**Springboard--DSC**
**Capstone Project 3**

# Understanding School Performance on Pennsylvania Standardized Tests

By Christopher Chung
February 2022

# 1. Executive Summary

Pennsylvania's public school system is home to a wide range of schools in terms of performance, demographics, and funding. It's no question that there exist sharp race/ethnic differences in student achievement. There are significant achievement gaps based on family economic status as well. Given this background information, the goals of this project are:

1. To investigate the magnitude of gaps in school performance across multiple demographic groups.
2. To understand what other factors are associated with school performance (including funding and expenditures per student).
3. To create a model that will establish specific connections between the independent variables and the dependent variable, which is the percentage of students that pass (proficient or above) in each school. Two models will be created, one for Mathematics assessment and the other for English Language Arts (now referred to as ELA) assessment.

To measure school performance, we'll rely on data from the Pennsylvania Department of Education on the percent proficient or above in Math and ELA for each school in Pennsylvania. The Pennsylvania System of School Assessment (PSSA) is a standards-based, criterion-referenced assessment which provides students, parents, educators, and citizens with an understanding of student and school performance related to the attainment of proficiency of academic standards. Every Pennsylvania student in grades 3 through 8 attending district schools, charters, and cyber charters is assessed in English Language Arts and Math.

By using data from the 2018-2019 school year on school information, demographics, and expenditures, we trained two models, one to predict the percent proficient or above in ELA and the other to predict the percent proficient or above in Math. Our best model for ELA was XGBoost, with an $r^2$ score of 0.81, a mean absolute error of 6.41, and a 95% worst case interval for residuals from −16.45 to 15.87. Our best model for Math was also XGBoost, with an $r^2$ score of 0.72, a mean absolute error of 8.91, and a 95% worst case interval for residuals from −22.09 to 23.24.

After analyzing the impact of features on our models using the SHAP library[1], we discovered that the most important features were the percentage of economically disadvantaged students in a school, the percentage of black students, the percentage of special education students, and the percentage of students that had regular attendance records. We will discuss the significance of these results and our subsequent recommendations later in this report based on SHAP values and counterfactual explanations combined with the Bayesian Optimization package.[2]

The intended stakeholders of this project are school administrators and policy makers in the Pennsylvania Department of Education who are interested in improving the performance of underperforming schools. This problem is relevant to them because PSSA results factor heavily into the School Performance Profile, which is the state's school rating system. Furthermore, school scores are used by districts in decisions about school turnaround interventions and school closings, as well as in the charter school renewal process.

The implementation details can be found in the notebooks, along with all the project deliverables, in this GitHub repository.

[1] https://shap.readthedocs.io/en/latest/index.html
[2] https://github.com/fmfn/BayesianOptimization

# 2. Approach

**Data Acquisition and Wrangling**

We used two datasets from https://futurereadypa.org, which contains a collection of school progress measures related to school and student success, created by the Pennsylvania Department of Education.

The first dataset is on School and Performance Data for the 2018-2019 School Year. The level of granularity for this dataset is by school and it contains general and demographic information for each school including (but not limited to):

- School name
- School ID
- School zip code
- Enrollment
- Percent male/female
- Race/ethnic composition of school
- Percent economically disadvantaged
- Percent of gifted students
- Percent of students in foster care
- Percent of students that are homeless
- Target: **Percent Proficient or Advanced in ELA**
- Target: **Percent Proficient or Advanced in Math**

The second dataset is on School Fiscal Data for the 2018-2019 School Year. The level of granularity here is also by school and this dataset contains information about the expenditures per student from various sources for each school, including:

- School ID
- Federal - Non-Personnel
    - *The federal amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on non-personnel.*
- Federal - Personnel
    - *The federal amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on personnel.*
- Local - Non-Personnel
    - *The local amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on non-personnel.*
- Local - Personnel
    - *The local amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on personnel.*
- State - Non-Personnel
    - *The state amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on non-personnel.*
- State - Personnel
    - *The state amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on personnel.*

After an initial inspection, we realized that both of our datasets were in a “narrow” format, where feature values were listed in separate rows for each school, as shown below.

| | DistrictName | Name | AUN | SchI | DataElement | DisplayValue |
|---|---|---|---|---|---|---|
| **0** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | 2 or More Races | 5.67 |
| **1** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | American Indian/Alaskan Native | 0.32 |
| **35** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | Asian | 1.46 |
| **36** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | Black/African American | 10.36 |
| **37** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | Economically Disadvantaged | 17.25 |
| **39** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | English Learner | 0.24 |
| **127** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | Female (School) | 60.65 |
| **128** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | Foster Care | 0.08 |
| **130** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | Hispanic | 7.85 |
| **131** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | Homeless | 2.19 |
| **134** | 21st Century Cyber CS | 21st Century Cyber CS | 124150002 | 7691 | Male (School) | 39.35 |

| | DistrictName | SchoolName | SchI | AcademicYearId | AUN | DataElement | DisplayValue |
|---|---|---|---|---|---|---|---|
| **0** | Albert Gallatin Area SD | Albert Gallatin Area SHS | 6001 | 19 | 101260303 | Federal - Non-Personnel | 35.22 |
| **1** | Albert Gallatin Area SD | Albert Gallatin North MS | 7607 | 19 | 101260303 | Federal - Non-Personnel | 41.89 |
| **2** | Albert Gallatin Area SD | Albert Gallatin South MS | 7608 | 19 | 101260303 | Federal - Non-Personnel | 41.81 |
| **3** | Albert Gallatin Area SD | Friendship Hill El Sch | 4922 | 19 | 101260303 | Federal - Non-Personnel | 160.13 |
| **4** | Albert Gallatin Area SD | George J Plava El Sch | 2129 | 19 | 101260303 | Federal - Non-Personnel | 108.38 |

Our first step was to convert these long format datasets to a wide format, where each row corresponds to one school and the features are shown in the corresponding columns of that row. We achieved this by using pandas .pivot() function to reshape each dataframe. Then we joined these two datasets on School ID. We used an inner join because we were only interested in schools that had information in both datasets. Onced merged, our dataframe had 2804 rows and 32 columns. The first few rows of the resulting dataframe is shown below.

| | SchoolName | SchoolNumber | School Zip Code | DistrictName | AUN | Title I School | School Enrollment | Male (School) | Female (School) | Percent Regular Attendance (All Student) | ... | Homeless | Military Connected | Local Non-Personne |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 21st Century Cyber CS | 7691 | 19335 | 21st Century Cyber CS | 124150002 | No | 1235 | 39.35 | 60.65 | 99.8 | ... | 2.19 | 0.73 | 0.0 |
| **1** | ASPIRA Bilingual Cyber Charter School | 8148 | 19140 | ASPIRA Bilingual Cyber Charter School | 181519176 | Yes | 365 | 54.79 | 45.21 | 68.1 | ... | 1.92 | 0 | 15709.0 |
| **2** | Abington Heights HS | 5091 | 18411 | Abington Heights School District | 119350303 | No | 1031 | 53.83 | 46.17 | 86.9 | ... | 0 | 0 | 2756.5 |
| **3** | Abington Heights MS | 6839 | 18411 | Abington Heights School District | 119350303 | Yes | 1119 | 52.28 | 47.72 | 90.9 | ... | 0.09 | 0 | 2722.0 |
| **4** | Clarks Summit El Sch | 7570 | 18411 | Abington Heights School District | 119350303 | Yes | 339 | 57.52 | 42.48 | 96.2 | ... | 0 | 0.88 | 2763.4 |

Now that the dataframe is established, we then checked for null values. In this scenario, we will make the assumption that we have informed the client and stakeholders of all null values that we discovered and have consulted with them regarding how to handle them.

There were five schools in our data frame with null values in almost every feature. We dropped these rows because it did not make sense to impute this many features in these observations. After that, there were only three columns that had null/invalid values:

- Percent Regular Attendance (All Student)
- **Target**: Percent Proficient or Advanced ELA/Literature (All Student)
- **Target**: Percent Proficient or Advanced Mathematics/Algebra 1 (All Student)

There were 126 rows that were missing the target values. After searching for additional information on these schools, we recognized that these were primarily early education institutions, serving grades 2 and below. The PSSA begins in the 3rd grade, which is why there was no data for our targets. As a result, we dropped these rows.

There were 25 rows that listed Percent Regular Attendance as 'IS'. We *could* look this information up on https://futurereadypa.org for the current school year. However, since this is a study using data from the 2018-2019 school year, the current data may not accurately represent the missing data from the 2018-2019 school year. In a real situation, we would first consult our client to see if the abbreviation **IS** means anything. This will give us a better understanding of the nature of this 'missing value', whether it's missing at random or missing for a reason. Assuming we have consulted with the client, we will impute these values with either the mean, median, or mode of this column so that we can keep these rows. To determine which is the better measure of central tendency, we constructed a histogram for this column as shown below in **Figure 1**.

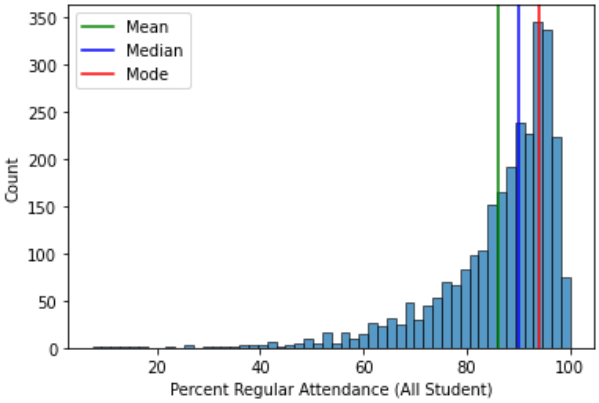


**Figure 1:** Distribution of Percent Regular Attendance

Given the skewed nature of this distribution, we decided to use the mode to impute the missing values of 'Percent Regular Attendance'.

Now that our dataframe no longer has any missing values, we checked our numeric features for any outliers by using the pandas .describe() method. Given that most of our features represented percentages, we knew that the minimum and maximum values of these features had to be 0 and 100, respectively, and this was true for all of our percentage columns. When looking at our expenditure columns, we realized that there were three features that had a minimum that was negative. Those three features were:

- Local - Non-Personnel
- State - Non-Personnel
- Federal - Non-Personnel

This affected 4 rows in our dataframe. In a real situation, we would consult with the client to gain further information on these four schools and their values for expenditures. It's possible that these negative numbers are correct and are representative of something significant in these schools. Or it's possible that the numbers should just be interpreted as 0. It's also possible that the numbers were incorrectly entered as negative, and should be positive. Given that we are unable to determine which situation we are in, plus the fact that it only affects a small portion of our dataframe, we will leave these values as is.

