Predicting NYC Public School Graduation Outcomes

Nathaniel Lanier
Brown University

https://github.com/ncl11/Data1030_project

Introduction

For this project we will be using the "2005-2010 Graduation Outcomes - School Level" dataset from the NYC OpenData website. The dataset contains a number of variables about the 2001-2006 cohorts broken down on the school level. Cohort is defined as the year the Regents Exam was taken. We will be using the feature variables available in the dataset for a given school and cohort to predict graduation rates for the next year. All features are clearly explained in Feature_Explanations.xlsx. The features that we will consider are Boro, Cohort, Total Cohort, Total Grads - n, Total Grads - % of cohort, Total Regents - n, Total Regents - % of cohort, Total Regents - % of grads, Advanced Regents - n, Advanced Regents - % of cohort, Advanced Regents - % of grads, Regents w/o Advanced - n, Regents w/o Advanced - % of cohort, Regents w/o Advanced - % of grads, Local - n, Local - % of cohort, Local - % of grad, Still Enrolled - n, Still Enrolled - % of cohort, Dropped Out - n, Dropped Out - % of cohort and Income. We drop the school names because there are over 400 unique schools. Percent of students who graduated is a continuous variable so this is a regression problem.

In New York State there are several types of high school diplomas. Regents diplomas are available to those students who pass the Regents exams and satisfy several other requirements. Local diplomas are available to those who are unable to pass the Regents test. These diplomas are significant because they allow students who have trouble passing standardized tests to apply to college and they are accepted by trade unions, the armed forces and many other such institutions who generally require the applicant to be a high school grad. It is important to note that passing the Regents Exam does not guarantee graduation and failing the Regents Exam does not guarantee that one will not receive a diploma. This problem is significant because possession of a high school diploma has a great impact on income and overall life outcomes. Having a model that can effectively predict graduation outcomes for a specific school can help the city shift fiscal and other resources towards schools who need it the most. The dataset originally comes with 25,096 rows but we end up dropping many of these for reasons described below.

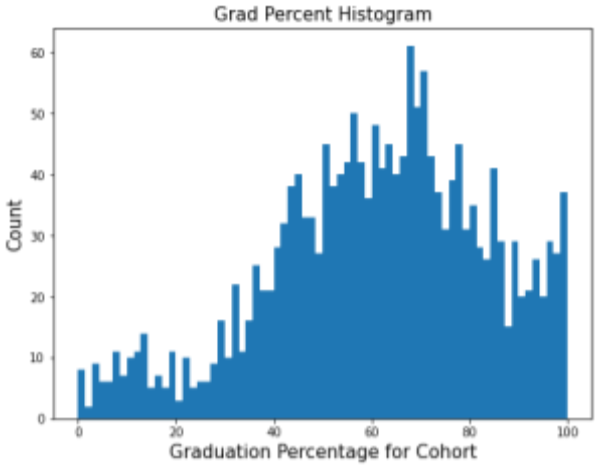

Figure 1. Histogram of graduation percentages for entire cohort over all years of data

Exploratory Data Analysis

Before we begin with EDA we must perform some data cleaning. We notice that for each school within a year there is a Total Cohort row and then a separate row for Asian, Black, White, Hispanic, English Language Learners, English Language Proficient Student, Female, Male, Special Education Student and General

Education. However it is not always clear how these variables intersect. If a student is a mixed race, Male, English Language Proficient, General Education student, for example, they could appear in the dataset five times. Does a student from Egypt not appear in race demographics at all? There are many nuances to consider that can't be ironed out based on the documentation. To solve this issue we filter out rows that are not listed as Total Cohort. Also there is a 2006 cohort and a 2006 Aug cohort in the original dataset. From the documentation and inspecting the dataset we can tell that 2006 Aug is just an updated version of 2006. In order to ensure that we don't have the same data counting twice we drop the 2006 cohort, replace it with 2006 Aug and change the name of 2006 Aug to 2006. We also recognize that the second character of each entry in the DBN row corresponds with the borough of the school. We create a new column with just this character(M-Manhattan, X-Bronx, R-Staten Island, K-Brooklyn, Q-Queens).

