Stacked Regression and Classification Models for NYC Elementary School Analysis

1. Introduction & Problem Statement

Our analysis focuses on data from New York City Public Elementary Schools for the academic years 2018-2019 and 2019-2020. Specifically, we examine the relationships among school size, racial composition, and poverty levels, derived from demographic surveys, and their impact on academic performance, measured by mathematics test scores.

The primary research question guiding our study is whether an ensemble modeling approach, in which we stack diverse base learners, can achieve superior performance compared to traditional single-model baselines. We aim to evaluate the model's effectiveness across two critical tasks. First, we predict schools' performance scores using linear and non-linear regression models. Next, we classify schools into distinct performance tiers. By examining these modeling techniques, we aim to provide analytical tools and insights for educational policymakers and school administrators, supporting informed decision-making and educational outcomes.

2. Data Collection & Preparation

Description of Datasets

The datasets used are all extracted from the **NYC Open Data** platform and the **official website of the NYC DoE.**

They are as follows:

- 2017-2018 2021-2022 Demographic Snapshot (5-6 datasets combined for three years
- Math Test Results 2013-2023
- NYC School Survey for the academic years 2017-2018, 2018-2019, 2019-2020.

The Demographic Snapshot includes school level data about the number of students enrolled in the school by grade (from 3k up to 12th grade), as well as proportions of different ethnic groups, gender, disability, and economic groups among students for each academic year.

The Math Test Results dataset includes school-level results for the New York State Math exams for each academic year across 4 levels of performance for the exam.

The New York City Survey data includes parents, students, and teacher's responses at the school level to the survey questions, in addition to the 6 key elements used by the department to score schools on a scale of 1.00 to 4.99:

- Rigorous Instruction;
- Collaborative Teachers;
- Supportive Environment;
- Effective School Leadership;
- Strong Family-Community Ties, and
- Trust.

Data Consistency as a Challenge

The analysis is limited to the specified academic years due to the unavailability of consistent survey data and scoring methodology for a broader time range.

On one hand, the newly developed "Framework for Great Schools" had been adopted as recently as 2016, making the first available rating for the 2017-2018 school year. On the other hand, the COVID-19 pandemic has led to a pause in these ratings from the 2020-2021 school year because of the inadequacy of the "elements" under the conditions of teaching brought about by the pandemic (e.g. measuring absenteeism).

We then discount the 2017-2018 school due to the difference in scoring for the New York State Math exams compared to 2018-2019 and 2019-2020 academic school years

Data Cleaning Process

Prior to merging the datasets for analysis, significant cleaning and preparation steps were required to ensure consistency across variables and years.

First, the **Demographic Snapshot** dataset, which compiles data across multiple academic years, was standardized. Variable names were harmonized across the different files to allow for proper alignment. Certain columns required manual adjustment, notably the fields reporting poverty and economic need. In cases where poverty rates were indicated as "Above 95%," values were replaced with a numerical estimate of 97.5% to enable quantitative analysis. For these schools, the absolute number of students classified as living in poverty was recalculated accordingly, assuming a poverty proportion of 97.5% applied to the total enrollment. All percentage fields were then converted into numeric format by extracting the leading numeric values, and demographic indicators were uniformly cast as numeric variables.

Next, the **Math Test Results** dataset, which contained school-level performance results on the New York State mathematics exams, was assembled by joining multiple sheets within the master file covering different years. Each sheet was merged into a unified dataframe. Test score variables were cleaned by converting all score fields to numeric and rounding values to two decimal places to ensure consistency in measurement. These math scores serve as the proxy for academic achievement in the subsequent analysis.

For the **NYC School Survey** data, school-level survey results from 2017, 2018, and 2019 were individually imported and processed. Non-data rows (such as headers or blank rows) were removed, and the key survey response fields — particularly those associated with the six Framework for Great Schools elements — were converted into numeric variables and rounded. A "Year" field was added manually to each survey subset to allow them to be merged across years.

After cleaning each dataset individually, the datasets were merged sequentially using the **DBN** (District Borough Number) and **Year** fields as the joining keys. To maintain data quality, columns with more than 60% missing values were excluded from the final dataset. Additionally, specific columns deemed redundant or problematic were removed based on manual inspection. All entries lacking mathematics test performance data were dropped, as the math test scores constitute the dependent variable of the study.

Finally, an additional geographic variable was created to indicate the borough associated with each school. The borough was inferred from the school DBN code — for example, DBNs containing "M" were classified as Manhattan, "Q" as Queens, "K" as Brooklyn, and "X" as Bronx. Schools that did not fit this categorization were temporarily assigned a placeholder label.

The final cleaned dataset includes data for 448 New York Public Schools in the academic years 2017-2018, 2018-2019 and 2019-2020, across 67 variables.

3. Exploratory Data Analysis (EDA)

Given the richness of our dataset, it becomes essential to explore the structure of our variables.

Correlation Matrix

Our initial diagnostic focused on identifying potential multicollinearity among variables by computing pairwise Pearson correlations. As expected, several variables exhibited strong, and in some cases near-perfect, correlations (Figure 1).

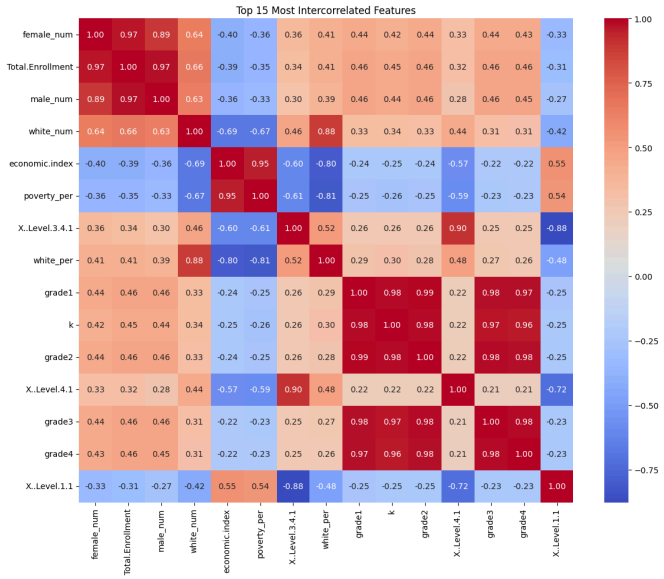

Figure 1: Correlation Matrix of 15 Most Correlated Variables

First, enrollment counts across early grades (Kindergarten through Grade 4) were almost perfectly correlated. This reflects the standardized structure of elementary schools in NYC, where enrollment typically remains consistent across lower grades.