We also saw abnormally large maximum values in three of our features, where these maximum values were significantly larger than the value in the 75th percentile of these features. These features were:

- Local - Non-Personnel
- Local - Personnel
- State - Personnel

To visualize the extent to which these values may be outliers, we graphed the histograms of these three features, as shown below in **Figure 2.**

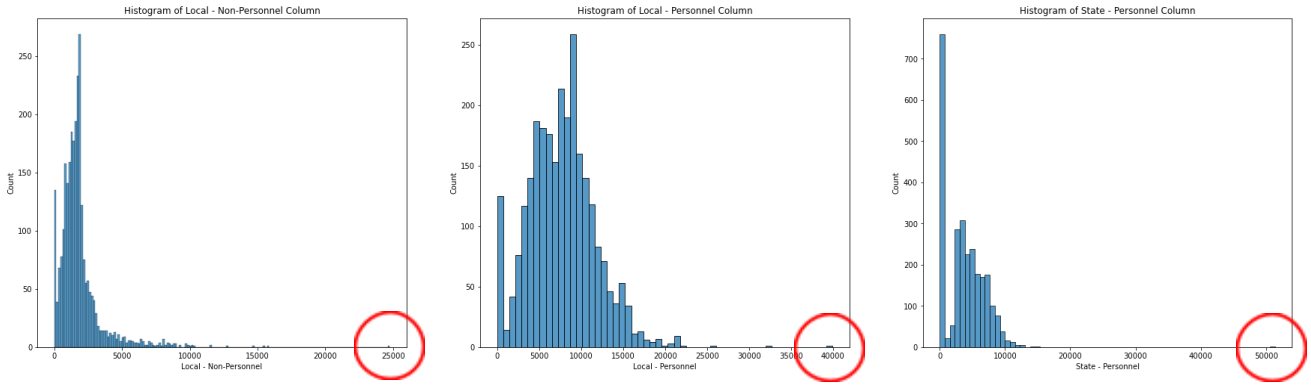


**Figure 2:** Distributions of Local - Non-Personnel, Local-Personnel, State - Personnel expenditures

Certainly these three features do seem to have outliers on the high end of their distributions. This gave us some concerns about the validity of these values, so we did some additional research. We identified the schools that had these large values of expenditure features and looked them up on SchoolDigger. According to this website, these abnormally large values were actually valid. As a result, we decided to leave them in our dataset.

Our cleaned dataset has 2673 rows and 32 columns, where each row corresponds to a unique school.

### Exploratory Data Analysis and Initial Findings

Our exploratory data analysis was driven by questions regarding the data, the distributions of features, and the relationships of the features with our target variables. The first part of our EDA was done in a jupyter notebook, using Python packages including pandas, matploblib.pyplot, and Seaborn. The second part of our EDA was done in Tableau. The link to the slides on Tableau Public is included here, and the Tableau workbook file is included in the GitHub repository.

We started by engineering additional features based on our expenditure (per student) variables. We added expenditures by groups to create six new features as follows:

| | | | | | | |
|---|---|---|---|---|---|---|
| Local - Personnel | + | State - Personnel | + | Federal - Personnel | | **Total Personnel Expenditure** |
| + | | + | | + | | **+** |
| Local - Non-Personnel | + | State - Non-Personnel | + | Federal - Non-Personnel | | **Total Non-Personnel Expenditure** |
| **Total Local Expenditure** | **+** | **Total State Expenditure** | **+** | **Total Federal Expenditure** | | **Total Expenditure** |

*From which sources do schools tend to spend more money per student? Local, state, or federal? Do schools tend to spend more money per student on Personnel or Non-Personnel?*

To address this, we created boxplots for expenditures separated by expenditure source to compare, as well as boxplots separated by expenditure categories, as shown below in **Figures 3 and 4**.

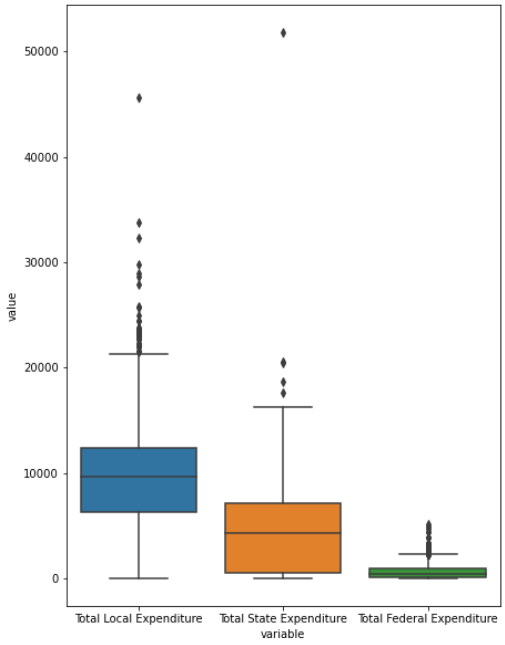


**Figure 3:** Boxplots of Expenditure by Source

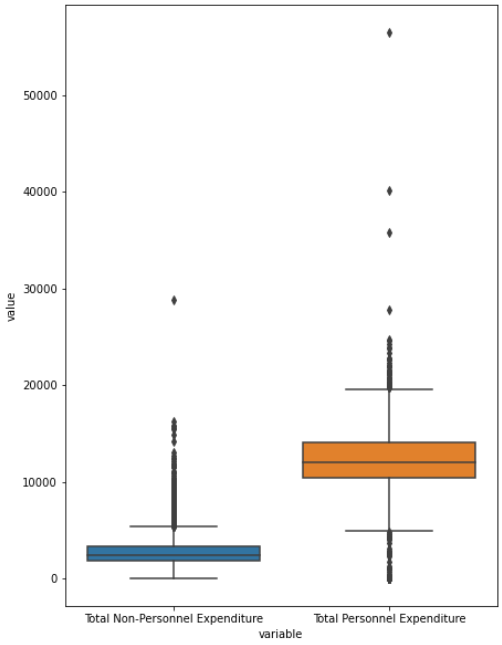


**Figure 4:** Boxplots of Expenditure by Category

As evident in the boxplots above, on average schools seem to spend the most amount of money from local sources, then state sources, then federal sources. This may be due to the fact that a substantial part of a school's funding comes from local property taxes. Furthermore, on average schools spend more money per student on Personnel than Non-Personnel. This can be explained by the fact that salaries are a significantly larger cost to schools compared to Non-Personnel related expenditures such as materials, books, school building maintenance, etc.

The remainder of our EDA was done in Tableau, and we'll discuss the key findings from this exploration here.

*How does attendance relate to school performance on Math and ELA?*

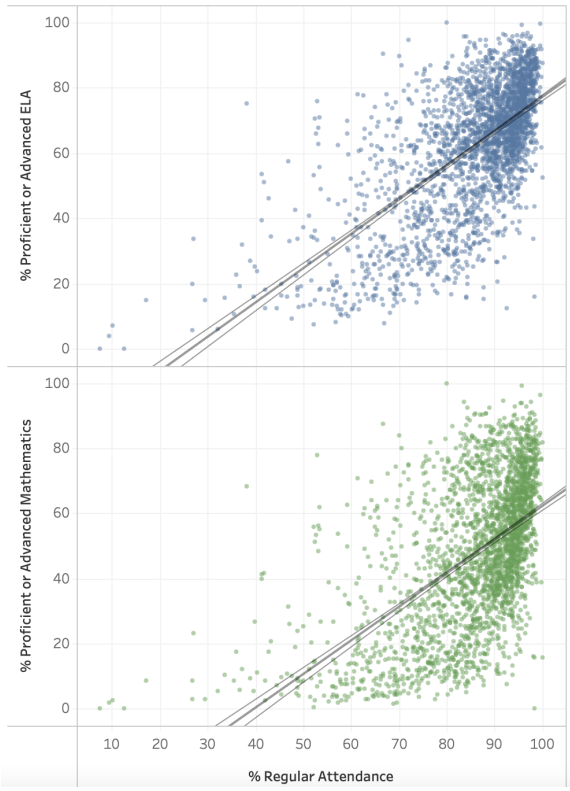


**Figure 5:** Scatterplots of Percent Regular Attendance and School Performance

Based on our scatter plots in **Figure 5** above, we see a strong positive correlation between percent of students with regular attendance in a school and school performance. The $r^2$ score for ELA is 0.41 and for Math is 0.30.

*How do racial/ethnic compositions of schools relate to school performance on Math and ELA?*

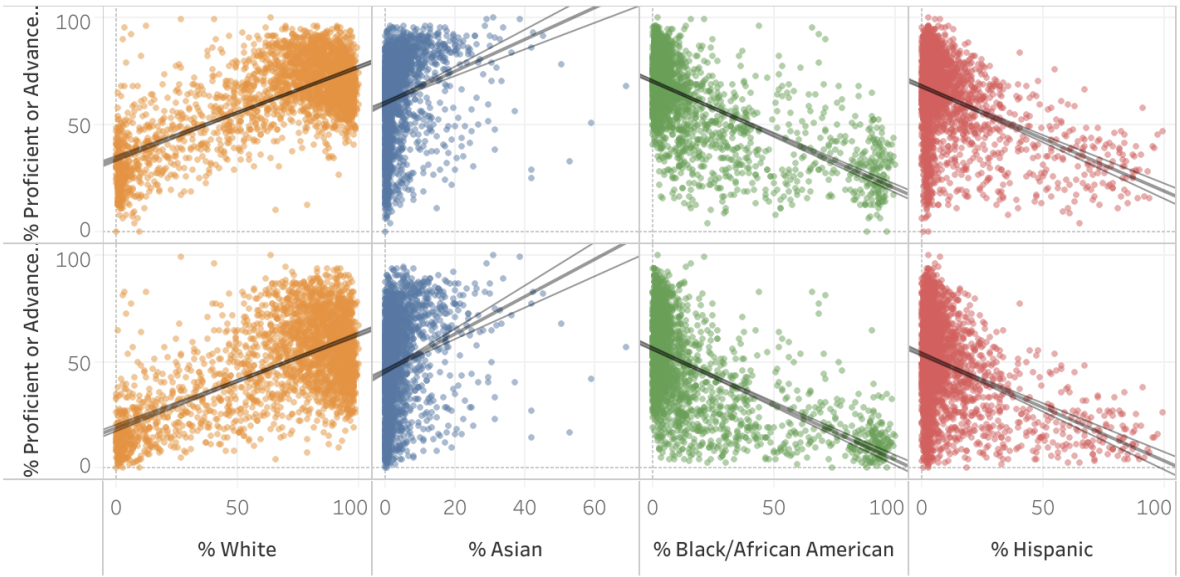


**Figure 6:** Boxplots of Racial/Ethnic Composition and School Performance

As seen in **Figure 6**, schools with a higher percentage of white students tend to perform better. The $r^2$ score for the percentage of white students and ELA performance is 0.51, and for Math, it is 0.41. On the other hand, schools with higher percentages of black and hispanic students tend to perform worse. The $r^2$ scores of the percentage of black students with ELA performance and Math performance are 0.42 and 0.35, respectively. The $r^2$ scores from the percentage of hispanic students are 0.18 and 0.15 with ELA performance and Math performance, respectively.

