Achievement Prediction Ensemble Models

Predicting Early Reading and Math Outcomes Using Interpretable Ensemble Learning

1. Executive Summary

This project investigates the determinants of early academic achievement among kindergarten students using a dataset containing student, teacher, classroom and school-level predictor characteristics for the target of student's reading and mathematics scores. The primary objective was to analyse which factors most strongly predicted these reading and mathematics scores and to identify the most accurate predictive models using econometric and modelling approaches.

We began with composing an Exploratory Data Analysis (EDA) to understand our given data quality distribution and key features such as missing values and imbalance data. We discovered that the dataset consisted of 5060 observations and 14 variables spanning 3 conceptual levels of characteristics, students' demographics, teacher's characteristics and school attributes including classrooms. The EDA revealed that some variables contained more missing values in comparison to others. Specifically, `ladder`, `score_read` and `score_math` contained noticeable missing values and greater imbalance data noticeably for `ethnicity` and school-related categories. The target variables, `score_read` and `score_math` displayed distribution of reading score being more stable right skewed than the math score. Correlation analysis suggested that there is relatively a high correlation between reading and mathematical score of r = 0.71 and a weak linear relationship between most predictors and outcomes indicating potential non linear structure of correlation between the proxies.

The data was preprocessed through median imputation for numeric variables, categorical imputation using an 'unknown' category, collapsing of sparse categories and one hot encoding of categorical variables. Redundant or unstable variables such as `ladder` and `schooldistrict_id` were removed. The data was then split into training and test sets of 80 to 20 splits.

The model was trained on both target variables and then on the combined target variables for OLS, Ridge, LASSO, Elastic Net, Decision Tree, Random Forest, LightGBM and XGBoost and Neural Network Models performance was evaluated using Mean Squared Error (MSE). From the result, we found that regularised linear models performed better than OLS, confirming multicollinearity in the predictors. Neural networks underperformed due to dataset size and sparse one hot encoded inputs. Decision trees also showed a strong overfitting most likely as it only uses one tree. With Random Forest showing better results that counteracts the decision tree overfitting. Ensemble boosting models such as LightGBM and XGBoost showed the best result with XGBoost achieving the lowest combined MSE of 990.22 and was selected as our final mode.

Feature importance analysis was using both XGBoost and SHAP showing consistent findings across subjects. The Socioeconomic status proxied by the `lunch` status showed the strongest driver performance. With `free-lunch` showed the strongest influence on high academic performance for both reading and mathematics scores. Followed by the school level effects such as `school_id` and `locations`. Schools located, `rural` areas showed a stronger reading performance whereas the `suburban` school showed better achievement for math scores. This suggested that the learning environment interacts differently for reading and mathematics differently. The classroom variables showed that `small-class` had a positive influence and were consistent with project STAR findings. Whereas Teacher's `experience` and students' demographic variables only had moderate effects. These findings highlighted that students performing based on the targeted variables such as class size and width were correlated with having low SES hence increasing the score of the targets.

Overall the investigation highlights that early achievement is shaped by a combination of socioeconomic background, school environment and class compositions. The result showed and established educational research showing that class size, school quality and SES are key determinants of learning outcomes. XGBoost proved most effective at capturing these multidimensional relationships and providing robust predictive models and interpretable insights via SHAP.

2. Data Descriptions

The given data has a source size of 5060 rows and 14 different columns. These columns include both the target variables and the features, including both numerical and categorical data. The data is designed to examine factors that would relate and influence the early academic performance for students more specifically to the reading and mathematical scores.

2.1 Variables

The dataset is composed of three conceptual levels reflecting the hierarchical nature of the educational environment.

At the student relation (`gender`, `ethnicity`, `birth`, `lunch`, `score_read`, `score_math`) attribution captures the individual demographic, socioeconomic characteristics and academic outcomes.
At the teacher relation level (`degree`, `ladder`, `experience`, `t_ethnicity`), it describes the educator's professional background and experiences. Together, this relation provides insight into how the teacher's qualifications and demographics may influence the student's achievements.
The school relation level (`class_type`, `school`, `schooldistrict_id`, `school_id`) includes features that characterise the broader learning context and institutional settings informing the attributes of the school-level effects on students through different learning environments.

