New York School Success

Alex Chen and James Kuang

12/12/21

Contents

The code to reproduce this report is available on Github.

1 Executive Summary

Problem. Graduation rates are one of the most used indications of how well a school is developing their students. Schools are essential to the growth of people in living morally, creatively, and productively in today's society. Despite the national high school graduation rates hitting all-time highs during recent years, some schools are still not seeing the same success. So, for our final project, we have decided to look into the various measures of schools in New York throughout 2019 and analyze which factors were most predictive of high graduation rates in those schools. These factors inform about areas of focus that will contribute most to a school's success in advancing student achievement. Although we are only exploring data encompassed in New York schools, we believe that our results can be extended to schools in other states and the US as due to the variety of schools in New York and their average overall graduation rate.

Data. Our datasets are all pulled from the New York State Education Department (NYSED) and merged together to consider the different types of explanatory variable. A large amount of data for high schools in 2020 were missing, so we focused on 2019. These datasets include both public schools and charter schools and statistics regarding each school's categorized standing, funds and expenditures, characteristics of the staff and students, and the overall demographics of the student population. Our primary response variable will be graduation rate, an illustrative factor of a school's student development.

Analysis. Before exploring our data or analyzing it, we split our data into a training and testing dataset, where the testing dataset is utilized for assessing and comparing our statistical model performances. Then, we explore the data to summarize the main characteristics and learn of any correlations between variables. We built 3 different predictive models: ridge regression, random forest, and boosting. To determine which model was the best predictor, we calculated the root mean squared error for each model, and we determined that the **random forest model???** had the lowest test error.

Conclusions.

2 Introduction

Background. High schools are designed as a center of education and development for their students, and their success can oftentimes be measured by their graduation rates. The importance of high school graduation can never be underestimated, as those who graduate can expect better opportunities in their future. The overall graduation rates of high schools in the US has been gradually increasing and reaching all time highs in recent years at 88% for the 2018-2019 academic year. However, high schools that are most successfully getting students through graduation are spread across certain states, and the average state graduation rates can dip down to as low as 75% as a result.[1] Moreover, although overall graduation rates in the US are at all-time highs, there are still states in the US struggling to achieve that same success.

In addition, the number of individual schools that are considered federally as having low graduation rates, or schools of 100 or more students where fewer than two-thirds earn diplomas in four years, have not been decreasing as overall graduation rates improve. In fact, the number had actually increased in 2016, revealing that it is currently incredibly important to develop methods of improving less successful schools.[2] These large disparities in graduation rates should be a concern within educational policy, and the solution isn't as easy as simply increasing funding. Many schools are limited in resources, so it is important to identify features that are most predictive of graduation rates. Furthermore, a thorough analysis of graduation rates and their respective schools will highlight the key qualities of schools with high graduation rates.

Analysis goals. Many factors are considered to be indicative of a quality high school, such as teacher/student ratios, effective school leadership, and a supportive teaching environment. Furthermore, by building statistical models, we seek to investigate how important these different features are in determining a high school's graduation rate. We hope to find which types of school features are most predictive of graduation rates, and

[1] See High School Graduation Rates By State (2021). usnews.com/education/best-high-schools/articles/see-high-school-graduation-rates-by-state.

[2] Number of High Schools With Low Graduation Rates Is Rising (2018). https://www.edweek.org/teaching-learning/number-of-high-schools-with-low-graduation-rates-is-rising/2018/06.

determine whether these features can be targets of improvement, as variables like student demographics of a school are not feasible to change.

Signficance. Our analysis will contribute to finding ways of improving our education system by finding features that are most important to graduation rates. These findings will produce targets for development and reform in our schools, ultimately with the goal of improved student development and opportunities in our society.

3 Data

3.1 Data Sources

Our data set is a combination of several publicly available data sets from the New York State Education Department's website (data.nysed.gov). Given our purpose, we decided on the Student and Educator Database, the Enrollment Database, and the Report Card Databases for the 2019-2020 school year. These data sets are centered around proving information about the ~1300 high schools as well as a scattering of schools serving other grade levels in the state of New York. This data is collected by the state specifically for parents and, thus, is focused towards featuring data points, such as class size and budget, that would be of interest to parents.