*How do socioeconomic demographics of schools relate to school performance on Math and ELA?*

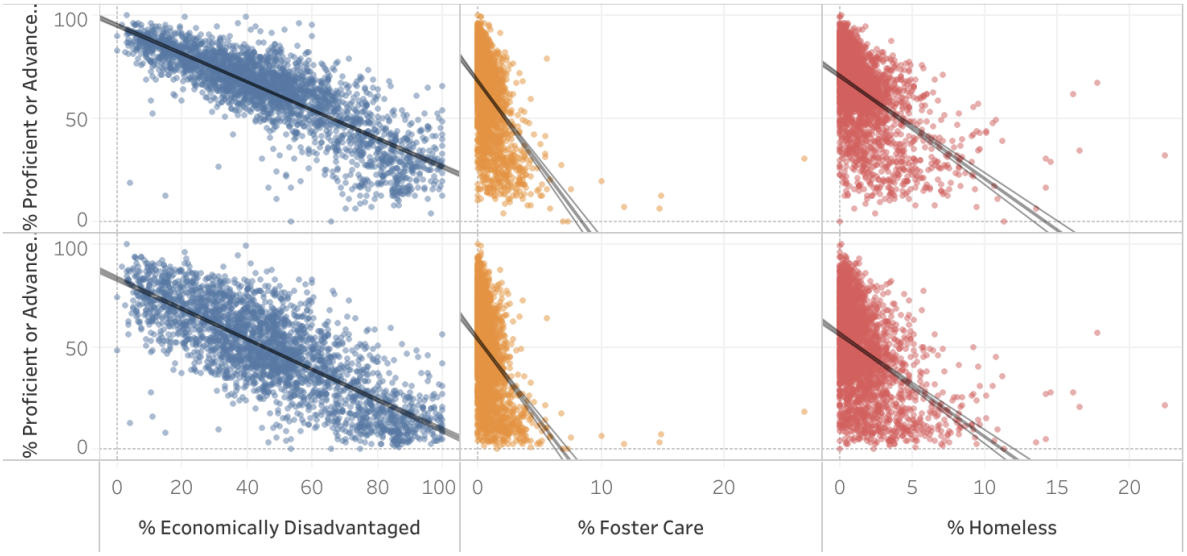


**Figure 7:** Boxplots of Socioeconomic Composition and School Performance

Based on our scatter plots in **Figure 7** above, there is a clear and negative correlation between the percentage of students that are economically disadvantaged in a school and school performance. The $r^2$ scores of the percentage of economically disadvantaged students with ELA performance and Math performance are 0.69 and 0.61, respectively.

It is possible that students who come from families with less resources and financial stability are less likely to prioritize learning. There can be a variety of reasons such as older students needing to also provide childcare for their younger siblings while their parents are working, or that families that are less wealthy may not be able to afford additional resources such as tutoring or test prep courses.

*What is the relationship between expenditures and school performance in Math and ELA?*

We first plotted scatterplots of local expenditures with school performance in ELA and Math, as shown in **Figure 8** below.

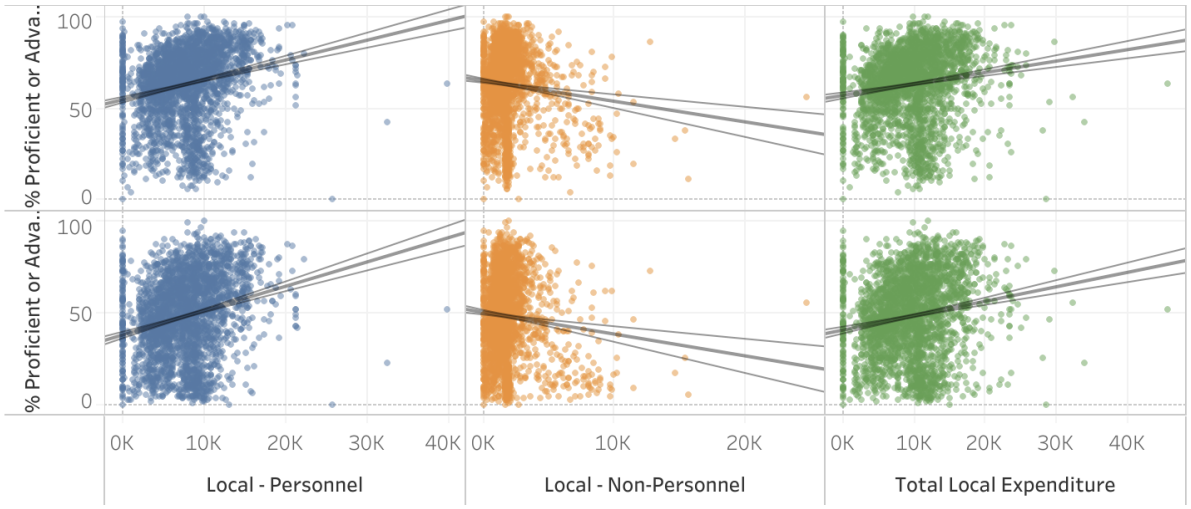


**Figure 8:** Scatterplots of Local Expenditures and School Performance

The trends here do not have as high of a correlation strength as our previous scatterplots. This is rather unexpected because we would have thought that higher performing schools are typically located in wealthier areas with higher property taxes. While we do see a *slight* positive trend on our scatter plots for **Local - Personnel** and **Total Local Expenditure**, the $r^2$ scores hover around 0.05 or below. These will be interesting features to investigate once we can analyze the feature impact from our model.

We also plotted scatterplots of federal expenditures with school performance, as shown in **Figure 9** below.

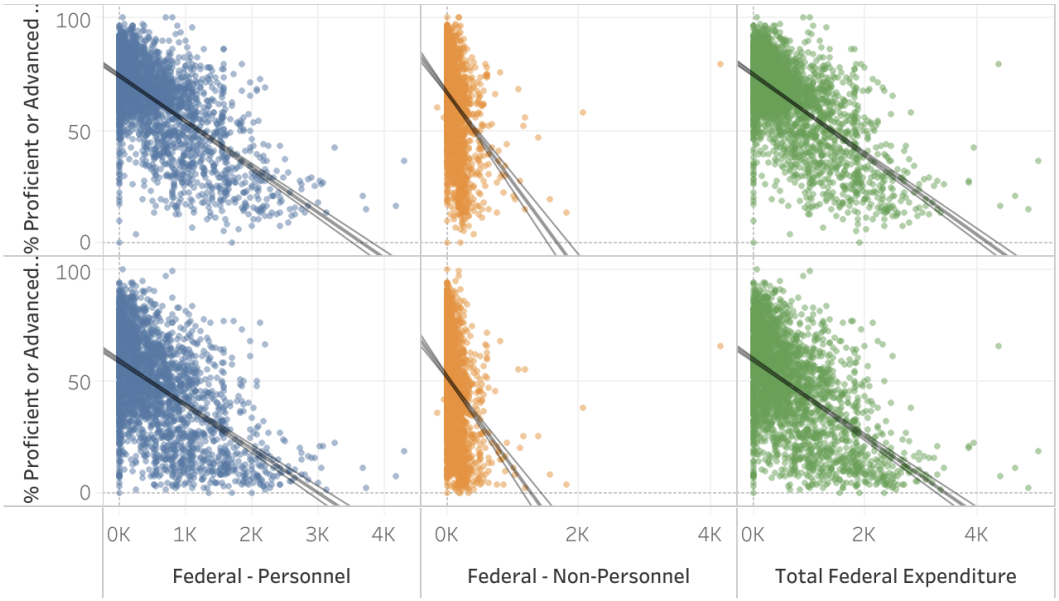


**Figure 9:** Scatterplots of Federal Expenditures and School Performance

The relationship between federal expenditures and school performance does seem to be stronger ($r^2$ scores for Federal - Personnel and Total Federal Expenditure hover around 0.30 and above). Interestingly, we're seeing *negative* correlations between federal expenditures and school

performance.  In other words, schools that spend more money from federal sources tend to perform worse.

Certainly, this is not an indication of causation and we are not suggesting that the federal government should cut back on providing funding to underperforming schools.  Most likely, we are seeing this trend because schools that perform worse are more likely to be Title I schools, and these schools receive money from the federal government. Title I schools are schools with high percentages of students from low income families. As a result, the federal government provides financial assistance to these schools to help ensure that all children meet academic standards.

This brings us to our next set of questions regarding Title I schools.

*How does the performance of Title I schools compare to that of non-Title I schools?*

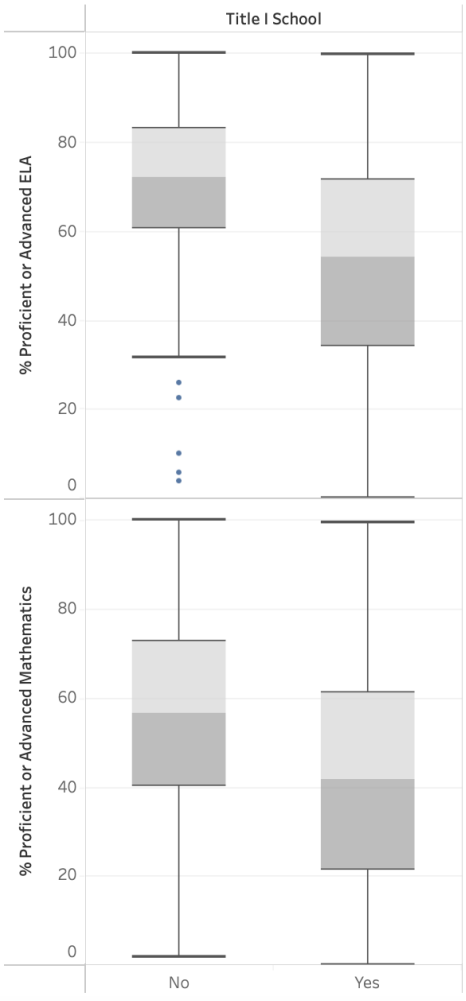


**Figure 10:** Boxplots of School Performance by Title I Status

As we can see from the boxplots in **Figure 10** above, Title I schools and Non-Title I schools both span most of the range from 0-100% on school performance in Math and ELA. However, if we look at the middle 50% of each, we see that Title I schools tend to perform worse than their non-Title I counterparts.

*How do Title I schools' expenditures compare to non-Title I schools?*

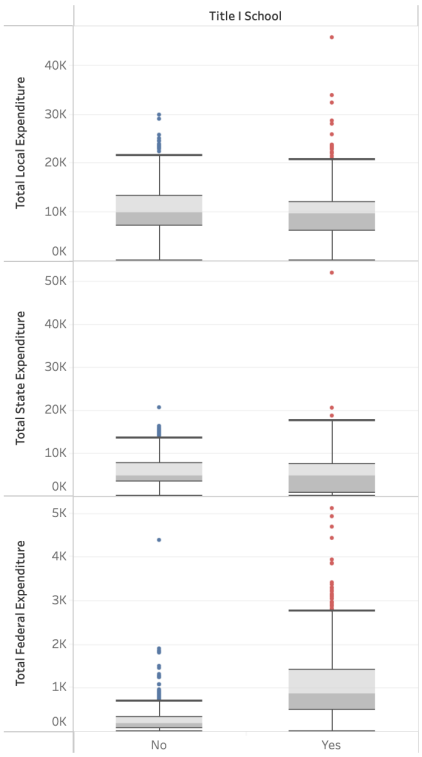


**Figure 11:** Boxplots of Expenditures by Title I Status

The boxplots for expenditures, separated by expenditure source and Title I status, are shown above in **Figure 11**. As suspected, Title I schools spend more money from federal sources per student than non-Title I schools. What's more interesting is that Title I schools and non-Title I schools seem to spend the same range of dollars from local sources. In fact, there are more high end outliers among Title I schools than non-Title I schools.