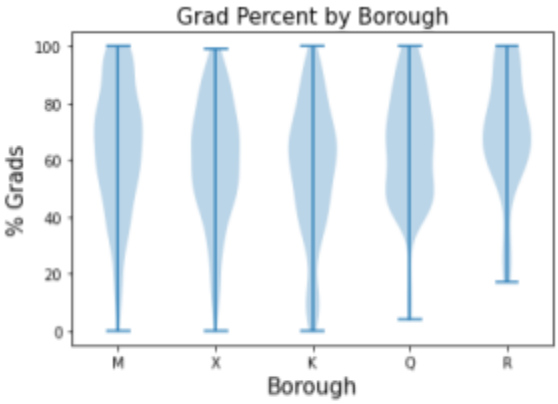

Figure 2. Violin plot of grad percentage broken down by Borough

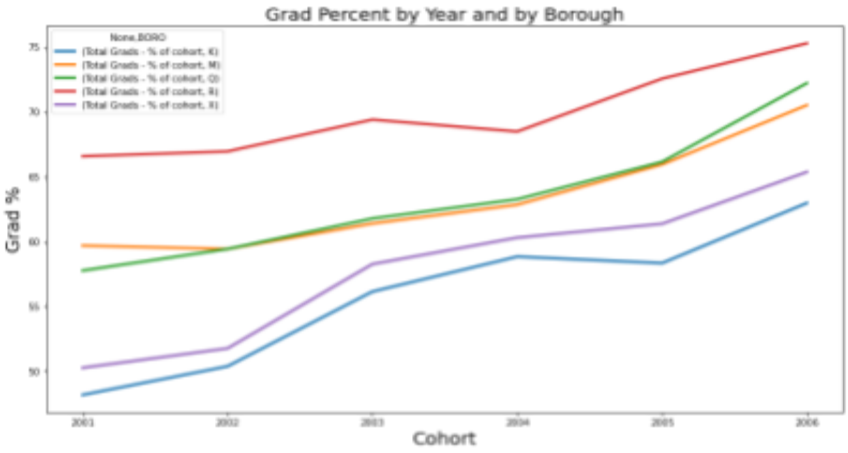

Figure 3. Line plot with year on x-axis, grad percent of the y-axis, broken down by year

During the exploration of our data we notice many trends and relationships between variables that will likely yield significant and predictive results. Firstly, the histogram of graduation percentages above shows a distribution with a mean of approximately 61, a large bump around 100 percent and a decent amount of mass at the lower end of the spectrum. In the violin plot we can see that there is some variation between the distributions when broken down by borough. Staten Island has the highest mean graduation rate and Brooklyn has the lowest. Both Manhattan and Staten Island have a second smaller peak of mass at the higher end of the distribution. We notice a consistent upward trend of graduation rates across all boroughs from 2001-2006 and interestingly in 2002

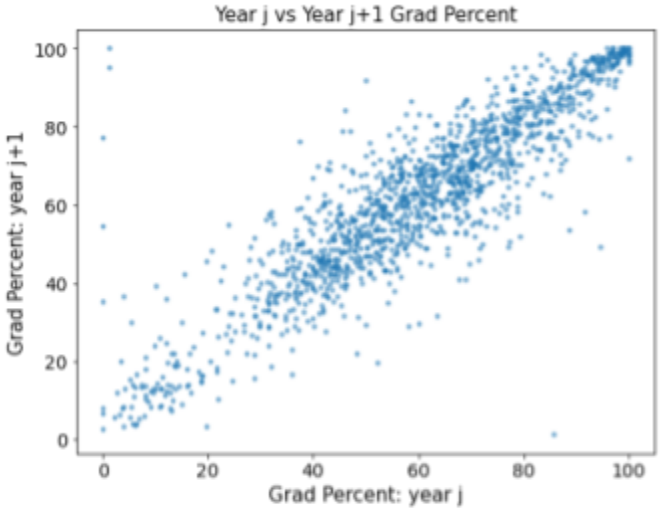

Figure 4 - Scatter plot with Grad Percent for year j on the x-axis and Grad Percent for year j+1 on the y-axis

Queens surpassed Manhattan in graduation rate. This is shown in the line chart on the right. We can tell from this visualization and from computing directly that the mean and variance change as a function of time so our data is not stationary. We have made a scatter matrix of many of our continuous variables in figure 8 shown at the end of the report. In the scatter matrix, y refers to the graduation rates for the next year. The features that have the largest correlation with y in absolute value are Grad Percent(0.89) and Still Enrolled Percent(-0.81). The scatter plot for Grad Percent vs y is highlighted above on the right.