The first data set, the Student and Educator Database, possesses information on the number of each staff, the attendance rate, the suspension rate, and the percentage of free/reduced price lunch students. The second data set, the Enrollment Database, has demographic data for each school. For example, this data set lists the percentage of parents whose parents are in the armed forces. The final data set, the Report Card Database, hosts information on the number of teachers out of certification or inexperienced, fiscal data concerning school budgets, the status of each school, and, most importantly, graduation rate which is our response variable.

3.2 Data Cleaning

We downloaded the data from the data.nysed.gov, but it came in the form Microsoft Access Databases. We manually had to extract the data into excel files preserving column names. On the extracted data, we filtered for data only from 2019. The team made this choice after noticing that much of the 2020 data was incomplete, likely an artifact of delays on the end of NYSED. There were also several columns that were missing most if not all of their data. We decided to also remove those features. Further, some data sets had data for each demographic group within the school in addition to numbers for the entire student body. As this was only present in some of the data, we decided to use only observations for entire school bodies. Once we had filtered the data, we merged them together using the Entity_CD identifier.

3.3 Data Description

3.4 Data Exploration

4 Modeling

4.1 Regression-Based Models

4.2 Tree-Based Models

After modeling our data using regression-based models, we decided that we want more powerful and better predicting models using tree-based models. Although tree-based models can be unstable and worse at predicting certain sets of data, we believe this state-of-the-art prediction method will be able to uncover the most important factors in determining graduation rate of high schools.

4.2.1 Random forest

Random forests represent one tree-based model that have been shown to perform exceptionally in certain sets of data. It uses a random subset method and bagging in order to prevent overfitting, which gives it great predictive power. One issue with random forests is their lack of precision in predicting continuous response variables depending on how well the true model fits a tree-based approach. However, it is important to test and analyze the predictive power of the random forest because of its proven predicting power.

Moreover, there are methods of tuning the random forest to build better random forests that are catered to our data. Although random forests have low interpretability compared to some tree-based models, it measures variable importance which is valuable for our purposes in learning the most important factors for school graduation rates.