This concludes our exploratory data analysis. Next, we applied final preprocessing steps in our data and evaluated a baseline model.

## Baseline Modeling

After dropping columns with labels, we converted the Title I status variable (with original string values of 'Yes' and 'No') to a binary variable.

Then, we took a quick look at the heatmap of the absolute value of correlation coefficients to determine if any of our variables were highly correlated with another. We are cautious about this, especially after seeing a few common trends in our exploratory data analysis. There are concerns with collinearity, such as an important predictor becoming less important in our model as that feature may have a collinear relationship with another predictor. The correlation heatmap is shown in **Figure 12** below.

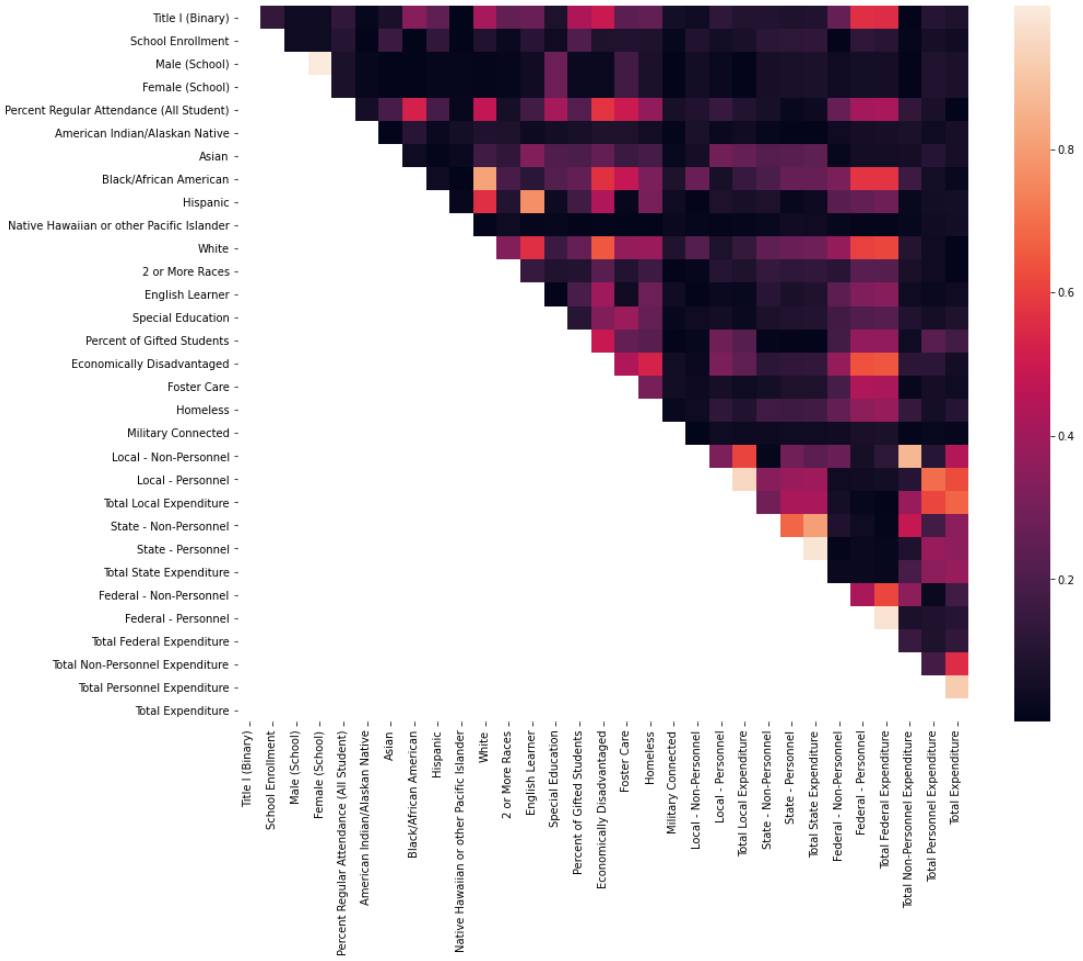


**Figure 12:** Heatmap of Absolute Value of Correlation Coefficients

The correlation heatmap definitely shows that there are a series of features that have high correlation with another. We iterated through the correlation coefficient values and filtered for pairs that have a correlation coefficient whose absolute value is greater than 0.69.

(We originally looked for pairs whose correlation coefficient had an absolute value greater than 0.70, and noticed that all of our engineered expenditure features were in the list with the exception of Total Personnel Expenditure. Upon looking at the list of correlation pairs between 0.5 and 0.7, we saw that Local - Personnel and Total Local Expenditure had the highest correlation coefficient from this group, of 0.694.)

This produced the following pairs and the absolute value of their correlation coefficients:

```
[('Male (School)', 'Female (School)', 1.0),
 ('Black/African American', 'White', 0.8143),
 ('Hispanic', 'English Learner', 0.7677),
 ('Local - Non-Personnel', 'Total Non-Personnel Expenditure', 0.8637),
 ('Local - Personnel', 'Total Local Expenditure', 0.9461),
 ('Local - Personnel', 'Total Personnel Expenditure', 0.694),
 ('State - Non-Personnel', 'Total State Expenditure', 0.8016),
 ('State - Personnel', 'Total State Expenditure', 0.9832),
 ('Federal - Personnel', 'Total Federal Expenditure', 0.9747),
 ('Total Personnel Expenditure', 'Total Expenditure', 0.9164)]
```

From these pairs, we selected the second item of each pair as our high correlation features.

*High correlation features to remove:*

- 'Female (School)'
- 'Total Federal Expenditure'
- 'English Learner'
- 'White'
- 'Total Personnel Expenditure'
- 'Total State Expenditure'
- 'Total Expenditure'
- 'Total Non-Personnel Expenditure'
- 'Total Local Expenditure'

For our baseline models, we will retain all of these features in our design matrix. However, we will use this list in the extended modeling phase of our project to compare models with all of our features included and models with high correlation features removed.

We split our data into train and test sets using an 80/20 train/test split ratio. For our target variables, we split them the same way for both ELA performance and Math performance.

Afterwards, we applied a standard scaler to our data since our features had dramatically different scales. We chose to standardize our data since some of our variables do not have a defined maximum value. To prevent potentially leaking information from our test set into the model, we fit the scaler on just the training data and then standardized both the training and test sets with this scaler.

We used **Linear Regression** (without any regularization) as our baseline model so that we can generate metrics to use as a reference point for extended modeling. As for our metrics, we will use the $r^2$ score, mean absolute error, and the 95% worst case upper and lower bound of residuals.

The metrics for the train and test sets for both ELA and Math models are shown in the table below.

| Subject | Train $r^2$ | Train MAE | Train lower worst case residual | Train upper worst case residual | Test $r^2$ | Test MAE | Test lower worst case residual | Test upper worst case residual |
|---|---|---|---|---|---|---|---|---|
| ELA | 0.80 | 6.68 | -16.55 | 17.18 | 0.771052 | 6.948462 | -18.039068 | 17.495580 |
| Math | 0.70 | 9.68 | -22.83 | 24.12 | 0.655321 | 10.117124 | -24.413786 | 23.483108 |

The performance of our initial model is certainly an acceptable baseline, and given its strong performance already, we are eager to see if we can improve upon this by testing alternative algorithms.

The residual plots and the distribution of residuals are shown below in **Figures 13 and 14**.

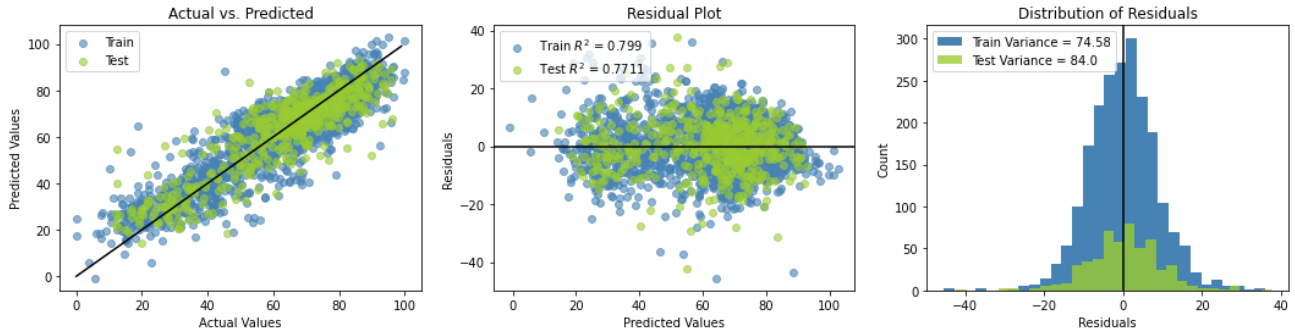


**Figure 13:** Residual plots and distribution of baseline model predicting school performance in ELA

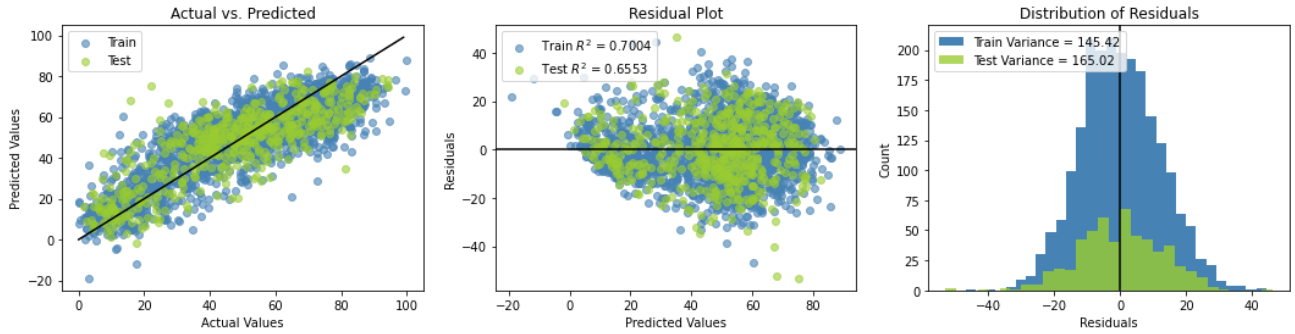


**Figure 13:** Residual plots and distribution of baseline model predicting school performance in Math

Curiously, our math school performance model seems to perform worse relative to our ELA school performance model, with a lower $r^2$ score, higher mean absolute error, and wider interval for worst case residuals. This could be due to a variety of reasons, and beyond the scope of this project.

**Extended Modeling**

With an understanding of where our baseline metrics stand, we will outline our plan for extended modeling. For each of our target variables, we will generate a model that combines one of two design matrices with one of three regression algorithms. Given that our baseline model does not seem to show any signs of overfitting, we did not feel the need to try Ridge Regression or LASSO Regression algorithms and opted for tree-based algorithms instead.