Methods

Given that the data is time series it is non-iid and we split our data for train-test-split based on year. Because at the time of prediction we will have no knowledge about how the data will trend in future years, we must only train on a given subset of the data up until time t and validate on data that is t+1 then test on t+2. We have six years from 2001-2006 so we did three splits. Starting with the year 2002 as validation and 2001 as train, for each split we moved to the next year as our validation data and all the previous years as our training data. Cohort year is the only group structure that we were concerned about because the borough data is distributed reasonably well. Additionally we don't have to worry about separating schools in order to ensure zero school overlap between train, val and test because we aren't concerned with the model performing well on previously unseen schools. This is because the number of schools from one year to the next is fixed. We have applied onehotencoder to our Borough column because it is categorical and standardencoder to our continuous columns. Our final processed dataset contains 23 feature columns. Missing values appeared in two places in our dataset. For purposes of anonymity, graduation data was suppressed for any cohort with fewer than 20 students. Additionally, because we are predicting graduation rates for year j+1 given data from year j, the data from 2006(the latest year) has no target variable available to train on. Because in both of these cases the target variable doesn't exist, our only option was to drop these rows. For this reason the model should only be deployed when the school in question has more

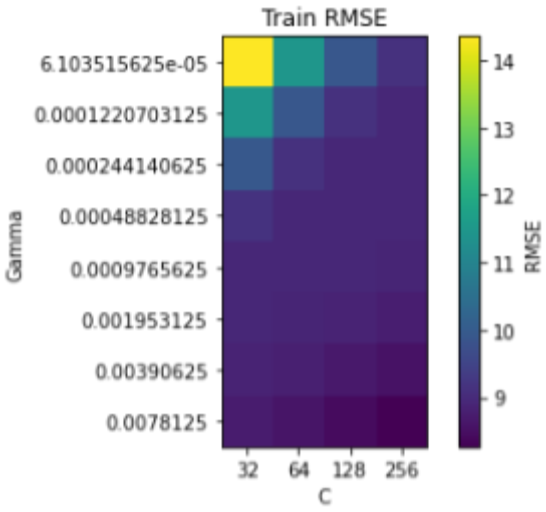

Figure 5 - Parameter grid and RMSE for best SVR model - train

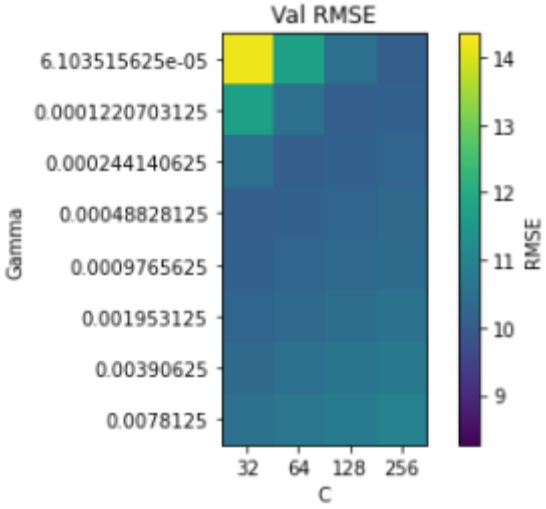

Figure - 6. Parameter grid and RMSE for best SVR model - Val

than 20 students. The model will most likely not be as accurate when used on these smaller specialty schools.