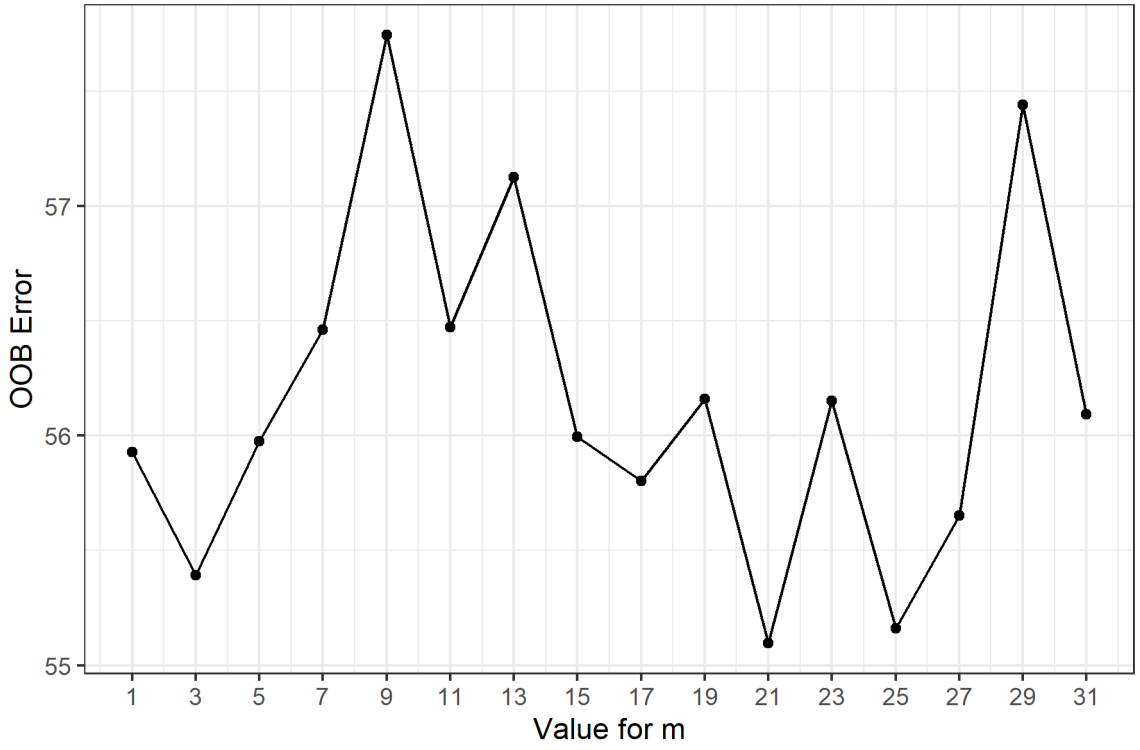

Figure 1: Tuning the value for m in a random forest model

We tuned our random forest models based on variable m: the subset of features at each split point of the tree. By tuning against out-of-bag error, we determine an optimal m-value, as shown in Figure 1. The value that minimized the OOB error was 21, which is much higher than the default m value of 10. With this tuned parameter, we are able to build a better random forest, and this random forest when checked against number of trees to ensure had an out-of-bag error that was stabilized.

Then, we calculate the levels of variable importance using the tuned random forest as shown in Figure 2. The clear most important variable in this random forest is the attendance rate of students at the school. This variable has a much higher value than the rest on both measures of importance, indicating that it is incredibly important to the model and calculating graduation rates. This makes sense because attendance is important to doing well in school, so schools with higher attendance rates tended to have more students perform better and graduate.

For the OOB prediction based importance, the top three measures were attendance rate, overall status of the school, and the percent of the students that were Asian. Although not as importance as attendance rate, overall status and percent Asian also make sense when determining graduation rates. If the school is in good

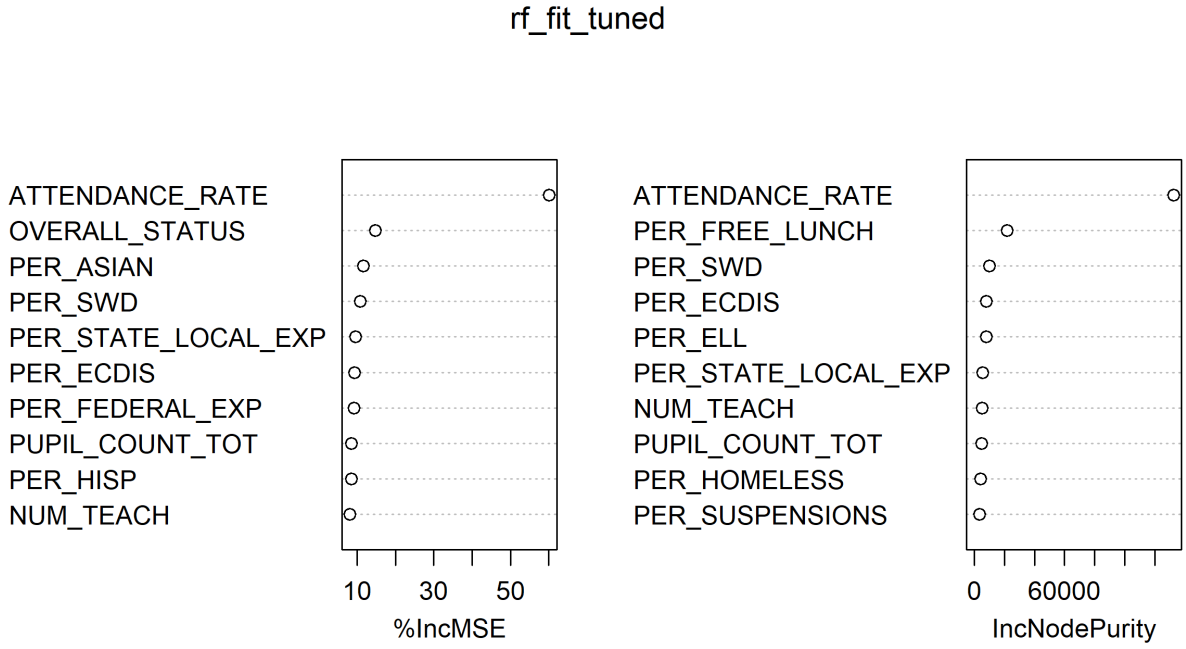

Figure 2: Plot of the variable importance using the purity based importance (on the right) and the OOB prediction based importance (on the left).

standing, then we would expect it to have higher graduation rates. Also, Asians are known to prioritize school and studying, which means that schools with higher percent Asians will have a slight boost in graduation rates. For the purity based importance, we also see percent of students with free lunch and percent students with disabilities being important. In contrast to the others, a higher value in these categories might represent lower income areas, which might lead to lower resources in the school and lower graduation rates. Overall, these values do make intuitive sense as to why they might be predictive of graduation rates.