The two design matrices are as follows:

- All features included
- High correlation features removed

The three algorithms we will use are:

- Random Forest
- LightGBM
- XGBoost

This will give us six models to compare for each of our two target variables (ELA and Math). Along the way, we applied hyperparameter tuning for each model using RandomizedSearchCV. The metrics for our models are shown in the table below:

| Subject | Features | Algorithm | Test $r^2$ | Test MAE | Test lower worst case residual | Test upper worst case residual |
|---|---|---|---|---|---|---|
| ELA | All features (baseline) | Linear Regression (baseline) | 0.771052 | 6.948462 | -18.039068 | 17.495580 |
| | All features | Random Forest | 0.792703 | 6.540850 | -16.757434 | 14.956027 |
| | | LightGBM | 0.801945 | 6.444087 | -16.650741 | 15.755405 |
| | | XGBoost | 0.806167 | 6.408681 | -16.448528 | 15.871565 |
| | High corr features removed | Random Forest | 0.798437 | 6.614020 | -16.460128 | 15.613353 |
| | | LightGBM | 0.792678 | 6.652352 | -16.909029 | 15.794854 |
| | | XGBoost | 0.800601 | 6.484040 | -16.613331 | 17.044295 |
| Math | All features (baseline) | Linear Regression (baseline) | 0.655321 | 10.117124 | -24.413786 | 23.483108 |
| | All features | Random Forest | 0.704263 | 9.139749 | -22.957004 | 22.360568 |
| | | LightGBM | 0.710077 | 9.103636 | -23.904595 | 23.129591 |
| | | XGBoost | 0.721477 | 8.959168 | -24.104502 | 22.505217 |
| | High corr features removed | Random Forest | 0.691764 | 9.366147 | -24.674917 | 22.678837 |
| | | LightGBM | 0.701980 | 9.232163 | -23.166101 | 22.786185 |
| | | XGBoost | 0.719577 | 8.911894 | -22.089569 | 23.241234 |

Across all of the new models we trained, we're seeing improved performance over our baseline models for both ELA and Math. In general, our $r^2$ scores are higher, our mean absolute errors are lower, and our interval for worst case lower and upper bound of residuals is smaller.

We created bar graphs of our metrics to compare them more easily and potentially identify if there is a "best" model. The metrics for our ELA models are visualized below in **Figure 14**.

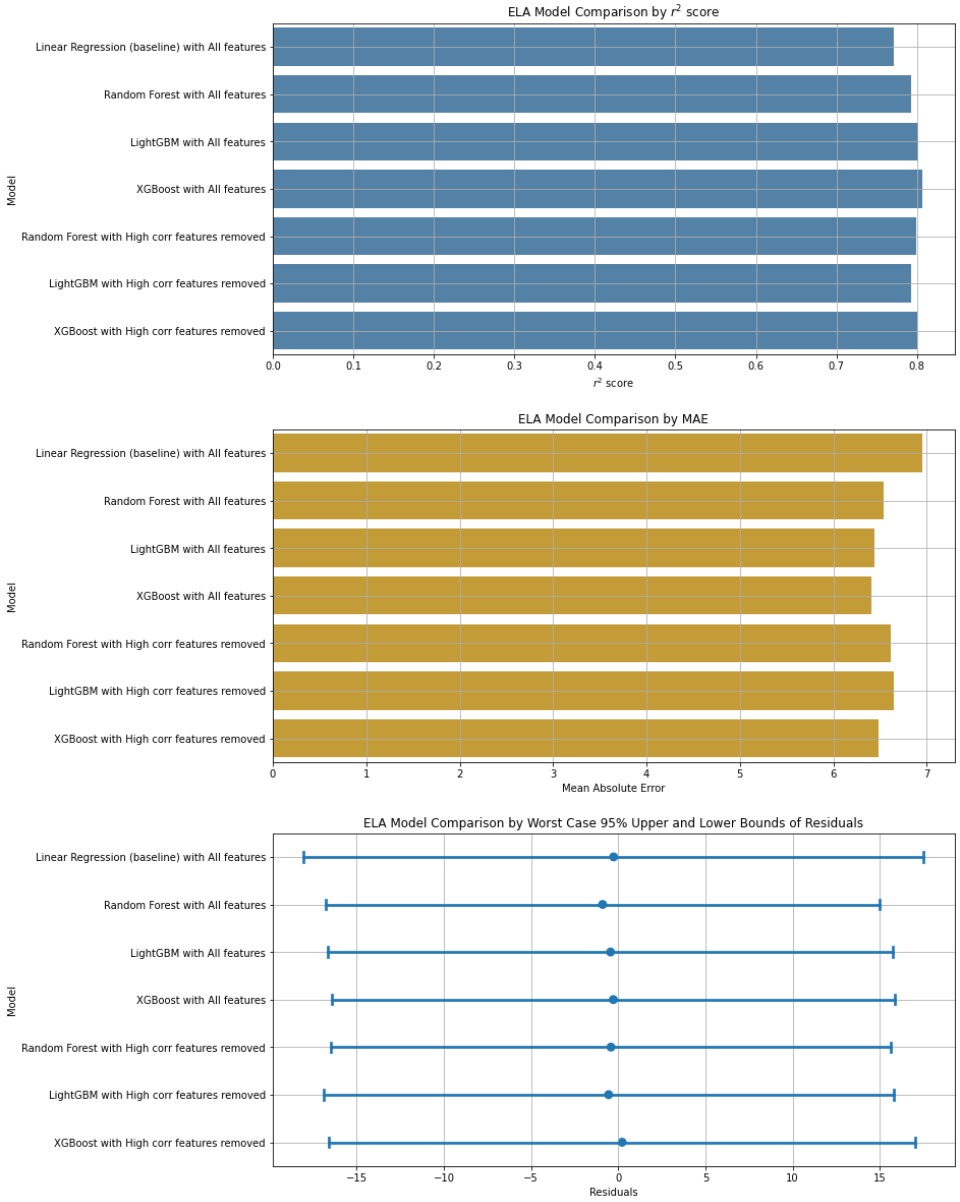


**Figure 14:** Metric comparison for models predicting school performance in ELA

Our models for ELA had very similar metrics, and while there is not any one clear absolute "winner", our XGBoost model with all features included does have the highest $r^2$ score and the lowest mean absolute error. We will select this model for our further analysis of school performance in ELA. The residual plots and distribution of residuals for this model are shown below in **Figure 15**.

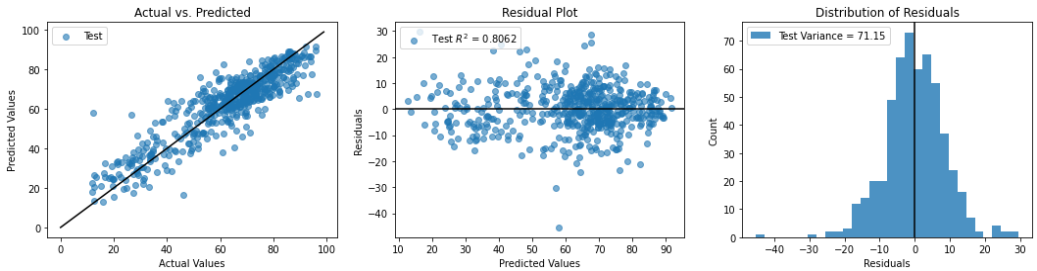


**Figure 15:** Residual plots and distribution of residuals of XGBoost model predicting ELA performance

Next, the metrics for our math models are visualized below in **Figure 16**.

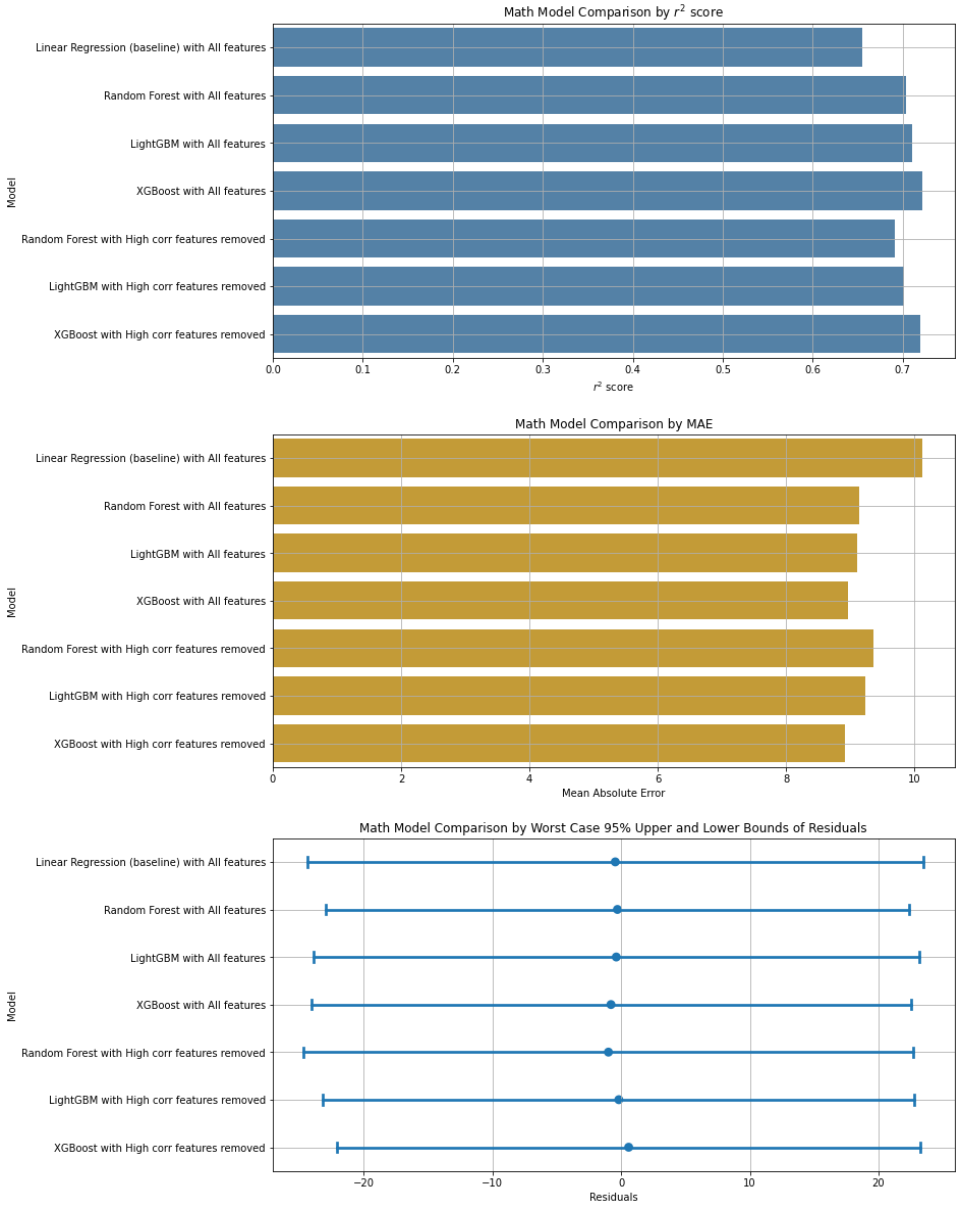


**Figure 16:** Metric comparison for models predicting school performance in Math

Our models for Math had very similar metrics and no clear absolute "winner". With that said, our XGBoost model with high correlation features removed does have a higher $r^2$ score and the lowest mean absolute error. We will select this model for our further analysis of school performance in Math. The residual plots and distribution of residuals for this model are shown below in **Figure 17.**

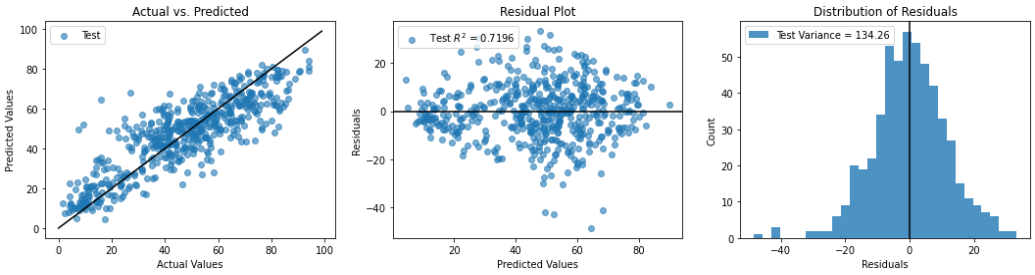


**Figure 17:** Residual plots and distribution of residuals of XGBoost model predicting Math performance

# 3. Findings

For our findings, we used the SHAP library to gain a better understanding of the feature impact of our black box model.  Then, we selected a school whose target values were 0 for both ELA and Math (meaning 0% of students were proficient or above) and tested alternative feature values to produce counterfactual explanations.