Additionally, Total Enrollment showed near-perfect correlation (rho > 0.98) with the number of female and male students combined — a definitional relationship:

$$\text{Total.Enrollment} = \text{female\_num} + \text{male\_num}.$$

We also observed a strong correlation (rho = 0.95) between the percentage of students living in poverty and the Economic Need Index, which is expected given that both metrics quantify overlapping aspects of economic disadvantage.

These redundancies suggest that dimensionality reduction or careful feature selection will be necessary to mitigate multicollinearity risks, which could otherwise destabilize model estimates or inflate variance.

Demographic Variables

Drawing from literature on the effects of socioeconomic disparities on academic performance, we expect the poverty rate in schools to be a strong leading indicator for low performance on maths regents exams.

However, there is a possibility for the true effect of economic hardship on the student population to be hard to detect due to the reality of the economic profiles of students in NYC public schools. According to our data, most schools appear to have over 60% poverty rate (Figure 2).

This limited variability could diminish the observed marginal effect of poverty on performance, potentially masking more granular socio-economic impacts.

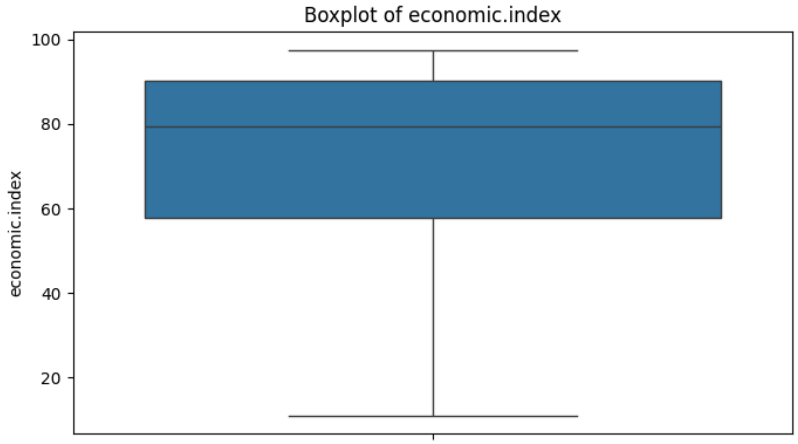

Figure 2: Economic Index Boxplot

We also anticipated that the concentration of poverty rates across NYC public schools would coincide with limited diversity in racial and ethnic distributions. To avoid overstating variations among minority groups, we instead focused on visualizing the percentage of white students as a representative demographic indicator (Figure 3).

The graph indicates that 75% of NYC public schools have a minority student body, specifically less than 25% white students. This is consistent with what we would expect in terms of correlation between high poverty levels and high black and hispanic student populations.

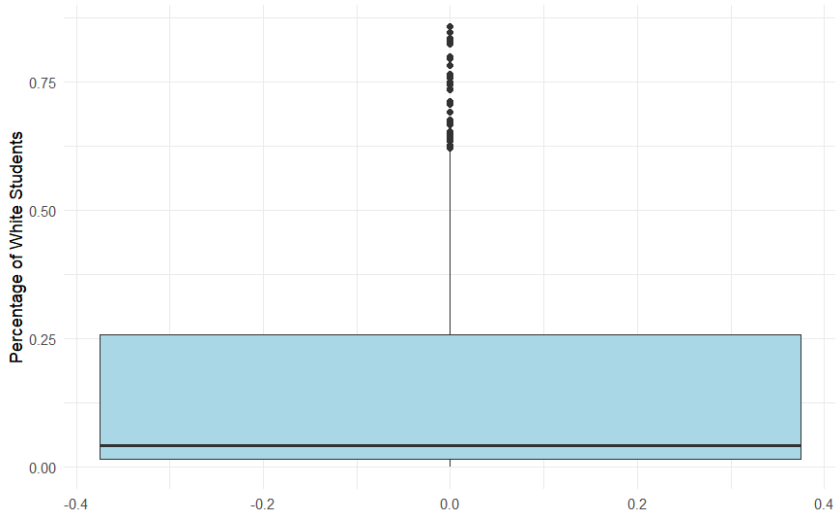

Figure 3: Percentage of White Students Boxplot

Given the high poverty rates, we also expect general medium to low performance for the majority of NYC public schools in our dataset. This is what we observe in Figure 4, where 75% of schools have less than 25% of exam takers scoring within the 4th level of the exam.

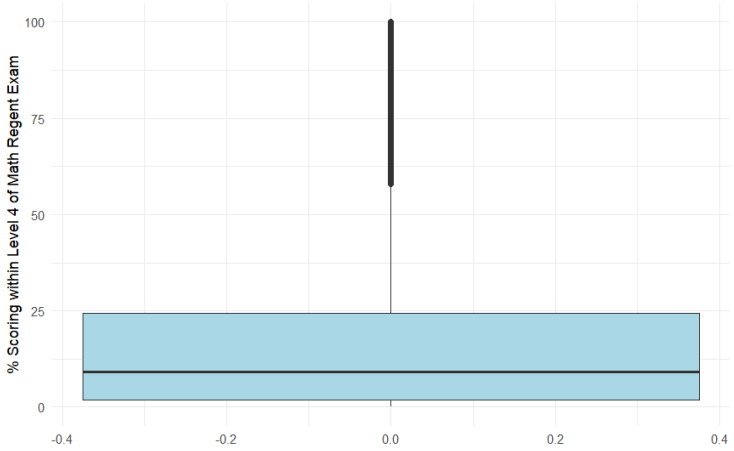

Figure 4: Boxplot of Percentage of Students Scoring within Level 4 of NYC Math Regent Exam

School Environment Variables (Survey Scores)

In terms of school environment variables (scores used are explained in the Appendix), we opted out of the segregation of scores between parents, students and teachers and instead rely on overall scores reflectives of the environment of the school.