For this model we will use root mean squared error to evaluate our model. We have made this decision because RMSE is much easier to interpret than mean squared error or R2. This is because RMSE is in the same units as our target variable meaning that we have an easy real world interpretation for our train, val and test scores. Specifically the RMSE will tell us how many percentage points off our prediction is on average. With these decisions being made we move to training our models. We trained regular regression, lasso, ridge, elastic net, random forest, support vector machines, k nearest neighbors and XGBoost. We will describe our parameter choices and results for our top three models. The best model was the SVR. The three test scores were approximately 9.92, 10.44 and 9.44. We tuned three hyperparameters for this model, kernel, gamma and C. The best hyperparameters were, kernel: rbf, gamma: 0.0001220703125 and C: 128. Heatmaps of the train and val scores with C and Gamma on x-axis and y-axis respectively are shown above in figures 5 and 6. Our second best model was Lasso regression with test scores of 9.88, 10.78 and 9.48. The only hyperparameter that we tuned was alpha and the best alpha value was .1. A plot of alphas vs RMSE for train and val for the split that was validated on the 2004 cohort can be seen to the left in figure 7. Our third best model was elastic net with test scores of 9.86, 10.56, 9.54. We tuned alpha and l1 and the best values were .1 and 1e-10. Because of the time series nature of our data there was no randomness due to splitting and all of our best three models were completely deterministic. However when training non-deterministic models we did set a random seed to ensure reproducibility. Also its worth noting that we experimented with using sklearn polynomial features for feature engineering but it only seemed to improve our score marginally and we decided that the cost that we pay with respect to interpretability wasn't worth a very slight increase in predictive power. It is also possible that the slight improvement was just due to random variation.

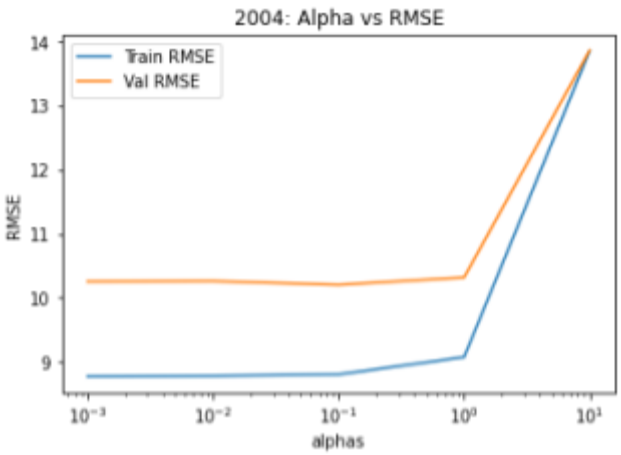

Figure 7 - Alpha vs RMSE for best Lasso model

Results

The test scores for all models listed above were significantly better than baseline. In order to quantify this difference we calculate the mean and standard deviation for all three test scores and the mean and standard deviation for all three baseline scores on the test data. We then calculate the difference of the means divided by the sum of standard deviations. We find that the difference in means divided by the sum of standard deviations is 16.8. This is a very significant improvement. These RMSE scores lead us to believe that this model will be very effective in practice to assist with fund allocation decision making. We have calculated feature importance using permutation and have discovered that the three most important features were Total Grads - % of Cohort, Still Enrolled - % of Cohort and Total Regents - % of Cohort. The top three most important features were unsurprising because they all had very strong correlation with the target variable. However we were

surprised that Still enrolled - % Cohort was more important than Total Regents - % of Cohort by a fairly significant margin because the correlations were similar. This difference in feature importance can be seen in figure 9. We were also surprised to see that several features actually slightly increased the performance(decreased RMSE) of the model when permuted. This is likely an anomaly due to the way the feature vector was permuted.

Outlook

In order to deploy this model there are several steps that we recommend. First we would like to perform the same analysis on more recent data. The 2001-2006 data was the most comprehensive that we could find that was publicly available. We are confident that this same pipeline could be applied to more recent data to build a highly predictive model. Also we think investigating local feature importance would be an important step in deployment. For example if the level of funding for a particular school changed significantly from one year to the next it is likely that the administration at the school would want to know why. Additionally we believe the model could be improved if more feature data was collected on each school. For example feature data such as level of parental involvement, accurate average income data for each school, data on current expenditures on school employees and social workers etc could potentially improve model performance but would require additional funding. All things considered, we are very happy with the performance of the model and are confident that it could play a crucial role in determining allocation of funds for New York City public schools.