The SHAP library will take our model and produce SHAP values for each feature of each row of data. For each row and feature in our dataset, the SHAP value represents the impact that the specific feature has on the overall value of the model's output. To put it another way, the sum of the SHAP values for all the features added with the average prediction for the whole dataset must equal the value of the target.

We can use the shap summary beeswarm plots, combined with color, to understand how high/low values of features can positively or negatively impact our target variable.

We'll first discuss our findings from the feature impact analysis, and then discuss counterfactual explanations.

**Feature Impact - ELA**

The SHAP summary Beeswarm plot for our XGBoost model predicting school performance in ELA is shown below in **Figure 18**.


**Figure 18:** Beeswarm summary plot of SHAP values for model predicting ELA performance

Based on the top features that impact our model, we can interpret their impact on ELA school performance as follows:

| **Positive Impact on ELA** | **Negative Impact on ELA** |
|---|---|
| Lower 'Economically Disadvantaged'<br>Higher 'White'<br>Lower 'Black/African American'<br>Lower 'Special Education'<br>Higher 'Percent Regular Attendance' | Higher 'Economically Disadvantaged'<br>Lower 'White'<br>Higher 'Black/African American'<br>Higher 'Special Education'<br>Lower 'Percent Regular Attendance' |

What is interesting and unexpected is that our expenditure features are not one of our top most important features. 'Local - Personnel' is listed as the 13th most important feature according to the SHAP summary plot. This may be an indication that schools that are able to spend more money per student are not necessarily going to see better performance in ELA.

We also created a few SHAP dependence plots to see how a feature might interact with another feature. When given a feature, the SHAP dependence plots will automatically identify a second feature that interacts with the given feature the most. We've included a couple of these dependency plots below in **Figure 19**.

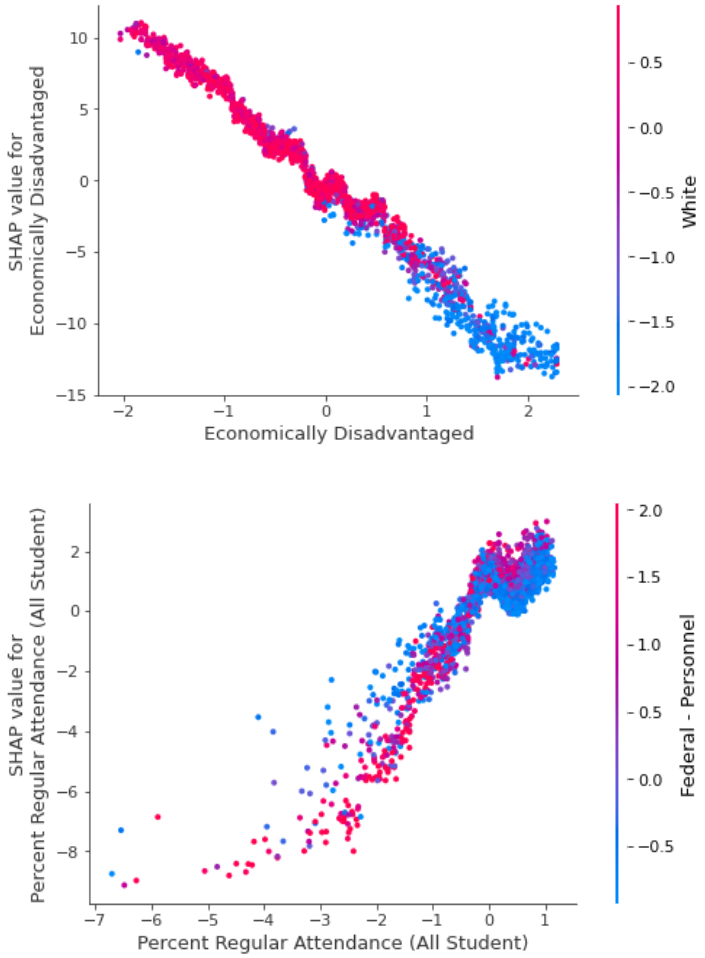


**Figure 19:** SHAP dependence plots for model predicting ELA performance

Our first graph shows the clearest pattern, indicating that low values of 'Economically Disadvantaged' tend to have high values of 'White' as well as have a positive impact on the target. Additionally, high values of 'Economically Disadvantaged tend to have low values of 'White' as well as a negative impact on the target.

We also see that schools with low values of ‘Percent Regular Attendance’ tend to have high values of ‘Federal - Personnel’.  Again, this is likely to be a result of a school’s Title I status.

**Feature Impact - Math**

The SHAP summary beeswarm plot for our XGBoost model predicting school performance in Math is shown below in **Figure 20**.

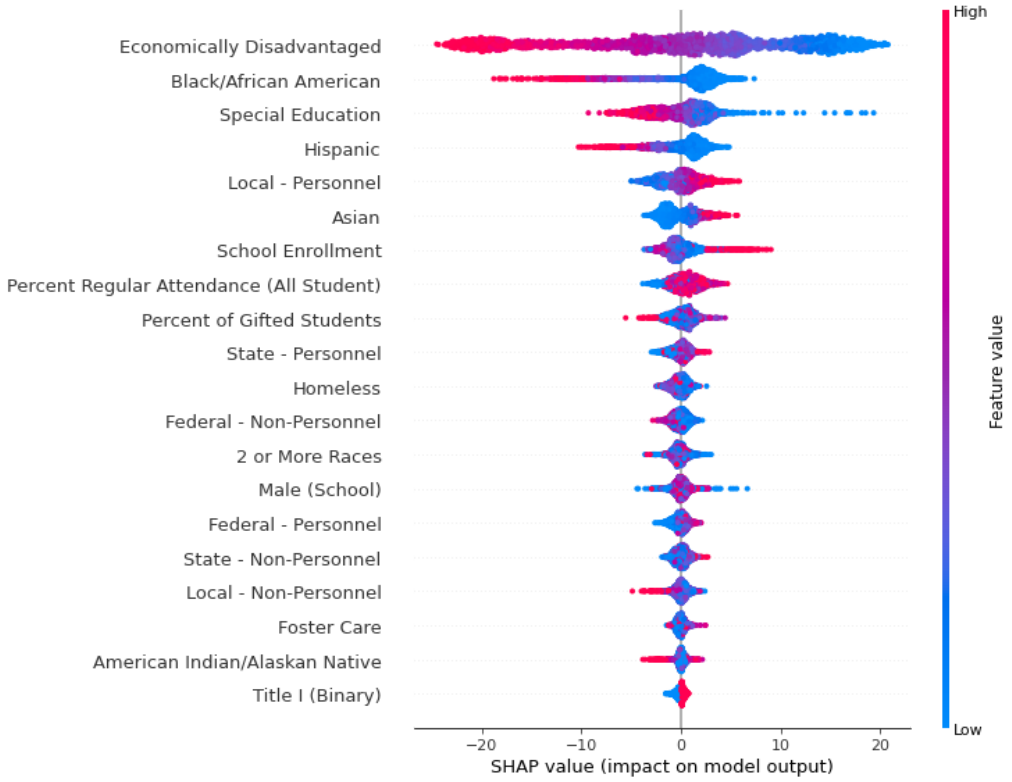


**Figure 20:** Beeswarm summary plot of SHAP values for model predicting Math performance

Based on the top features that impact our model, we can interpret their impact as follows.

| **Positive Impact on Math** | **Negative Impact on Math** |
|---|---|
| Lower 'Economically Disadvantaged'<br>Lower 'Black/African American'<br>Lower 'Special Education'<br>Lower 'Hispanic'<br>Higher 'Local - Personnel'<br>Higher 'Percent Regular Attendance' | Higher 'Economically Disadvantaged'<br>Higher 'Black/African American'<br>Higher 'Special Education'<br>Higher 'Hispanic'<br>Lower 'Local - Personnel'<br>Lower 'Percent Regular Attendance' |

Like before, we created SHAP dependence plots for our Math model as well, a few of which are shown below in **Figure 21**.

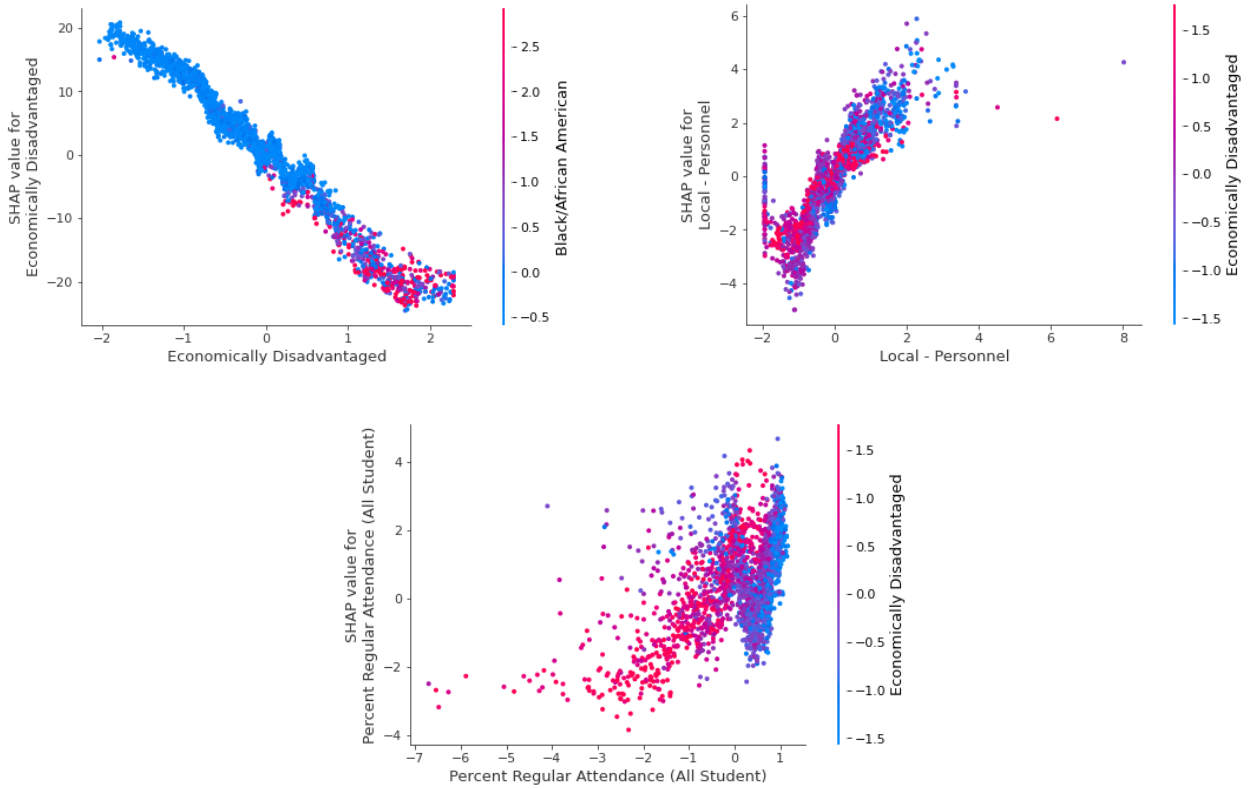


**Figure 21:** SHAP dependence plots for model predicting Math performance

The first dependence plot shows that low values of 'Economically Disadvantaged' tend to have low values of 'Black/African American' and also have a positive impact on the target variable. Higher values of 'Economically Disadvantaged' have more low values of 'Black/African American' and also a negative impact on the target.