These three different related structures allow the exploration of both students' individual and contextual influences on their outcomes for the target variables `score_read` and `score_math`. This dataset takes into account how the external environment, teachers, and the students themselves are factors that impact their academics.

More specifically, there are differences in the data characteristics that are used for this data. `Lunch` and `ethnicity` are socioeconomic indicators that show the status and diversity of the student and the environment they are in. `degree`, `experience,` and `ladder` will capture the instructional quality and professional development of the teacher's attributes. `School_id` and `schooldistrict_id,` moreover, connect to the hierarchical identification of individual students in broader institutional contexts.

2.2 Data Quality

Noticeably, there are missing variables across the majority of the categories and numerical factors. Certain categories display imbalanced data and outlier distributions that will be further discussed in the Exploratory Data Analysis section. These findings contributed to findings in the further modelling sections.

3. Exploratory Data Analysis (EDA)

3.1 Missing Values

A missing value analysis was conducted to evaluate the data completeness as shown in Figure 1. Our findings show that `ladder`, `score_read` and `score_math` have significantly more missing values in comparison to the other variables. Specifically, it was observed that 8.24% of `score_read` and 6.99% of `score_math` were missing values. Since these are the target variables, any observations lacking these values will need to be removed during the data cleaning and preprocessing stage.
Among the predictors, the `ladder` variable had the highest proportion of missing values (over 10%), followed by `t_ethnicity`, `lunch`, `degree` and `experience`, which all had smaller gaps. To preserve records and avoid unnecessary data loss, missing values in these predictor variables will be treated as a separate category during preprocessing.

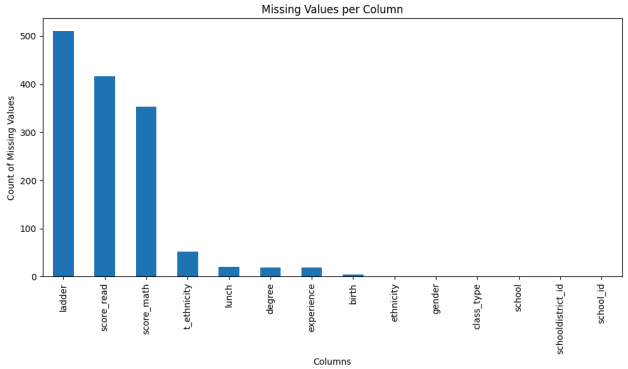


Figure 1: Missing Values per variables

3.2 Distributions of Target Variables

Figure 2 and 3 displays histograms for the two target variables - `score_math` and `score_read`.

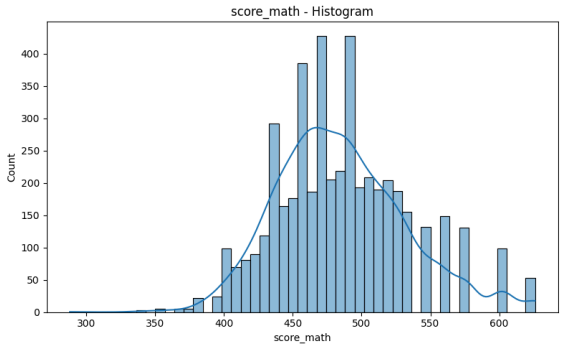


Figure 2: `score_math` to count histogram

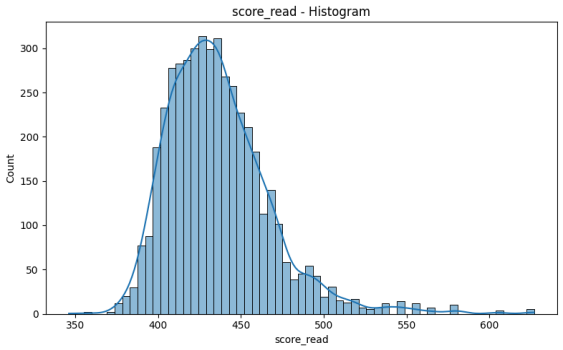


Figure 3: `score_read` to count histogram

The `score_math` distribution appears roughly bell-shaped with a peak at around 450 - 500. The tail for this distribution is slightly heavier toward the right in comparison to the left. This pattern indicates that a regression-type model would be suitable for predicting `score_math`.