Generally, it appears that all categories from the Framework for Great Schools show a similar distribution of scores, skewed to the right (Figure 5).

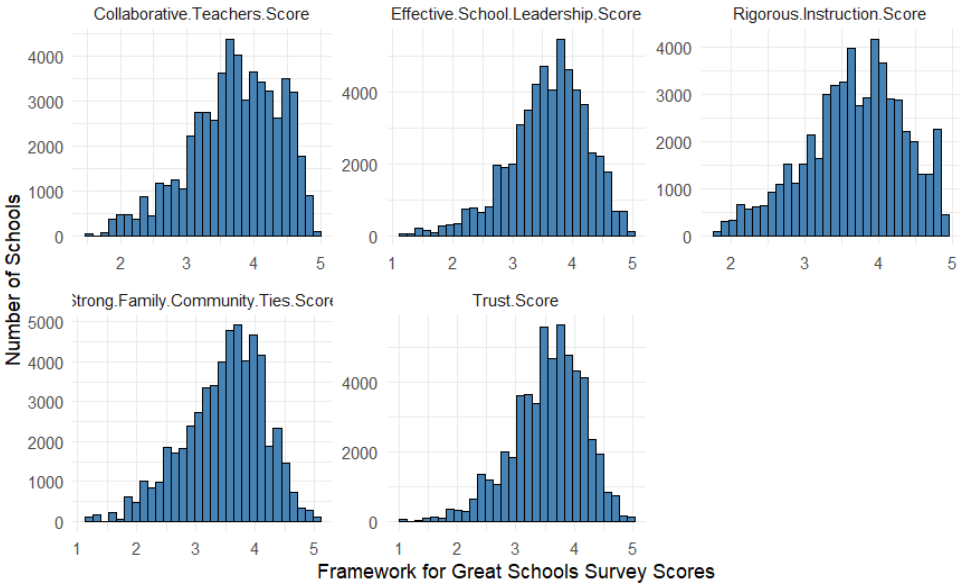

Figure 5: Histograms of Framework for Great Schools Survey Category Scores

Patterns using K-Means and PCA

To further explore structure and reduce dimensionality, we conducted a PCA followed by K-Means clustering, revealing three distinct school groupings differentiated by size, poverty, and performance levels.

The PCA1 component most likely measures a performance-poverty gradient, in that schools on the positive side are doing better academically and have less economically disadvantaged student enrollment. And PCA2 could be a combination of measures like student body size, trust measures, or teacher collaboration measures.

The K-Means clustering shown on Figures 7, has three distinct cluster regions that almost do not overlap with each other implies that the dataset has strong internal structure, based on distinct combinations of school characteristics.

We notice that cluster 1 (Orange) is set apart from the rest along the PCA1 axis and most likely depicts large enrollment, high-achieving schools with low economic need. And schools in cluster 2 (Blue) are mostly small schools with greater poverty rates and lower academic achievement. These schools are likely to have greater structural and socioeconomic problems. Finally, Cluster 0 (Green) is between the two extremes and is that group of schools with more varied profiles—mixed size, economic need, and performance.

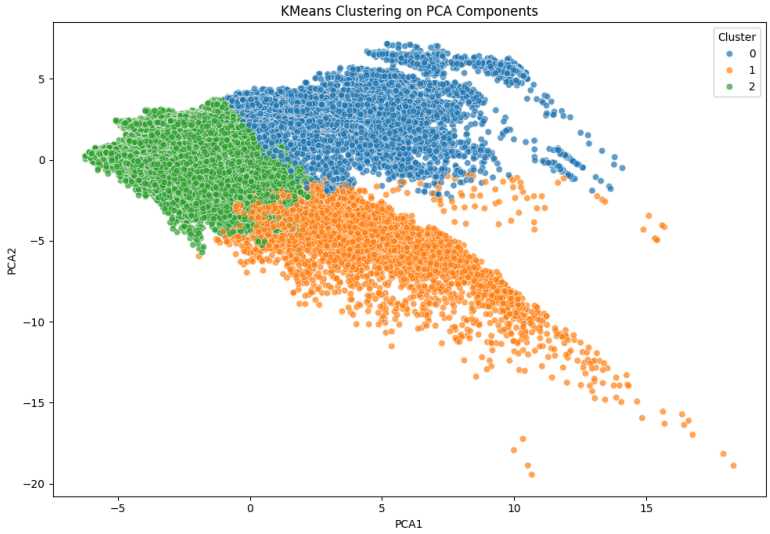

Figure 6: K-Means Clustering on PCA Components (k = 3)

Implications for Model Selection:

Structured clusters → Use classification models
Since clear groupings emerge after dimensionality reduction, models like: Logistic Regression; Random Forest; Gradient Boosting

If PCA already separates groups well, a **linear model** (like logistic regression) may perform decently.

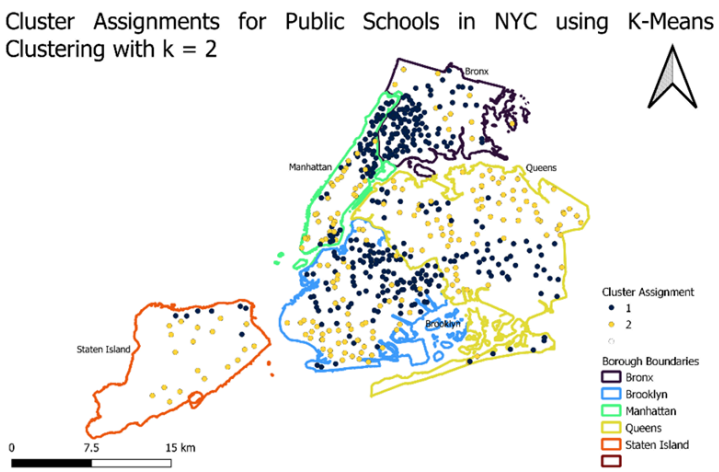

Figure 7: Geographical Representation of K-Means Clusters on NYC 5 Boroughs Map

To further drive home the patterns of public school profiles across boroughs, namely as it relates to socio-economic and performance profiles, we used QGIS software to represent the clusters on a map, limiting our clusters this time to only 2 for cleaner geographical representations (Figure 8).