Scatter Matrix-Continuous Variables

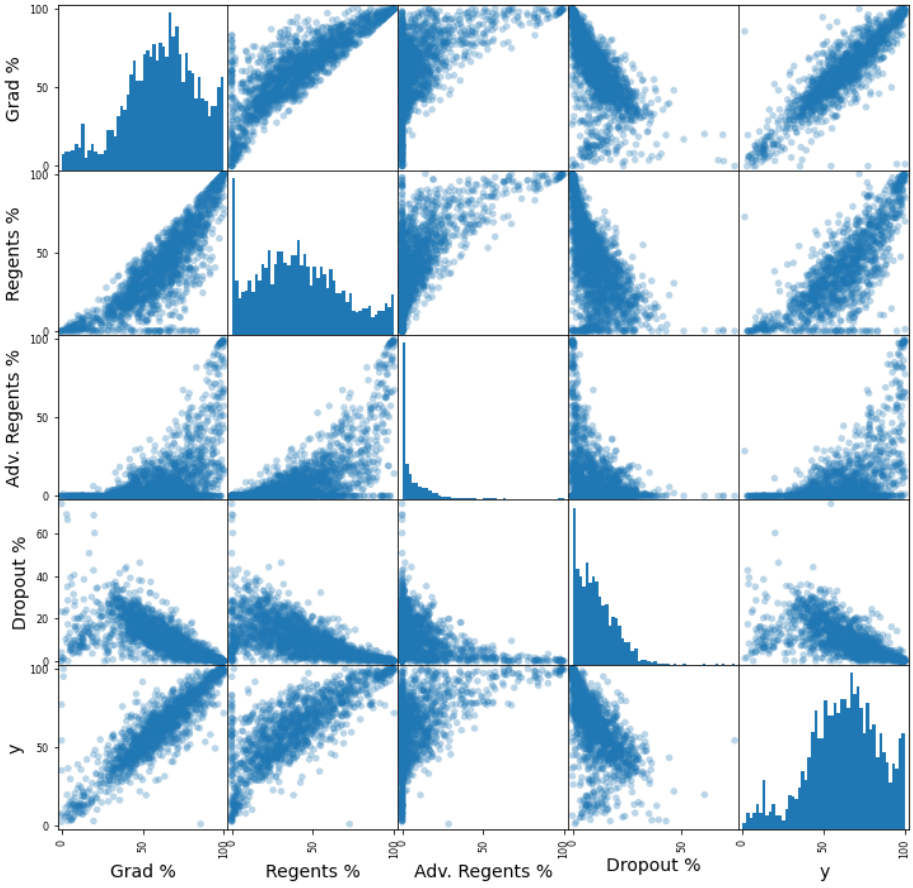

Figure 8- Scatter matrix of most important continuous features being used in the model