In contrast, the `score_read` distribution is close to normal, centred around 400-450, but with a longer right tail, which extends beyond 600. This suggests that while most students perform within a consistent range, a small number achieve exceptionally high scores and very few fall below 400. This indicates that reading performance is generally more stable and consistent across students compared to math performance.

Overall, these distributions show that reading scores are more consistent, as it displays less variability across students. However, math scores vary more widely, implying that performance in math may be more influenced by external factors such as environment or teaching quality. As a result, models predicting `score_math` may require more features or non-linear algorithms to capture these differences effectively.

3.3 Distributions of Predictor Variables

(See plots in Appendix)
Figure 4 - 15 exhibits the distribution of the predictor variables, where categorical variables are presented as bar graphs and numerical variables are presented as histograms on the *appendix* below.

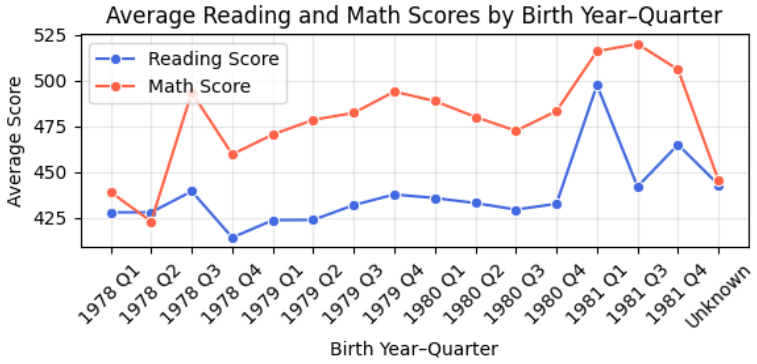
The distributions show well-balanced data for `gender` (Figure 10), `lunch` (Figure 8), `class_type` (Figure 14) and `school_id` (Figure 6). However, there are clear imbalances in *`ethnicity`* (Figure 12), *`ladder`* (Figure 9) and `school` (Figure 7). Since some of these key predictors have very few data points, this imbalance could cause issues during modelling. Therefore, grouping smaller categories into broader ones is necessary to improve model stability and reliability for future analyses.

Teacher-related features, such as `experience` (Figure 11), show a right-skewed distribution, peaking around 10 - 15 years of experience, with only a few teachers having exceptionally high experience levels.

It is also observed that `school_id` and `schooldistrict_id` are not strongly correlated. `school_id` is uniformly distributed while `schooldistrict_id` shows a highly uneven distribution, suggesting minimal relationship between the two identifiers.

Finally, the `birth` variable (seen below) displays a heavy skew toward the 1980 cohort, with fewer data points outside that range. The birth year quarter shows that different quarters are associated with meaningfully different average reading and math scores. For instance, some quarters such as 1981 Q1 shows substantially higher scores in math while others (e.g. 1980 Q3) show noticeably lower scores. These fluctuations are not smooth and

do not follow a simple linear trend across the years. If quarters were collapsed into broader categories (e.g. only birth year), these nuanced score differences would disappear.



Overall, the predictor variables display a mix of both balanced and imbalanced distributions, indicating that further data cleaning and transformation are required to accurately assess the relationship between `score_math` and `score_read` target variables.

3.4 Correlation Analysis

Next, we constructed a label-encoded correlation matrix using all features to examine the relationships between the target variables and key predictors of student achievement. This analysis helped identify potential linear dependencies and redundancies among predictors, as well as the factors most strongly associated with `score_read` and `score_math`.

As observed from Figure 16 of our heatmap (Seen in Appendix), the dataset shows moderate correlations between the two target variables and generally weak relationships between other features, suggesting that achievement is influenced by multiple interacting factors rather than any single dominant predictor.

Key Correlation Findings

- `score_read` **and** `score_math` **(r = 0.71)**

The strong positive relationship indicates that students who perform well in one subject also tend to perform well in the other, validating that both outcomes reflect general academic ability.

- `lunch` **and** `score_read` **(r = 0.25)** / `lunch` **and** `score_math` **(r = 0.22)**

Since categorical variables were label-encoded alphabetically (**free < non-free**), the positive correlation directly implies that **lunch_non-free** students performed better, reflecting the influence of socioeconomic advantage on achievement.