We can clearly see that the blue dots are concentrated in specific areas (namely The Bronx and the area between Brooklyn and Queens), and the yellow dots are concentrated in areas away, namely Staten Island, south of Manhattan and north of Queens).

Each dot represents a school with the yellow dots having a relatively affluent profile and better performance score, unlike schools represented by the blue dots.

4. Feature Engineering & Preprocessing

Feature Variables:

For the feature variables, a number of feature engineering operations were performed to improve data quality, dimensionality reduction, and model explainability. Grade-level enrollment variables were reduced to three aggregated categories—elementary_enrollment, middle_enrollment, and high_enrollment—reflecting school level differences without redundancy. A number of ratio-based features were also formed to control for school size such as male_female_ratio, swd_ratio (students with a disability), ell_ratio (English language learners), and poverty_ratio (students in economically disadvantaged homes).

For reduction in skewness and smoother model fit, key variables in terms of school size such as Total.Enrollment and Number.Tested were subjected to a log transformation. Categorical variables like borough were one-hot encoded and non-informative and unnecessary fields were dropped to prepare the dataset for supervised learning purposes.

Response Variables:

For the Response variable, we aggregate all the X.level. columns to a response variable to capture academic performance for each school in each year. The calculation formula below:

$$\text{Performance Score} = \sum_{i=1}^{4} i \times \% \text{ of students who scored in Level } i \text{ range}$$

.

The performance score is a weighted average of student proportions across four proficiency levels in their exam. The performance score would be a continuous number from 1 to 4, which is suitable for the regression model in the later sections. Then performance_score_boxcox is also created by normalizing the raw performance score through a Box-Cox transformation to correct non-normality (Figure 8).

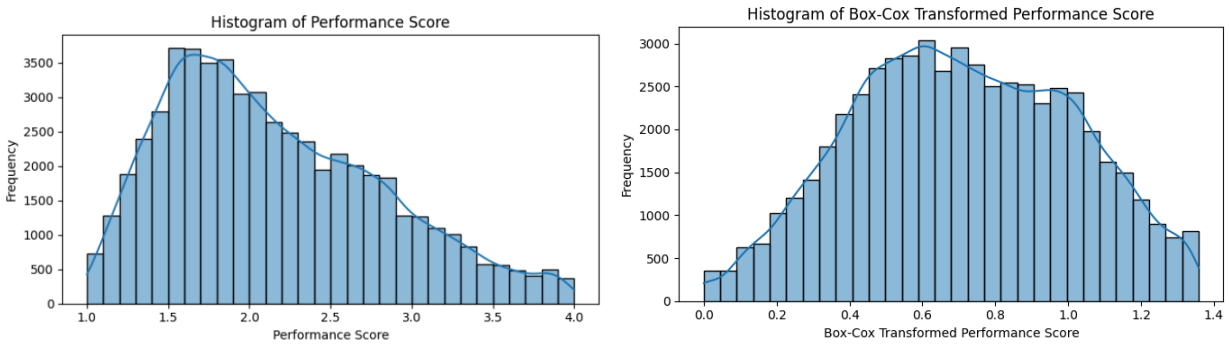

Figure 8: Histogram of Performance before & after Box-Cox transformation

For classification models, a label variable (performance_tier) was also generated by binning raw performance score into four interpretable tiers: Low, Basic, Proficient, and Advanced. Finally, the original percentage variables based on levels were dropped to avoid data leakage after response feature construction.

5. Supervised Modeling & Validation Strategy

Our analysis pursued two objectives. First, we estimated each school's mean Mathematics Regents scale score, a continuous outcome that ranges from 1 to 4. Second, we translated those scores into four policy-relevant tiers—Low, Basic, Proficient, and Advanced. To ensure that performance estimates would generalise, the data were divided into distinct subsets: 60 percent for model fitting, 20 percent for hyper-parameter selection, and the remaining 20 percent for final testing. Linear baselines were tuned with five-fold cross-validation; computationally heavier tree-based and k-nearest-neighbour models were evaluated with three folds to keep run-time within practical limits.

We began the regression analysis with linear techniques—Ordinary Least Squares, Ridge, Lasso, and Elastic Net. These methods are quick to train and offer coefficients that can be interpreted directly. However, because they assume a strictly linear relationship between predictors and outcomes, they struggle with the nonlinear interaction effects that often arise in data. To address this limitation, we next evaluated k-nearest neighbours, a Gradient-Boosting Regressor, and a Random Forest. The Random Forest proved most effective: its ensemble of decision trees uncovered links among demographic characteristics, school climate indicators, and achievement while remaining stable in the presence of overlapping or correlated predictors.

- **Random Forest**

A focused grid search showed that a forest containing 200 trees provided the best balance of accuracy and efficiency (refer to **Figure 1**). At that size, the root-mean-squared error fell from 0.108 (RMSE at 100 trees) to 0.107, explaining 97.7% variance. Larger forests increased computation without major benefit, so the 200-tree configuration was retained.