Our next plot shows that low values of 'Local - Personnel' tend to have high values of 'Economically Disadvantaged', as well as a negative impact on the target variable. High values of 'Local - Personnel' tend to have lower values of 'Economically Disadvantaged', as well as a positive impact on the target variable.

Our final plot shows that low values of 'Percent Regular Attendance' tend to have high values of 'Economically Disadvantaged' as well as a negative impact on the target. High values of 'Percent Regular Attendance' seem to have a mixture of high/low values of 'Economically Disadvantaged' as well as a mixture of positive/negative impact on the target variable.

### Counterfactual Explanations

In preparation for counterfactual explanations, we need to assess which of these features have practical potential to be altered. It does not make sense to adjust the racial demographic percentages of a school or number of special education, homeless, and foster care students, nor is it *ethical*, to

increase or reduce the number of students from these categories purely to improve school performance.

Assuming our clients for this project are state education law makers and school officials, it may be worth emphasizing the impact of the features 'Economically Disadvantaged' and 'Percent Regular Attendance (All Student)'. Certainly, it is not feasible nor ethical to simply reduce the number of students that come from economically disadvantaged families (or increase the number of students that come from wealthier families). However, an argument can be made for investing in the *local communities* of underperforming schools, providing support services to families whose income fall below the poverty threshold and, as an idea, perhaps even offering microloans to families with small businesses to expand.

With this in mind, we will attempt to produce counterfactuals by altering the following features:

- 'Economically Disadvantaged'
- 'Percent Regular Attendance'
- 'Local - Personnel'

These features will be altered separately to see the effects of each feature independently. We also identified a row in our dataset whose target values for both Math and ELA are 0. Our goal will be to produce counterfactuals that will get us above or as close to a predicted target value of 10 in both ELA and Math. In other words, we would like to see if we can propose alternative scenarios where this school could be predicted to reach 10% school proficiency in ELA and Math. We used the insights gained from our SHAP analysis to propose the alternative values for each of these features. The original and alternative values of the three features of this observation are as follows:

| Feature | Original Value | Alternative Values |
|---|---|---|
| 'Economically Disadvantaged' | 66.09 | Decreasing values from 65 to 50, increment = 1 |
| 'Percent Regular Attendance' | 12.60 | Increasing values from 15 to 80, increment = 5 |
| 'Local - Personnel' | 25885.22 | Increasing values from 26000 to 40000, increment = 1000 |

Surprisingly, the original value for 'Local - Personnel', is already on the high end of the distribution of this feature, as shown below in **Figure 22**. It's interesting that this school is receiving/spending high dollar amounts per student while still performing so poorly.

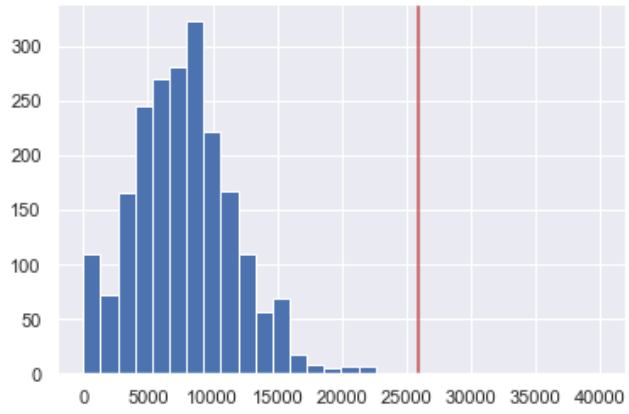


**Figure 22:** Distribution of 'Local - Personnel', red line indicating the value for the selected observation

**Counterfactual Explanations - ELA**

The results of our ELA predictions from alternative feature values are shown in **Figure 23** below.

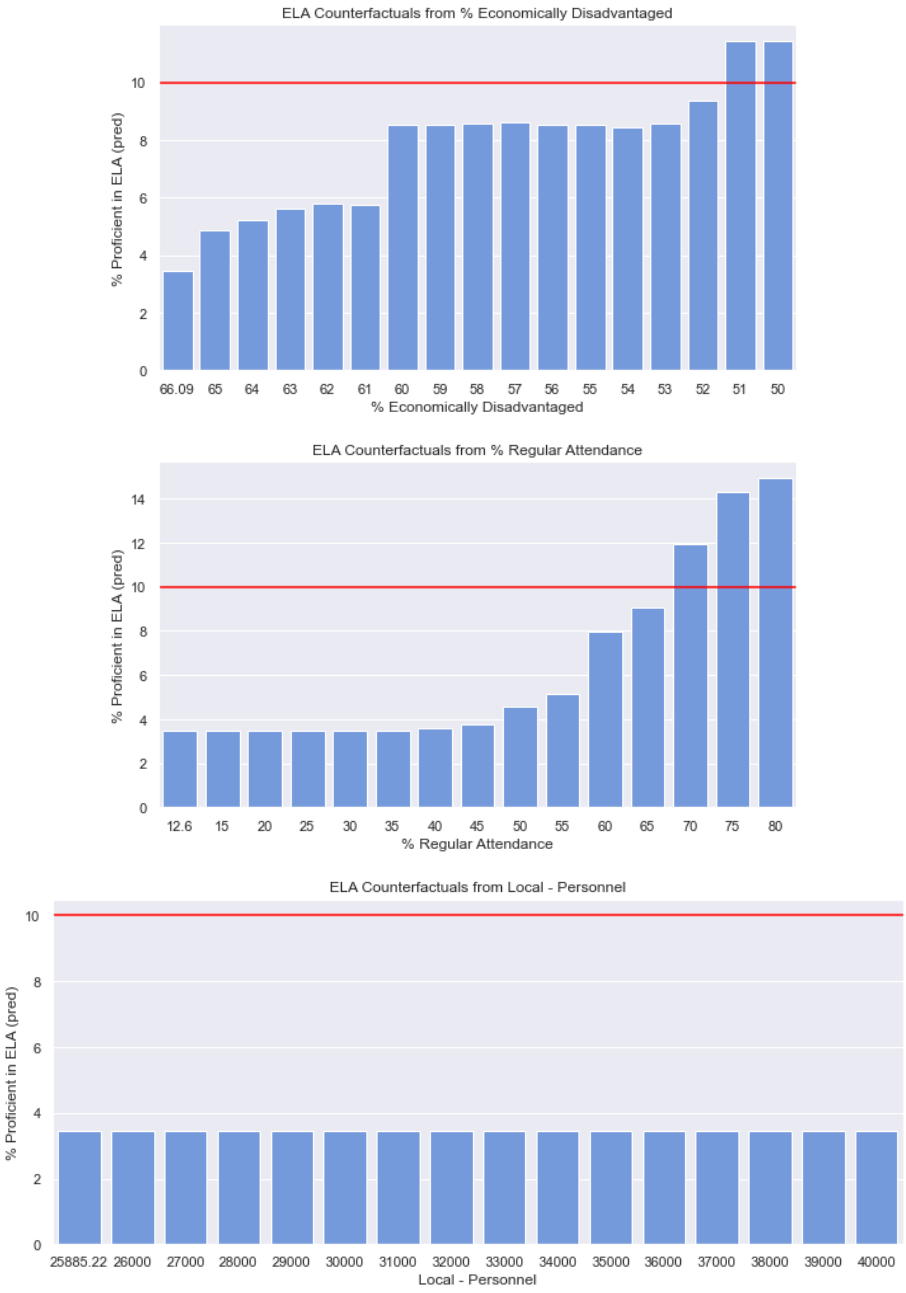


**Figure 23:** Counterfactual results for model predicting ELA performance

Our results show us that if we only reduce the percentage of families that are economically disadvantaged in this school, our model predicts improvements in school performance in ELA. In order to bring the target prediction above our threshold of 10%, it requires a reduction in 'Economically Disadvantaged' by 15, bringing it down to 51%.

As expected, sole increases in Percent Regular Attendance do not have a strong effect on our target variable. In order to start seeing any noticeable differences at all in our target variable, our Percent Regular Attendance needed to jump from 12.6% to 55%.  In order to reach our target

threshold of 10%, the Percent Regular Attendance needed to increase by approximately 60, bringing it to a value of 70%.  It may not be reasonable to expect such a significant increase in this feature in the short-term.

And finally, sole increases in 'Local - Personnel' do not cause our model's predictions for the target value to improve at all.

### Counterfactual Explanations - Math

The results of our Math predictions from alternative feature values are shown in **Figure 24** below.