- **`ethnicity` and `lunch` (r = 0.44) / `ethnicity` and `school_id` (r = 0.53)**

The moderate correlation between ethnicity, lunch eligibility, and school ID indicates clustering of socioeconomic and demographic characteristics across schools. This suggests that school-level effects may partially capture SES patterns.

- `experience` **and** `score_math` **(r = 0.09)**

A weak but positive correlation suggests that students taught by more experienced teachers tend to achieve slightly higher math scores. While the effect is small, it aligns with the idea that teaching experience may contribute modestly to student performance.

- `school_id` **and** `t_ethnicity` **(r = 0.33)**

A moderate positive correlation indicates that schools tend to have relatively homogenous teacher demographics, suggesting that teacher ethnicity is somewhat clustered within schools.

- **Other Features** (`gender`, `class_type`, `degree`, `birth`) have very weak or near-zero correlations with performance. These indicated minimal direct linear effect on scores and possible non-linear relationships better captured by advanced models.

These plots reinforce the earlier findings that socioeconomic background and school environment play meaningful roles in student performance, though individual correlations remain modest.

3.5 Outlier Detections

To identify the impact of outliers on model performance, we applied the Interquartile Range (IQR) method to the numerical variables in the dataset `score_read`, `score_math`, `experience`. Under this method, any value outside the interval

$$Lower\ Outlier\ =\ Q1\ -\ (1.5\ \times\ IQR),$$
$$Upper\ Outliers\ =\ Q3\ +\ (1.5\ \times\ IQR),$$

is classified as a statistical outlier. (See appendix for Boxplot of score_read, score_math and experience with outliers identified)

Interpretation of findings:

Variable	*Number of outliers*	*% of data*	*Interpretation*
`score_read`	114	2.25%	There were high and low performing students outside the lower and upper bounds.
`score_math`	166	3.28%	
`experience`	31	0.61%	All taking the same value - 27 years

As shown from Figure 19 and 20, there are extreme outliers that were above the 600 range for both `score_read` and `score_math`. Outliers can negatively affect model performance, particularly for methods such as OLS, Ridge, and Lasso, which are sensitive to extreme values. For instance, linear regression models can be heavily influenced by a few very large residuals, thus causing the fitted line to shift toward these extreme observations. This reduces accuracy for the majority of students and leads to poor generalisation.
Additionally, all 31 outliers in the `experience` variable correspond to teachers with exactly 27 years of experience. Since they represent a very small portion of teachers, they could cause the model to overemphasise the impact of highly experienced teachers.
Outliers tend to inflate prediction errors and can disproportionately dominate the loss function since squared errors amplify the effect. Removing these outliers would allow us to improve model stability and help reduce the mean squared error (MSE).

4. Methodology

4.1 Cleaning Data and Preprocessing

To prepare the dataset for modelling, we applied several preprocessing steps to ensure consistent and logical reasoning behind each of our methods. Each decision was guided by data characteristics identified in the Exploratory Data Analysis.

Handling Missing Values

- Target Variables

We identified missing values across both of our target variables (`score_read` and `score_math`) and the predictor variables. Observations with missing values in the target variables were removed, as models cannot learn scores against labels which don't exist. Keeping these missing target values would introduce structural bias, distort the loss surface, and produce unreliable metrics (i.e. MSE). Therefore, dropping these observations help maintain the validity and reliability of our model evaluations and guarantee a valid loss.

- Numerical Predictor Variables

For numerical features, missing values were imputed using the **median**, chosen based on the degree of skewness in their distribution. Specifically, `experience` was imputed with its median, as this measure is less sensitive to outliers. The `experience` variable showed a right-skewed distribution with a few extreme values (e.g. 27 years), making the median imputation more appropriate than the mean. We chose to input the median for these values over the mean because of the distribution to preserve the central tendency without being affected by the outliers. Median imputation is also more robust than the mean imputation in smaller or asymmetric datasets to maintain the integrity of teachers' experience data.

- Categorical Predictor Variables

For categorical predictors with missing values such as `ethnicity`, `birth`, `class_type`, `t_ethnicity`, `degree`, `ladder` and `lunch`, a new category labelled 'Unknown' was introduced.
This method was chosen for two main reasons:

1. **Data preservation**: Removing rows with incomplete categorical data would reduce the sample size and introduce bias. By creating an 'Unknown' category, this allows us to retain all observations

2. **Information value of missingness:** In educational and demographic datasets, the absence of data can be informative. The model can learn patterns associated with the 'Unknown' category to help prevent bias exclusion and worsening of class imbalance

Dropping Redundant Features
From Figure 1, we observed that there were larger amounts of missing values for the two variables: `ladder` and `schooldistrict_id`.