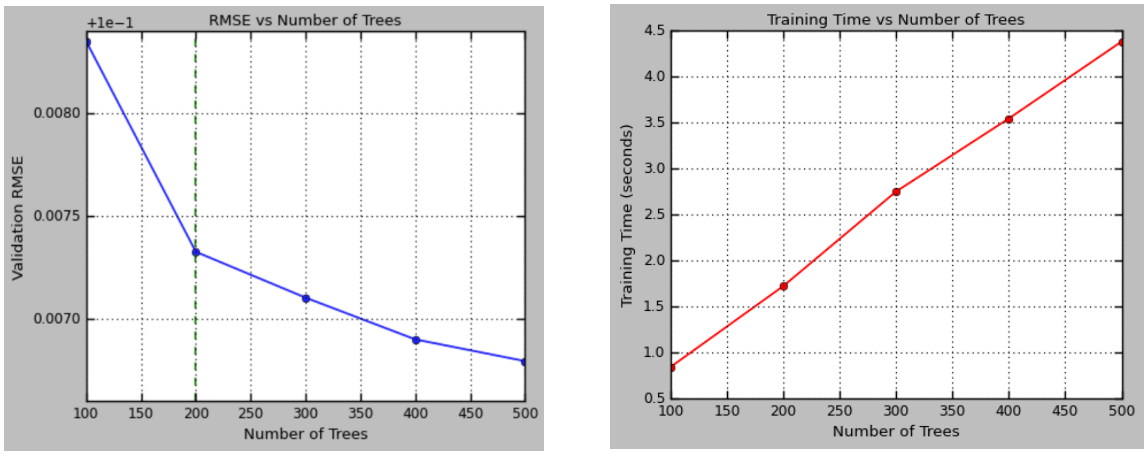

Figure 1: RMSE vs Number of Trees and Training Time vs Number of Trees

Random Forest Interpretation:

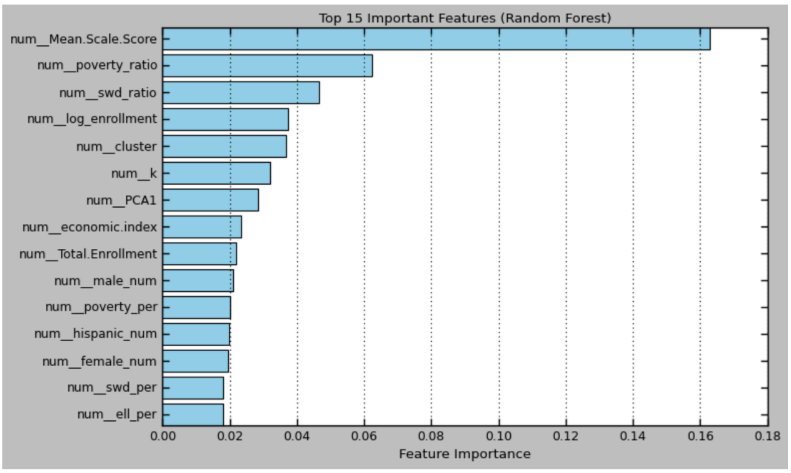

Figure 2: Top 15 most important features as determined by the Random Forest model

The feature Mean.Scale.Score(Average of total students tested) stands out as the most influential predictor, followed by poverty_ratio and swd_ratio (students with disabilities), which matches the analysis of EDA. These features likely capture core academic performance and demographic disparities across schools. Enrollment-related and socio-economic indicators such as log_enrollment, economic.index, and Total.Enrollment also contribute meaningfully to the model's predictions.

- **Logistic Regression and XGBoost**

Our baseline model for the classification portion was logistic regression. It fits a single linear decision surface in the feature space and converts the resulting log-odds into class probabilities. The model works reasonably well—just over 94 percent of Basic schools and nearly 93 percent of Proficient schools are placed correctly—but the model struggles when boundaries curve, most notably misclassifying about 16 percent of Advanced schools as Proficient. By contrast, XGBoost assembles hundreds of decision trees in sequence, letting later trees correct the mistakes of earlier ones and thereby learn higher-order interactions. The added flexibility trims misclassification rates across three of the four tiers: the true-positive rate for Advanced rises from 0.843 to 0.859, Low increases from 0.877 to 0.885, and Basic edges up from 0.938 to 0.941, while Proficient remains essentially unchanged (refer to **Figure 3**).

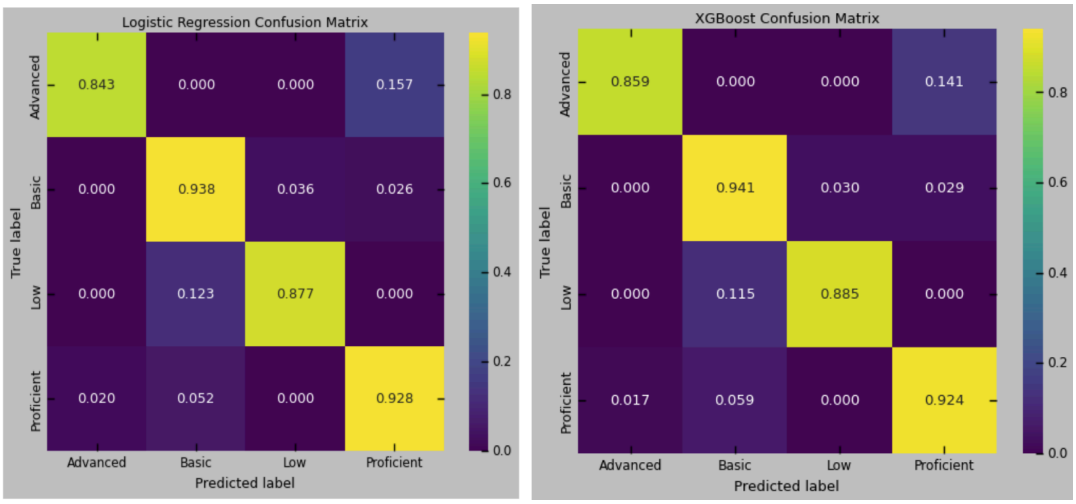

Figure 3: Confusion Matrices of XGBoost and Logistic Regression

Most errors in both models still occur on the two decision boundaries that are hardest to separate—Advanced vs Proficient, and Low vs Basic—but XGBoost narrows those margins without introducing new weaknesses elsewhere. Overall test accuracy rises only modestly, from 0.922 with Logistic Regression to 0.924 with XGBoost, yet that small gain is concentrated exactly where the linear model fails. The trade-off is complexity: interpreting hundreds of boosted trees is far less straightforward than reading a set of logistic coefficients.