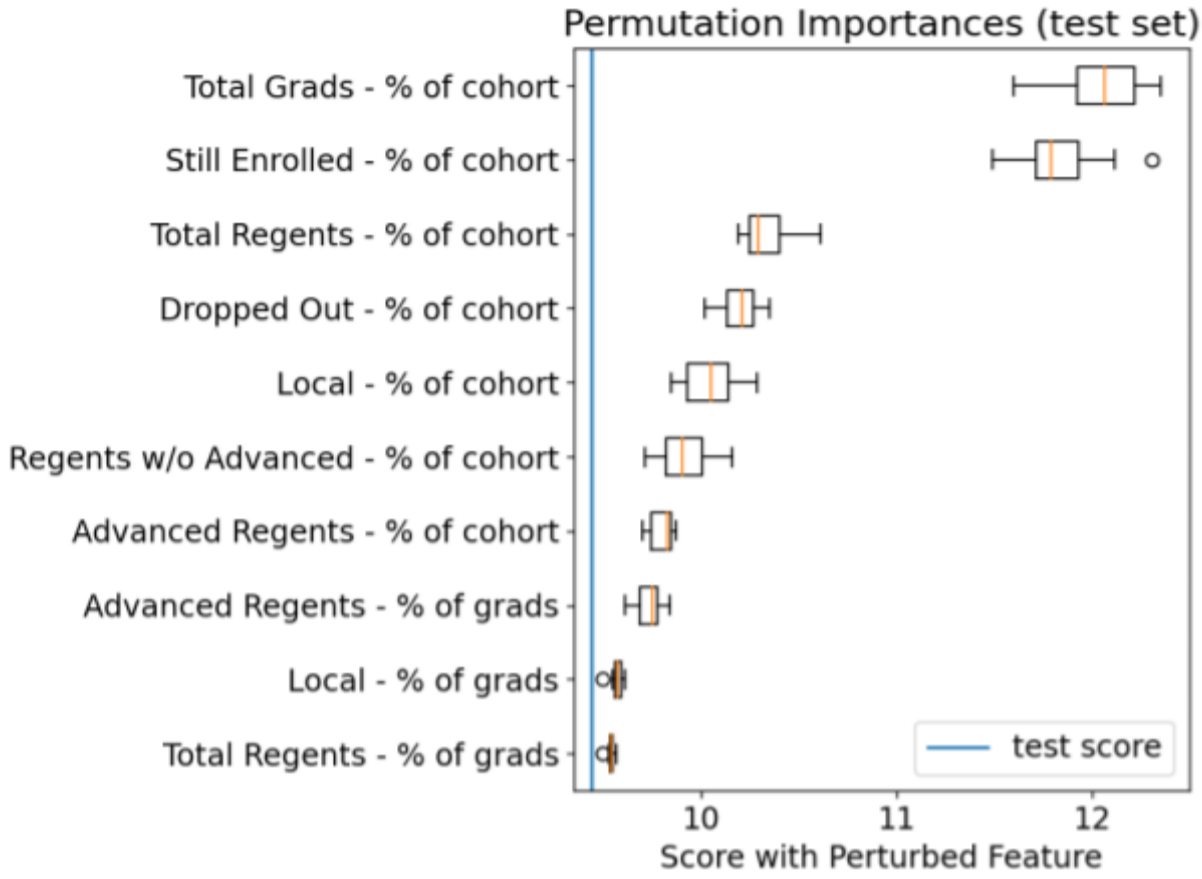

Figure - 9. Top ten most important features computed by permutation compared with pre-permuted test score

References

1. Miyaki, Keita. “Time Series Splits w/ SciKit Learn.” *Medium*, 5 Aug 2019, https://medium.com/keita-starts-data-science/time-series-split-with-scikit-learn-74f5be38489e. Accessed 11 Oct 2020.
2. Bureau of Labor Statistics. “Median weekly earnings $606 for high school dropouts, $1,559 for advanced degree holders.” *BLS*, 21 Oct 2019, https://www.bls.gov/opub/ted/2019/median-weekly-earnings-606-for-high-school-dropouts-1559-for-advanced-degree-holders.htm. Accessed 11 Oct 2020.
3. Walsh, Marion, and Sandi Rosenbaum. “Diplomas and Credentials Available to Students with Disabilities and their Impact on Eligibility for Higher Education and Other Opportunities.” *Special Needs New York*, 19 March 2014, https://www.specialneedsnewyork.com/2014/03/diplomas-and-credentials-available-to-students-with-disabilities-and-their-impact-on-eligibility-for-higher-education-and-other-opportunities/. Accessed 11 Oct 2020.

4. R G, Rajan. “Time Series - Exploratory Data Analysis & Forecast.” *Kaggle.com*, 2017, https://www.kaggle.com/rgrajan/time-series-exploratory-data-analysis-forecast. Accessed 11 Oct 2020.
5. “2005-2010 Graduation Outcomes - School Level.” 18 10 2011, https://data.cityofnewyork.us/Education/2005-2010-Graduation-Outcomes-School-Level/vh2h-md7a.
6. Cccnewyork.org. https://data.cccnewyork.org/data/map/66/median-incomes#66/39/3/107/3/a/a, Accessed 2 Dec 2020