on the type of student who may be more likely to opt for a GED or non-credential diploma and how to encourage these students to receive their high school diploma instead. According to the results of this study, having the resources and abilities to provide equal access to educational resources and educating students on the importance of a high school diploma will be greatly beneficial to encourage growth in high school graduation rates.

Limitations

Dataset Limitations

Unfortunately, the data from this dataset all come from the same school year (2019-2020) so there could be possible skew in the data from this particular year. Specifically, the COVID-19 pandemic truly changed the way traditional schools operated and forced most to engage in online education for at least some period of time. This change in how schools traditionally operated posed many challenges for both teachers, students, and school administration. Students were forced to engage in unfamiliar learning environments, and many had to pick up extraneous jobs (i.e.: caring for sick family members, working additional jobs to account for the high unemployment rate) that caused a major shift in routines and schedules. In addition to the peculiar school year of 2019-2020, the dataset causes some limitations because of the correlation between explanatory variables. Some of the explanatory variables high correlation could lead to masking other important variables without as much correlation. The high correlation of these variables is, unfortunately, hard to eliminate in a report of this type but it is important to understand these correlations as they could be important in future reports and to understanding the validity of this report.

Analysis Limitations

The dataset was split into training and test data sets using a random method to reduce bias; however, this technique could easily lead to randomness in results. The 80-20 split method could have given a specific split that led to this particular data but rerunning the analysis with a different selection of 80-20 split data could always lead to differing results than the ones presented in this analysis. Additionally, the large amount of observations was useful in making sure we had a large picture of the data in question. However, it is important to note that the data was so large some of the models were not able to run on the 80% training data, so it had to be simplified down to just 1000 observations for the sake of getting a result from the model. Although 1000 observations make for quite a large sample, it is not a good representation of the entire dataset and could have led to some analysis that was not fully representative of the entire training set.

Follow-Ups

In future studies, data should be collected from multiple years to see if the trends are yearly. This would be useful in the study because it would provide information on whether certain years were outliers and if the graduation rate was improving over time. Additionally, reports should look at other states besides just New York. Each state may have their own individual factors for graduation rates; however, looking at statewide factors like geography, demographics, and political leanings may be an interesting way to see trends in the data from an entire country perspective. However, this could easily lead to a case with class imbalance, so it is important to address that. Finally, future work should consider including a factor about highest parent education level and household income. These factors can greatly contribute to a students' value of education and could be beneficial resources into how and why students choose to graduate versus dropout of high school.

Appendix

Explanatory Variables

Below are descriptions of all the explanatory variables included in the dataset. The type of variable is listed below the name and description.

- REPORT_SCHOOL_YEAR: school year in which the data were collected
  - Categorical variable
- AGGREGATION_INDEX: numeric index assigned to assist in aggregating data at statewide (0), Need/Resource Category (1), County (2), District (3), and School (4)
  - Categorical variable
- AGGREGATION_NAME: the name of the entity (district of school)
  - Categorical variable
- NRC_CODE: need to resource capacity code, a measure of a district's ability to meet the needs of its students with local resources*
  - Categorical variable
- COUNTY_CODE: 2-digit county code**
  - Categorical variable
- COUNTY_NAME: county name
  - Categorical variable
- MEMBERSHIP_CODE: 1- or 2-digit code corresponding to the cohort membership being reported: 6 – 6-year outcome, June 8 – 5-year outcome, June 9 – 4-year outcome, June 10 – 5-year outcome, August 11 – 4-year outcome, August 18 – 6-year outcome, August
  - Categorical variable
- MEMBERSHIP_KEY: 3-digit code that is specific to the cohort and the school year being reported
  - Categorical variable
- MEMBERSHIP_DESC: description of the cohort membership
  - Categorical variable
- SUBGROUP_CODE: 2-digit code identifying the various demographic subgroups
  - Categorical variable
- SUBGROUP_NAME: name of subgroup. For more information on subgroups, see graduation rate terms glossary on data.nysed.gov
  - Categorical variable
- ENROLL_CNT: count of students in the cohort based on the last enrollment record as of June 30, 2020, with a first date of entry into grade 9 during the 2016-2017 school year, regardless of their current grade level
  - Continuous variable
- GRAD_CNT: number of students in the cohort who earned either a Regents of Local diploma
  - Continuous variable
- LOCAL_CNT: number of students in the cohort who earned a Local diploma
  - Continuous variable
- REG_CNT: number of students in the cohort who earned a Regents diploma