6. Combined Models and Regression Results

Two stacking ensemble regressors were implemented using a meta-learning approach to enhance predictive performance and leverage the strengths of multiple regression models. Both pipelines used Ridge Regression as the meta-learner due to its regularization capabilities and strong performance in combining model outputs.

6.1. Stacking: Random Forest + Gradient Boosting + Linear Regression

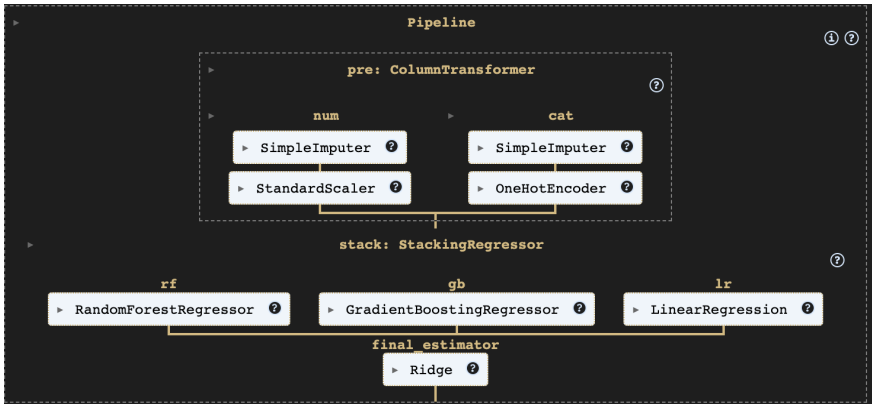

Figure 1: Architecture of the combined model (RF + GB + LR)

The base models used in the stacked regression model were Random Forest, Gradient Boosting, and Linear Regression. The ensemble was constructed using a pipeline that applied preprocessing before passing the data to the stacking model. These models were chosen based on their strong individual performances during baseline testing. However, the final ensemble produced only moderate results.

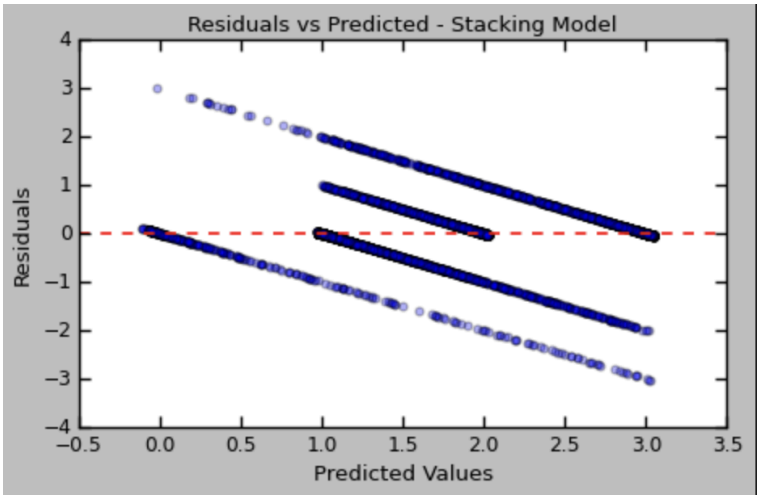

Figure 2: Residual Plot of the combined model (RF + GB + LR)

The performance metrics for the ensemble were as follows: RMSE of 0.443, MAE of 0.214, R^2 of 0.758, and a regression accuracy of 86.8%. While the model explains approximately 76% of the variance, including Linear Regression may have constrained the model's ability to capture the non-linear patterns in the data. The residual plot also reveals a poor model fit, with visible banding patterns and discrete predicted outputs, indicating underfitting and potential misuse of classification-style models in a regression setting. Despite Ridge helping to manage overfitting, this limitation likely impacted the overall model performance.

6.2. Stacking: Random Forest + Gradient Boosting + KNN

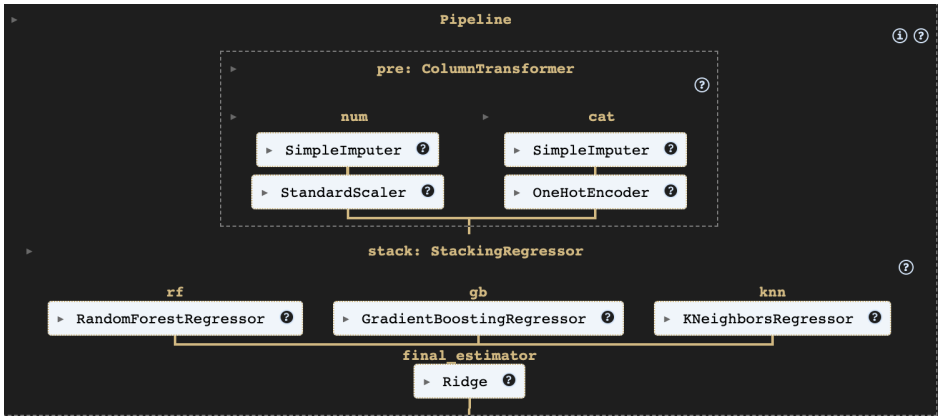

Figure 3: Architecture of the combined model (RF + GB + KNN)

The linear model was replaced with K-Nearest Neighbors (KNN) to improve this, introducing instance-based learning and greater model diversity. KNN is non-parametric and excels at modeling local relationships, which complemented the tree-based learners well. Including KNN as a base model introduces non-linear, instance-based learning into the ensemble, enhancing model diversity. While even though Linear Regression performed slightly better alone, combining different model types often boosts overall performance in stacking.