`The ladder` contained over 10% missing values which is a very high proportion, and displayed limited variability. The correlation analysis also indicated overlap with other proxies, such as `lunch` and `ethnicity`. Retaining this could introduce noise and multicollinearity; therefore, it was removed.

`Schooldistrict_id` was also dropped because it provided redundant geographical information already captured by `school_id`. Keeping both variables risk noise and adding unnecessary complexity.
Removing these redundant features helped improve model interpretability and reduced overfitting risks arising from correlated predictors.

4.2 Encoding Categorical Variables

All categorical variables were converted into numerical form using `one-hot encoding` to ensure compatibility with machine learning algorithms. Most regression and tree-based models require numerical input as they cannot directly interpret text labels or qualitative categories. By using `one-hot encoding,` it transforms each categorical feature into a series of binary 0 or 1, or further if there are more than 2 categories, indicator variables where each column will represent one category. This will also allow the use of regularisation methods such as lasso or tree-based models to automatically shrink, drop or split into categories without risk of overfitting. To prevent the dummy variable trap, one category from each feature was excluded as the reference level during encoding.

4.3 Addressing Class Imbalance

We observed moderate imbalance across some categorical variables, particularly `ethnicity` and `degree`. Instead of applying oversampling or undersampling, which could distort the dataset's distribution, we grouped rare categories into a single 'Other' group to maintain sufficient representation and reduce sparsity.

- `ethnicity` was collapsed into three groups: *Cauc*, *Afam*, and *Other* (combining smaller groups).
- `degree` was simplified to *Bachelor*, *Master*, and *Other* (grouping less common qualifications under 'Other')

This grouping preserved the dataset's demographic realism while ensuring that each level had sufficient observations for reliable estimations.

4.4 Outcome of cleaning and preprocessing

After cleaning, imputing , and encoding, the dataset was fully numeric and free of missing values with reduced redundancy and balanced categorical levels. These preprocessing steps ensured that the dataset was ready for

modelling, allowing subsequent regression and ensemble algorithms to operate on consistent, unbiased and interpretable features.

4.5 Train-Test Split and Model Setup

Before fitting models, the dataset was divided into training and test subsets using an 80/20 split. The training set was used to fit and tune all the models, and the test set remained completely unseen from fitting and training the model and was used in our final evaluation. This ensured that the model performance metric was reflected based on generalisation rather than being overfit.

All the cleaning processes were implemented using a `scikit-learn` pipeline, fitting only on the training data to prevent any data leakages. Numerical variables were standardised using `StandardScaler,` and categorical variables were one-hot encoded to suit all types of models
As our dataset contains two separate target variables, all models were trained twice. Once to predict the `score_read` and `score_math.` Then the model performance was evaluated using Mean Squared Error (MSE) on the test set as well as using cross-validation for hyperparameter tuning.

5. Models

5.1 Overview of Modelling Strategy

To ensure our models are treated fairly and are compared in respect to each other, we implemented a diverse set of models ranging from simple linear models to highly flexible linear models. In each model, our goal was to evaluate how well simple linear effects explain variation in scores. Whether more complex nonlinear interactions improve accuracy, and which features are most influential across models.

5.2 Linear Regression and Regularised Models

Linear regression was implemented as a baseline model to predict reading and math scores. It provides a simple, interpretable model where the relationship between predictors and outcomes can be directly understood through coefficient estimates. Establishing a linear baseline allows us to compare the results of a baseline model with more complex non-linear models later on.

To account for risks of multicollinearity, overfitting and high model variance, regularisation methods of linear regression were also tested: Ridge, LASSO, and Elastic Net Regression. These models add penalty terms to the loss function and restrict the magnitude of coefficients to improve generalisation and interpretability.

Ordinary Least Squares (OLS):
OLS regression estimates model coefficients by minimising the sum of squared errors between the observed and predicted values. It fits all features simultaneously without any penalty terms. The model can be expressed as:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$$

where Y_i represents the dependent variable (test score), X_{ij} are predictor variables, B_j are the coefficients, and ϵ_i is the error term.
OLS provides a clear interpretation where each coefficient represents the expected change in the predicted test score. However, it is sensitive to multicollinearity and noise and could possibly inflate coefficient variances and reduce predictive accuracy.