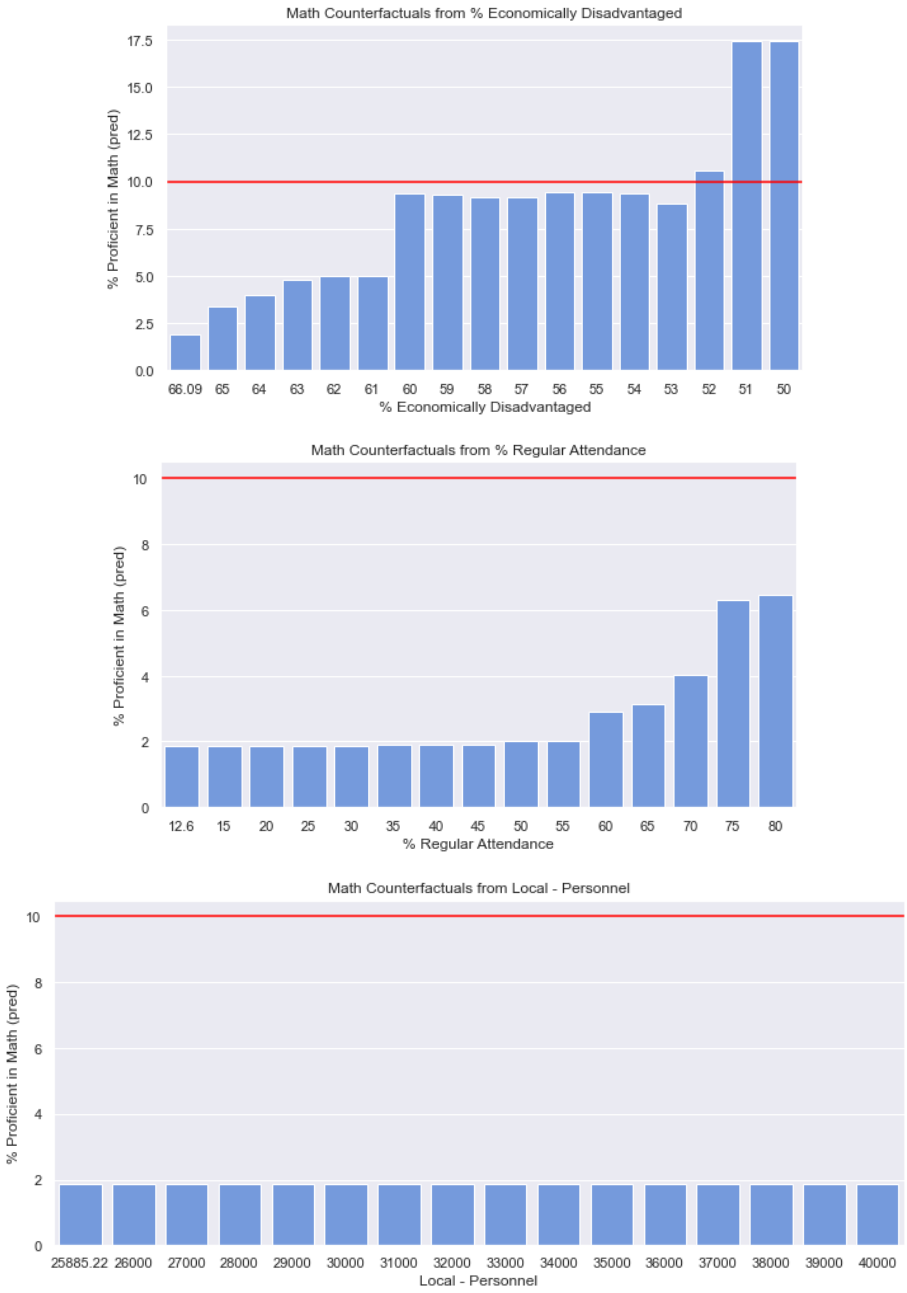


**Figure 24:** Counterfactual results for model predicting Math performance

With our Math model, we see similar counterfactual results as our ELA model. When we only decrease 'Economically Disadvantaged', the predicted school performance in Math does improve. To reach our target threshold of 10% for school math performance, the model requires that 'Economically Disadvantaged' decrease down to 52%.

When we only increase ‘Percent Regular Attendance’, the effect on school Math performance is not as significant. In fact, increasing it from the original 12.6% to 80% only brings the predicted target value up to 6.4%.

As we've seen with our ELA counterfactuals for 'Local - Personnel', increases in this feature do not lead to any improvements whatsoever in predicted school performance in Math.

**Counterfactuals with Bayesian Optimization**

While the above analysis was centered around altering only one feature value at a time, we were also curious whether there exist *combinations* of *smaller* feature changes that would also bring our predicted school performance (in both ELA and Math) close to 10%.

| Feature | Original Value | Alternative Values for Combined Analysis |
|---|---|---|
| ‘Economically Disadvantaged’ | 66.09 | Decreasing values from 65 to 55, increment = 1 |
| ‘Percent Regular Attendance’ | 12.60 | Increasing values from 15 to 30, increment = 5 |
| ‘Local - Personnel’ | 25885.22 | Increasing values from 26000 to 30000, increment = 1000 |

We created an optimization function to be maximized, defined as the predicted output subtracted by our target threshold of 10. To perform this analysis efficiently, we then performed Bayesian Optimization on this function using the search space indicated in the table above. We ran 5 initial points and 20 iterations.

For our ELA model, the best output of our function and the corresponding alternative feature values are shown in the table below:

| Bayesian Optimization, Result for ELA | |
|---|---|
| ‘Economically Disadvantaged’ | 56.58 |
| ‘Percent Regular Attendance’ | 21.23 |
| ‘Local - Personnel’ | 28041.70 |
| **Function output** | –1.406 |

For our Math model, the best output of our function and the corresponding alternative feature values are shown in the table below:

| Bayesian Optimization, Result for Math | |
|---|---|
| ‘Economically Disadvantaged’ | 56.11 |
| ‘Percent Regular Attendance’ | 13.68 |
| ‘Local - Personnel’ | 27372.17 |
| **Function output** | –0.57 |

Our analysis shows that combinations of smaller changes in these three features can get us close to our target threshold of 10% proficiency in both ELA and Math. For our ELA model, we were able to get up to 1.4 percentage points below 10%, and for our Math model, we were able to get up to 0.57 percentage points below 10%.

## 4. Conclusions and Future Work

We can summarize the key results of this project as follows:

| Target | English Language Arts | Mathematics |
|---|---|---|
| Model | XGBoost, all engineered features included | XGBoost, high correlation features removed |
| Metrics | $r^2$ score = 0.81<br>Mean absolute error = 6.41<br>95% Worst case residuals: [−16.45, 15.87] | $r^2$ score = 0.72<br>Mean absolute error = 8.91<br>95% Worst case residuals: [−22.09, 23.24] |
| Feature Impact | Positive impact on ELA performance:<br>• Lower 'Economically Disadvantaged'<br>• Higher 'White'<br>• Lower 'Black/African American'<br>• Lower 'Special Education'<br>• Higher 'Percent Regular Attendance’ | Positive impact on Math performance:<br>• Lower 'Economically Disadvantaged'<br>• Lower 'Black/African American'<br>• Lower 'Special Education'<br>• Lower 'Hispanic'<br>• Higher 'Local - Personnel'<br>• Higher 'Percent Regular Attendance' |
| Counterfactuals | Alternative scenarios to improve selected school’s predicted ELA performance from 0% to 10% (or as close to 10% as possible):<br>• ↓ ‘Economically Disadvantaged’ to 51<br>*OR*<br>• ↑ ‘Percent Regular Attendance’ to 70<br>*OR*<br>• Combination of:<br>↓ ‘Economically Disadvantaged’ to 56.58<br>↑ ‘Percent Regular Attendance’ to 21.23<br>↑ ‘Local - Personnel’ to 28041.70 | Alternative scenarios to improve selected school’s predicted Math performance from 0% to 10% (or as close to 10% as possible):<br>• ↓ ‘Economically Disadvantaged’ to 52<br>*OR*<br>• ↑ ‘Percent Regular Attendance’ to 80+<br>*OR*<br>• Combination of:<br>↓ ‘Economically Disadvantaged’ to 56.11<br>↑ ‘Percent Regular Attendance’ to 13.68<br>↑ ‘Local - Personnel’ to 27372.17 |

As for future work, there’s certainly room to potentially improve the performance of our models and add to the interpretability of our results. Here are some options for future work:

- During our exploratory data analysis, a few of our variables had some extreme outliers. In this project, we checked the validity of those values from a separate source and decided to keep them. It's possible that removing these outliers may help our models to better model the trend.
- As an additional metric, we can use the adjusted $r^2$ score to determine if the removal or inclusion of our high correlation features will result in better or worse performance.
- When finding counterfactuals using Bayesian Optimization, we can add more features and their alternative values to the solution space.
- It may also be interesting to consider looking at two datasets on PA school data from different school years, such as the 2016-2017 school year and then the 2018-2019 school year. Perhaps a dataset can be constructed with features representing the change in feature values of the two datasets and the target variable representing the change in percent proficiency in ELA and/or math. It would be interesting to create a model that will demonstrate which changes result in positive changes in school performance, as well as negative changes in school performance.
- Incorporate machine learning algorithms that can establish causal inferences between our feature variables and our target variables.

## 5. Recommendations for the Clients

As we consider recommendations for how to use these findings, it is important to keep in mind that *"correlation does not imply causation"*.[3] In other words, while we have certainly established links between certain feature values and school performance on the PSSAs, this does not necessarily mean that these inputs are the *cause* of improved or worsened performance of a school. Nonetheless, we can argue for some of the applications in taking what the correlations are indicating as a *hypothesis* for causation.

Furthermore, in creating recommendations from our results, it is imperative that we first understand the practicality and ethics of our results. As we mentioned previously, it is not feasible nor ethical to simply reduce (or increase) the number of students that come from one demographic group for the sake of improving school performance.

Unquestionably, the most important feature in our models and the one with the clearest trend is 'Economically Disadvantaged'. An argument can be made towards increasing community engagement and support services for areas with high poverty rates as a means of improving local school performance.

We also want to highlight the impact of higher values of 'Percent Regular Attendance'. While the improvement on school performance is not as significant as reducing 'Economically Disadvantaged', we can also make an argument for encouraging families to ensure that their children attend school as regularly as possible.

Furthermore, it's possible that by investing in support services for areas with high poverty rates, the reduction of the percentage of economically disadvantaged will naturally improve attendance. As families gain more stability, it may become easier for their children to consistently attend school. In

[3] https://www.theguardian.com/science/blog/2012/jan/06/correlation-causation

fact, we saw a similar pattern emerge in our dependence plots for Math where lower values of 'Economically Disadvantaged' were correlated with higher values of 'Percent Regular Attendance''.

As for our results from looking at alternative values for 'Local - Personnel', the lack of change in target value may point to the notion that simply pumping more money into underperforming schools does not necessarily improve performance. This, of course, is ***not*** an implication in any way that state, local, and federal sources should decrease school funding. Rather, we can argue that the problems of underperforming schools may not necessarily be a result of issues within the school but rather a result of issues that face the local communities of these schools. Therefore, the most effective solution involves connecting with the local communities of schools and improving the quality of life for struggling families as a means of reducing the percentage of families who are economically disadvantaged.

As a result, our recommendations to improve the performance of underperforming schools are:

- Research and invest in ideas to improve the quality of life for impoverished neighborhoods, such as:
    - Offering social services
    - Providing funding for adults to attend trade school or other educational opportunities
    - Providing micro loans to families with small businesses to expand their ventures
- Connect with the families of local communities of underperforming schools to emphasize the significance of consistent attendance. Identify any common barriers that families may face in preventing their children from attending school regularly.

## 6. Consulted Resources


https://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/default.aspx

https://www.ncsl.org/research/education/funding-approaches-the-property-tax-and-public-ed.aspx

https://www2.ed.gov/programs/titleiparta/index.html

https://blog.clairvoyantsoft.com/correlation-and-collinearity-how-they-can-make-or-break-a-model-9135fbe6936a

https://www.ncnewsonline.com/news/local_news/budget-report-on-pa-education-agency-provokes-spat-between-lawmakers-watchdogs/article_f332fd80-533c-5619-858d-fcecb3a8afb5.html