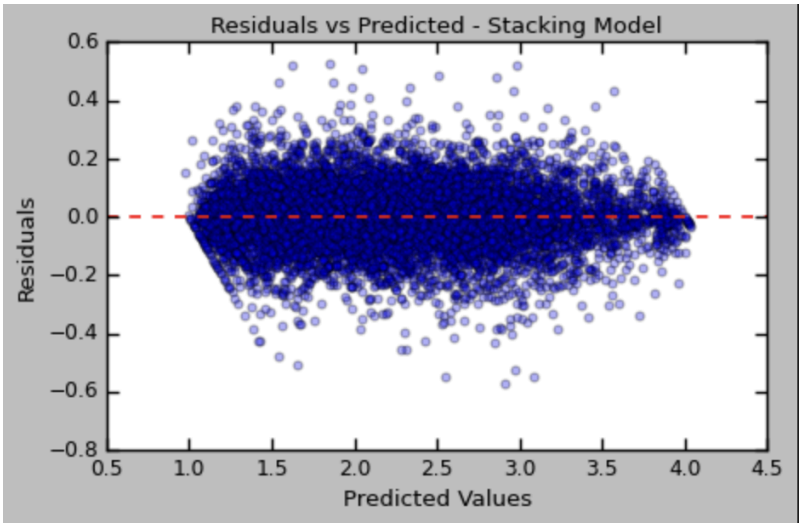

Figure 4: Residual Plot of the combined model (RF + GB + KNN)

This revised ensemble significantly improved performance, achieving an RMSE of 0.101, MAE of 0.073, R^2 of 0.977, and a regression accuracy of 96.6%. The model explained nearly 98% of the variance and outperformed all other regressors, indicating that the fusion of tree-based and distance-based models enabled the stacking ensemble to generalize better on unseen data. A stronger macro F1 score (91.9%) also indicates more balanced performance across all imbalanced classes. This is especially important in scenarios where class distribution isn't even. When we look at the residual plot, the residuals look randomly scattered, with no clear trends or

curvature, and roughly centered around 0, which suggests the model has captured the main structure of the data and isn't missing obvious nonlinearities, and that the stacking model is not systematically over- or under-predicting, which is good.

6.3. Regression Pipeline Architecture

The regression stacking pipeline was structured to follow a layered architecture comprising a preprocessing step, a set of base learners, and a meta-learner. In the preprocessing stage, the input data was cleaned and transformed to ensure it was suitable for modeling. Following this, three diverse base learners—Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbors (KNN)—were employed to capture various patterns within the data. Each algorithm brought a unique perspective: Random Forest captured broad, global trends; Gradient Boosting focused on learning from residual errors through an iterative process; and KNN leveraged local neighborhood information. The predictions from these base models were then passed to a meta-learner, specifically Ridge Regression, which integrated the outputs to produce the final prediction. This ensemble approach enhanced overall model performance by combining the strengths of different learning algorithms.

7. Combined Classifier and Classification Results

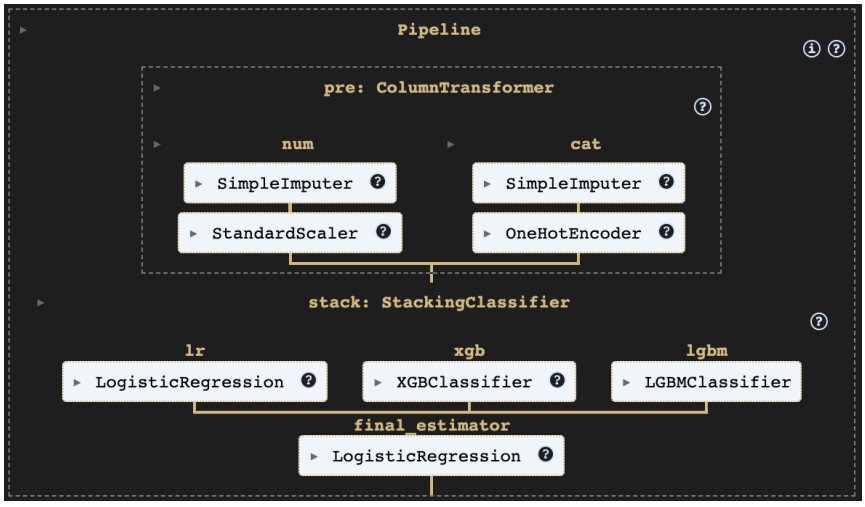

Figure 5: Architecture of the combined classifier

A stacking classifier system integrating Logistic Regression with XGBoost and LightGBM models was used for categorical school performance tier predictions. A model combination was created to manage linear prediction systems and non-linear modeling techniques. We included LightGBM in the stacking classifier to introduce algorithmic diversity and strengthen the

ensemble's ability to model complex non-linear patterns. While it wasn't used as a standalone base model earlier, its gradient boosting approach complements the linearity of Logistic Regression and the depth of XGBoost. A Logistic Regression was selected as the meta-learner for the stacking classifier due to its capable interpretation and stability.

The model demonstrated 93.7% test accuracy and a 91.9% Macro F1 score, which showed that it could be adequately generalized with high accuracy and balanced predictions between different classes. Such strong performance across all categories becomes essential because the dataset has an existing issue with class imbalance.

8. Model Comparison/Selection

8.1. Regression Comparison

Model Combination	RMSE	MAE	R^2	Regression Accuracy
RF + GB + KNN	0.101	0.073	0.977	0.966
RF+ GB + Linear Regression	0.443	0.214	0.758	0.868