  - Continuous variable
- REG_ADV_CNT: number of students in the cohort who earned a Regents diploma with advanced designation
  - Continuous variable
- NON_DIPLOMA_CREDENTIAL_CNT: number of students in the cohort who earned a non-diploma commencement credential (ex: CDOS credential, Skills & Achievement certificate)
  - Continuous variable
- STILL_ENR_COUNT: number of students in the cohort who were still enrolled as of June 30
  - Continuous variable
- GED_CNT: number of students in the cohort who entered an approved high school equivalency preparation program
  - Continuous variable
- DROPOUT_CNT: number of students who dropped out
  - Continuous variable
- GRAD_RATE: percentage of students in the cohort who earned either a Regents of Local diploma
  - Continuous variable
- LOCAL_RATE: percentage of students in the cohort who earned a Local diploma
  - Continuous variable
- REG_RATE: percentage of students in the cohort who earned a Regents diploma
  - Continuous variable
- REG_ADV_RATE: percentage of students in the cohort who earned a Regents diploma with advanced designation
  - Continuous variable
- NON_DIPLOMA_CREDENTIAL_RATE: percentage of students in the cohort who earned a non-diploma commencement credential (ex: CDOS credential, Skills & Achievement certificate)
  - Continuous variable
- STILL_ENR_RATE: percentage of students in the cohort who were still enrolled as of June 30
  - Continuous variable
- GED_RATE: percentage of students in the cohort who entered an approved high school equivalency preparation program
  - Continuous variable
- DROPOUT_RATE: percentage of students who dropped out
  - Continuous variable

*Need to Resource Capacity Codes Explanation

The Need to Resource Capacity Code is the ratio of the estimated poverty percentage to the combined wealth ratio. A district with both estimated poverty and combined wealth ratio equal to the state average would have an NRC index of 1.

- Code 1: New York City
- Code 2: Large City High NRC Districts
  - Buffalo, Rochester, Syracuse, Yonkers
- Code 3: Urban-Suburban High NRC Districts
  - All districts at or above the 70th percentile (1.1835) that have:
    - At least 100 students per square mile; or

        - An enrollment greater than 2,500 and more than 50 students per square mile

- Code 4: Rural High NRC Districts
    - All districts at or above the 70th percentile (1.1835) that have:
        - Fewer than 50 students per square mile; or
        - Fewer than 100 students per square mile and an enrollment of less than 2,500
- Code 5: Average NRC Districts
    - All districts between the 20th (0.770) and the 70th (1.1835) percentile on the index
- Code 6: Low NRC Districts
    - All districts below the 20th (0.770) percentile on the index
- Code 7: Charter Schools

**County Codes Description

Each number on the list is the county code that corresponds with the county name from the database.

1. Albany
2. Alleghany
3. Broome
4. Cattaraugus
5. Cayuga
6. Chautauqua
7. Chemung
8. Chenango
9. Clinton
10. Columbia
11. Cortland
12. Delaware
13. Dutchess
14. Erie
15. Essex
16. Franklin
17. Fulton
18. Genesee
19. Greene
20. Hamilton
21. Herkimer
22. Jefferson
23. Lewis
24. Livingston
25. Madison

26. Monroe
27. Montgomery
28. Nassau
29. NA
30. NA
31. New York
32. Bronx
33. Kings
34. Queens
35. Richmond
36. NA
37. NA
38. NA
39. NA
40. Niagara
41. Oneida
42. Onondaga
43. Ontario
44. Orange
45. Orleans
46. Oswego
47. Ostego
48. Putnam
49. Rensselaer
50. Rockland
51. Saint Lawrence
52. Saratoga
53. Schnectady
54. Schoharie
55. Schuyler
56. Seneca
57. Steuben
58. Suffolk
59. Sullivan
60. Tioga
61. Tompkins
62. Ulster
63. Warren
64. Washington
65. Wayne
66. Westchester
67. Wyoming
68. Yates