**Data Description and pre-processing steps:**

The main dataset we used includes a combination of performance and descriptive variables about high schools in Chicago (Chicago\_Public\_Schools\_-\_School\_Progress\_Reports\_SY1617 and Chicago\_Public\_Schools\_-\_School\_Profile\_Information\_SY1617). We added a variable that provides a number of libraries within 0.5mi of school location using ‘Libraries-Locations, Hours and Contact Information’. We also enriched the data with census data using ‘Census Data – Languages Spoken in Chicago, 2008-2012’ and ‘Census Data – Selected socioeconomic indicators in Chicago, 2008-2012’. All datasets were extracted from the Chicago Data Portal (data.cityofchicago.org).

Our question of research is attempting to predict graduation rates of a Chicago High School base on school’s qualitative variables and various socioeconomic variables. One of the assumptions we make is that socioeconomic measures have not changed too much since 2012, or if there is a change it’s proportional across neighborhoods.

Before diving into analysis, we had to do some data pre-processing:

1. Merging all the datasets into one file
2. Filtering out schools that don’t have information regarding graduation rate, our variable of interest
3. Adjusting counts of students to rates
   1. The original dataset reported volumes of students in the low-income group, special ed group, different ethnic backgrounds, etc. We used those counts, along with total count of students in the school to calculate rates.
4. Calculating YOY change for variables reported in year 1 (2015) and year 2 (2016)
5. Some variables had up to 10 null values, we filled in those as follows:
   1. YOY change variables were filled in with zeroes, assuming no change YOY
   2. All other continuous variables were filled in with mean values
   3. Categorical variables were filled in with ‘U’ for Unknown

Following data pre-processing we created three separate datasets:

1. Original values
2. Original continuous values and dummy values for categorical data
3. Scaled values using min-max from 0 to 1

**Exploratory Analysis:**

Our data set consists mostly of neighborhood schools, followed by charter schools. Those two groups out of 10 represent over half of the schools we have in our data set. Citywide-Option and Special Ed schools have graduation rate average significantly lower than other categories of schools.

Average income in the school’s neighborhood varied and some had statistically significant differences. Contract schools had a higher average income and was significantly higher that average income in Career Academy, Magnet, Small and Special Education schools. Going along with income, the lowest average hardship index was associated with Contract schools, which was significantly lower that that of Career Academy and Special Ed schools.

Library count within 0.5mi was the highest next to Contract Schools, but the difference from other schools was not statistically significant.

Most schools had below average student growth, followed by average results. Rating status was mostly Good for the schools in our data set.

Among neighborhoods that had non-English language as predominant, Spanish was the most common. There were no significant differences in the graduation rate among neighborhoods with different non-English predominant languages.

Graduation rate appears to have a strong positive correlation with College\_Enrollment\_Rate\_School (0.65). Student\_Special\_Ed\_Pct has strong negative correlation with graduation rate (-0.61). Few variables have weak positive correlation with graduation rate: College\_Persistence\_School\_Pct\_Year\_2 (0.47), Attainment\_ACT\_Grade\_1 (0.45), Growth\_ACT\_Grade\_11\_Pct (0.34), Average\_ACT\_School (0.46), Student\_Count\_Total (0.39).

Graduation Rate, College enrollment and College Persistence vs. Student count/Avg ACT/Growth ACT/Attainment ACT look like exponential relationships. Student Special Ed has negative correlation with all top selected variables. ACT related variables have positive correlation among each other, as expected. There is also a positive correlation of ACT variables with student count.

Graduation Rate and College Enrollment Rate have a distribution that is left skewed. Student Special Ed, ACT result variables and Student Count are all skewed to the right. College Persistence variable is approximately normally distributed.

**Random Forest Feature Importance Analysis:**

According to the Random Forest model most important model variables are:

* Student\_Growth\_Rating\_NO DATA AVAILABLE
* Average\_ACT\_School
* Bilingual\_Services\_U
* Attainment\_ACT\_Grade\_11\_Pct
* Overall\_Rating\_Inability to Rate
* Student\_Special\_Ed\_Pct
* Student\_Attainment\_Rating\_NO DATA AVAILABLE
* Refugee\_Services\_U
* College\_Persistence\_School\_Pct\_Year\_2
* School\_Type\_Citywide-Option

**PCA Analysis:**

Our dataset has 28 continuous variables. Using PCA with 10 components we were able to narrow down our number of variables to 10, while still accounting for 91% of variability in the data. First two components accounted for over 50% of the variability.