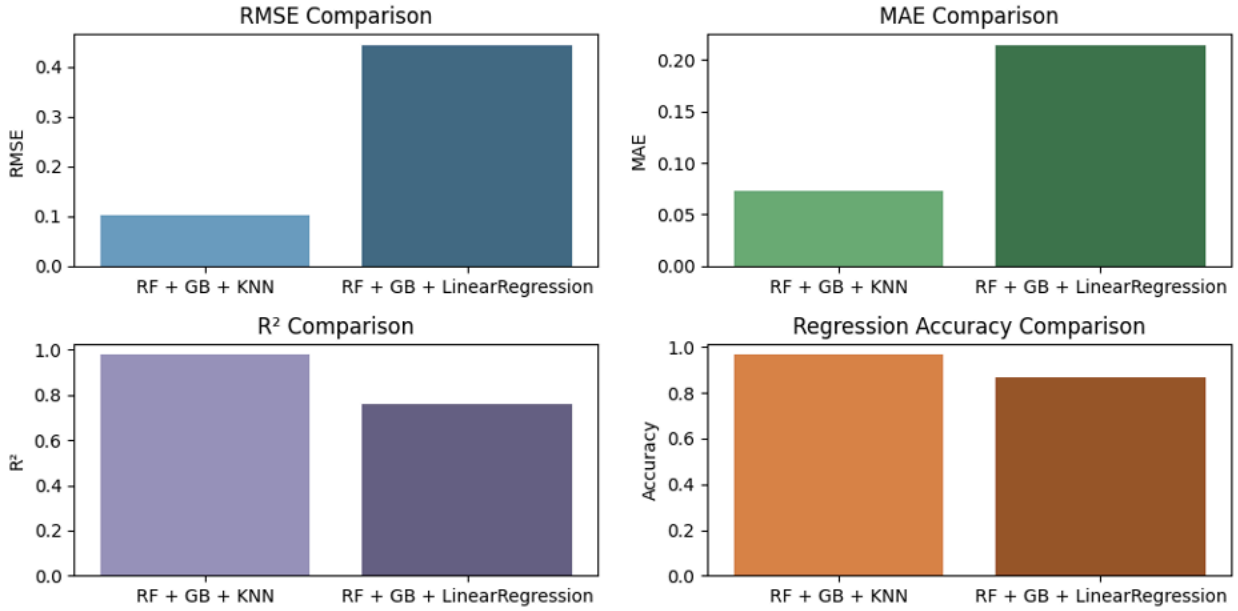

Figure 6: Model Comparison based on RMSE, MAE, R^2 and Accuracy

The bar plots indicate that the stacking ensemble consisting of RF + GB + KNN produced both

the lowest RMSE and MAE results in addition to acquiring the highest R^2 and accuracy levels. Including KNN in the ensemble results in superior performance due to its ability to detect intricate local relations.

8.2. Classification

Model Combination	Accuracy	Macro F1
Stacked Classifier	0.937	0.919

The classification stacking model performs very well, most especially the macro F1 score, showing a good balance in classifying not only the majority performance tier but evenly also the minority performance tier accurately.

9. Final Model Selection

The Random Forest (RF), Gradient Boosting (GB), and k-Nearest Neighbors (KNN) stacked model is the top-performing ensemble for regression tasks. Combining the strengths of each separate model to achieve high R^2 values and low error metrics results in a very effective method to extract complex relationships from data. RF's robustness to overfitting, GB's minimization of residuals, and KNN's simplicity and adaptability make the model fairly generalized on any given regression problem.

In classification tasks, the stacked ensemble of Logistic Regression, XGBoost, and LightGBM is chosen from various ensemble techniques for being highly accurate and having balanced F1 scores for different classes. The ensemble method exploits the interpretability of Logistic Regression, the ability of XGBoost to cope with very big data with copious quantities of the features, and the efficiency of LightGBM to work with big data. Combining these models ensures high accuracy and appropriate balance in handling imbalanced datasets.

Combining the strengths of diverse learning algorithms in both ensemble approaches is successful, and the resulting models are more robust and accurate than any individual model can achieve alone. Such a strategy allows the models to capture the underlying patterns in the data better and, therefore, be highly reliable for real-world applications.

10. Conclusion

In conclusion, this analysis has carefully explored the connection between school environment metrics and academic performance across New York City Public Elementary Schools. Through detailed data collection, rigorous cleaning procedures, and insightful exploratory analysis, we

identified important correlations among demographic variables, school climate indicators, and mathematics test outcomes. Our findings clearly highlight the substantial influence that socioeconomic and demographic factors, such as poverty rates and student enrollment demographics, have on educational performance.

Our modeling approach further deepened these insights. The stacked regression ensemble, integrating Random Forest, Gradient Boosting, and K-Nearest Neighbors, demonstrated superior accuracy and robustness in predicting mathematics achievement scores. This approach effectively captured both broad patterns and detailed, localized relationships in the data. Similarly, the classification ensemble—combining Logistic Regression, XGBoost, and LightGBM—proved highly effective in categorizing schools into meaningful performance tiers. This ensemble method notably balanced accuracy across all performance categories, effectively managing class imbalance.

The successful application of ensemble methods reinforces the value of combining diverse modeling techniques, each contributing distinct strengths to create comprehensive, reliable predictions. The resulting insights offer practical guidance for educational policymakers and school administrators, enabling targeted interventions and more effective resource allocation.

Appendix A - Understanding NYC DOE School Survey Elements

The NYC School Survey evaluates public schools across six key elements that are critical to fostering effective school environments.

These elements are **Rigorous Instruction,** which assesses the quality and challenge of academic programming; **Supportive Environment**, focusing on the emotional, physical, and academic support provided to students; **Collaborative Teachers**, measuring staff collaboration and professional growth opportunities; **Effective School Leadership**, which gauges the administration's ability to inspire and manage the school community; **Strong Family-Community Ties,** reflecting the engagement and partnership between schools, families, and the local community; and **Trust,** which aims at measuring the level of trust and respect among students, teachers, parents, and school leadership.

These elements are from the new Framework for Great Schools, aiming to promote student success and equity. We must note that the scoring is not dependent on survey responses alone. It includes assessments from Quality Reviews which are conducted by trained reviewers and designed to evaluate the effectiveness of schools in promoting student achievement and meeting the diverse needs of their communities. It also includes additional metrics, such as chronic absenteeism.

The NYC Department of Education provides a more detailed explanation and breakdown of the measures in its yearly scoring technical guides.

Task Distribution

Sara Hassani: data cleaning + eda + write report

Jingxi Li: feature engineering + eda + write report

Julieta Caroppo: design baseline machine learning models baseline + conducted initial model performance evaluation + tune hyperparameters + write report

Yuhan Lin: design the whole structure of ml models + broaden model baseline and stacking models + tune hyperparameters + write report

Ruijia Ge: build stacking models + model comparison + write report

Github link: https://github.com/linda664/5243-final-project/tree/main