4.2.2 Boosting

Boosting (gradient boosting) is another powerful tree-based method that we will be building based on these New York high schools graduation rates. This method is known for building great models and models that are better than even random forests. It also works well with continuous response variables, such as the graduation rate in our case. The model also reveals which variables are most important in determining graduation rate, as well as partial dependence plots to go alongside. So, we use boosting in order to build a model that will have great predictive power for graduation rates and will also highlight the important variables that we must consider.

5 Conclusions

5.1 Method comparison

5.2 Takeaways

5.3 Limitations

5.3.1 Dataset limitations

5.3.2 Analysis limitations

5.4 Follow-ups

A Appendix: Descriptions of features

Below are the 31 features we used for analysis. Words written in parentheses represent variable names. Unless noted otherwise, all variables are continuous.

Schools

- *School Standing*
  - Overall Status (`OVERALL_STATUS`): The status of the school: Good Standing, Targeted Support and Improvement, Comprehensive Support and Improvement, Closing/Closing School. (Categorical variable)
  - Needs index (`NEEDS_INDEX`): Need-to-Resource Capacity Category. The need-to-resource capacity (N/RC) index is a measure of a district's ability to meet the needs of its students with local resources.
- *Expenditures*
  - Federal expenditures (`PER_FEDERAL_EXP`): Per pupil expenditures using federal funds
  - State and local expenditures (`PER_STATE_LOCAL_EXP`): Per pupil expenditures using state and local funds
- *Staff*
  - Number of teachers (`NUM_TEACH`): Number of teachers as reported in the Student Information Repository System (SIRS), used for determining the percent of inexperienced teachers
  - Number of principals (`NUM_PRINC`): Number of principals as reported in the Student Information Repository System (SIRS), used for determining the percent of inexperienced principals
  - Number of counselors (`NUM_COUNSELORS`): Total number of school counselors
  - Number of social workers (`NUM_SOCIAL`): Total number of social workers
  - Percent teacher inexperience (`PER_TEACH_INEXP`): Percent of teachers with fewer than four years of experience in their positions
  - Percent principal inexperience (`PER_PRINC_INEXP`): Percent of principals with fewer than four years of experience in their positions
  - Teachers teaching out of certification (`PER_OUT_CERT`): Percent of teachers teaching out of their subject/field of certification

Students

- *Characteristics*
  - Attendance rate (`ATTENDANCE_RATE`): Annual attendance rate
  - Percent suspended (`PER_SUSPENSIONS`): Percent of students suspended
  - Percent reduced lunch (`PER_REDUCED_LUNCH`): Percentage of enrolled students eligible for reduced-price lunch
  - Percent free lunch (`PER_FREE_LUNCH`): Percentage of enrolled students eligible for free lunch
- *Demographics*
  - Percent female (`PER_FEMALE`): Percent of female students (K-12)
  - Percent male (`PER_MALE`): Percent of male students (K-12)

- Percent American Indian (`PER_AM_IND`): Percent of American Indian or Alaska Native students (K-12)
- Percent Black (`PER_BLACK`): Percent of Black or African American students (K-12)
- Percent Asian (`PER_ASIAN`): Percent of Asian or Native Hawaiian/Other Pacific Islander students (K-12)
- Percent Hispanic (`PER_HISP`): Percent of Hispanic or Latino students (K-12)
- Percent White (`PER_WHITE`): Percent of White students (K-12)
- Percent Multi (`PER_MULTI`): Percent of Multiracial students (K-12)
- Percent English language learners (`PER_ELL`): Percent of English Language Learners (K-12)
- Percent with disabilities (`PER_SWD`): Percent of students with disabilities (K-12)
- Percent economically disadvantaged (`PER_ECDIS`): Percent of economically disadvantaged students (K-12)
- Percent migrants (`PER_MIGRANT`): Percentage of migrant students (K-12)
- Percent homeless (`PER_HOMELESS`): Percent of homeless students (K-12)
- Percent foster care (`PER_FOSTER`): Percent of students in foster care (K-12)
- Percent parent armed forces (`PER_ARMED`): Percent of students with a parent on active duty in the Armed Forces (K-12)