Going forward we did regression analysis using PCS components and using actual continuous variables in combination with dummy variables. We analyzed our models with and without feature selection.

**Regression Model Analysis (Approach 1):**

We did few approaches in analyzing graduation rates. Our top models have r-squared result over 0.7.

With PCA first subset of models did not have any feature selection. The best result was achieved by Lasso (r-squared of 0.724) with alpha=0. 5264. Ridge performed slightly worse, with r-squared being 0.021 less than lasso. SGD was the worst performing model with r0squared result of 0.506. Elastic Net confirmed that Lasso is the best selection, returning alpha=0.5264 with L1 ratio = 1, essentially selecting Lasso. Without feature selection the model included all the variables available, which was 174 of them.

In the second PCA subset, we applied feature selection using best model identified above (Lasso with alpha 0.5264). Optimal percentile of features came out to be 32. However, r-squared increase from 9th percentile to 33rd was only 0.012, while having only 16 variables in the model instead of 55. To make this model more efficient and easier to interpret, we selected 9th percentile instead of 33rd.

Final model was Lasso (alpha=0.5264) with 16 variables and it’s r-squared was 0.714.

The list of variables in descending order are:

* Student Growth Rating – No Data Available (negative impact)
* Rating Status –Intensive Support (negative impact)
* PC1 (positive impact)
* Refuge Services – Unknown (negative impact)
* PC4 (positive impact)
* Refugee Services – No (positive impact)
* Overall Rating Level 1+ (positive impact)
* School Type – Citywide Option (negative impact)

From the above list, it looks like the schools that have little data collected about them (could be something not captured by the data set we had) did not perform as well when it comes to graduation rates. Two variables with ‘No Data Available’ and ‘Unknown’ selections had negative impact on the graduation rates. Intensive support variable also had negative impact on graduation rates, which goes in line with out exploratory analysis, where special ed schools had significantly lower graduation rate.

PC1 consists primarily of positive loadings from ACT variables, College persistence, and Hispanic and White ethnicity student ratios. This means that those values have positive impact on graduation rate. Negative loading in PC1 included percent HH below poverty, percent aged 16+ unemployed, hardship index, low income student ratio, and black ethnicity student ratio. This tells me that areas with high unemployment and low income tend to impact schools graduation rate in a negative way.

PC4 is drive by positive loading from ACT results, college enrollment, college persistence, percent 16+ unemployed, percent aged under 18 and over 64, teacher attendance change. Variable percent 16+ unemployed coupled with percent aged under 18 and over 64, make me think that it is more likely to describe areas with high ratio of full-time students. In this case, PC4 also suggests that higher ACT results also have a positive impact on graduation rate, which is expected. One new variable that surfaced is change in teacher attendance rate from 2015 to 2016, which also has a positive impact on graduation rate. PC4 also includes negative loadings from Student low income percent and student English learners’ percent. This suggests that schools in the areas with more students whose English is not first language, coupled with low income are at the risk of lower graduation rate.

**Regression Model Analysis (Approach 2):**

Second approach did not use PCA and included a series of models with and without feature selection. Scoring is done on RMSE instead of r-squared. R-squared of two best models are compared at the end.

Without feature selection best model came out to be Lasso with alpha=0.1438. This selection is in line with above approach suggesting that lasso is the best modeling selection for this data.

Performing feature selection further improved performance of Lasso, when optimal percentile (31) is used. This model ended up including 60 out of 192 available variables.

Some of the negative impact variables are:

1. Student Growth Rating - No Data Available
2. School Type – Citywide Option
3. Overall Rating – Level 3
4. Overall Rating – Inability to rate
5. School Survey Quality of Facilities – Very Strong
6. Rating Status – Intensive Support
7. School Survey Collaborative Teachers – Strong
8. School Survey School Community – Very Strong

Some major positive impact variables are:

1. College enrollment rate
2. Attainment ACT
3. Average ACT
4. Rating Status – Good Standing
5. School Type – Military Academy
6. School Survey Collaborative Teachers – Very Strong
7. School Survey parent teacher partnership – Neutral

Having too many variables makes for some counterintuitive results, like having negative impact on graduation rate from high scores related to quality of facilities, collaborative teachers, and school community.

**Comparison of two best models:**

Some common themes between two best models from two approaches include:

1. Schools with no data available for student growth rating tend to have lower graduation rate.
2. Rating status – intensive support has negative impact on school’s graduation rate.
3. School Type – Citywide option has lower graduation rate.
4. High ACT results correlate to higher graduation rate.
5. Higher college enrollment rates correlate to higher graduation rates.