Regularised Linear Models

To address multicollinearity and reduce overfitting, we implemented Ridge, LASSO and Elastic net where all using cross-validation versions (`RidgeCV`, `LassoCV`, `ElasticNetCV`) each with cv = 5.

Ridge Regressions (L2 Regularisation)
Ridge regression adds an L2 penalty term to the OLS loss function

$$\lambda \sum \beta_j^{\,2}$$

This shrinks the coefficient toward zero, reduces variance and stabilises estimates when predictors are highly correlated. Ridge regression is most effective when most features have small but non-zero effects.

LASSO Regression (L1 regularisation)
LASSO regression adds an L1 penalty term

$$(\lambda \sum |\beta_j| \ )$$

which shrinks coefficients and performs feature selection by driving some coefficients exactly to zero. LASSO is particularly useful for identifying the most influential variables and simplifying the model.

Elastic Net (L1 + L2)
Elastic net combines Ridge regression and LASSO regression penalties to balance variable selection and coefficient shrinkage. It is highly effective when there are correlated predictors or when the number of features is relatively large compared to the number of observations. Hence Ideal when predictors are correlated, which is common in one-hot encoded data.

Result and Observations
Observations from the results indicate that regularised models for both reading and math scores outperformed OLS which confirms the presence of multicollinearity. Ridge regression achieved the lowest MSE value for reading scores (627.1401), while Elastic Net achieved the lowest MSE value for math scores (1830.1289). However, Ridge regression achieved the lowest combined MSE value (1229.02). This could be due to the fact that Ridge shrinks coefficients smoothly without removing variables entirely.
All three regularised models performed similarly which reflects the effectiveness at stabilising coefficients and preventing overfitting. Overall, the regularised models performed soundly on this dataset since the majority of predictors are categorical and one-hot encoded. The one-hot encoding produces a structure that tends to produce many correlated binary indicators which regularisation methods are specifically designed to handle.

5.3 Neural Network (MLP Regressor)

A Multilayer Perceptron (MLP) Regressor was implemented to capture potential non-linear relationships between student, school and contextual features that may not be well represented by linear models. Neural Networks were chosen for their ability to learn complex feature interactions, specifically where the effect of one variable depends on the value of another which cannot be done by standard linear models.

The model used a feed-forward method with two hidden layers of 64 and 32 neurons with ReLU activation functions. The network was trained using Adam optimiser with an adaptive learning rate, with a maximum of 500 iterations and early stopping to prevent overfitting when validation loss does not improve. Hyperparameters were tuned using GridSearchCV with a 5-fold cross validation to ensure the model generalises well.
The optimal parameters for reading scores:

```
{'model__activation': 'relu', 'model__alpha': 0.0001, 'model__hidden_layer_sizes':
(64, 32),
'model__learning_rate_init': 0.003, 'model__solver': 'adam'}.
```

The optimal parameters for Math scores:

```
{'model__activation': 'relu', 'model__alpha': 0.001, 'model__hidden_layer_sizes':
(64, 32),
'model__learning_rate_init': 0.003, 'model__solver': 'adam'}.
```

These tuned models reduced the MSE compared to the untuned baseline MLP. However, even after tuning, the neural network achieved a slightly higher MSE compared to the regularised linear models (Ridge, LASSO, Elastic Net). This is likely due to the dataset's relatively small size and the fact that it predominantly consists of categorical features. Categorical variables are often better suited to simpler and more interpretable models such as Ridge and LASSO. Neural Networks generally require large datasets to learn stable non-linear representations, and with limited data they may overfit or fail to learn robust patterns.

Additionally, after preprocessing, most variables in the dataset become sparse One Hot Encoder indicators. Neural Networks do not handle sparse data well because information density is low, gradients become noisy, and interactions between dummy variables are weak and hard for the network to learn.
Despite using such a flexible model, Neural Networks did not outperform the linear baselines which shows that model complexity does not always mean better predictive performance.

5.4 Decision tree

A decision tree regression was implemented to explore whether simple nonlinear splits in the feature space could capture meaningful structure in the data. Decision trees recursively partition the data into homogeneous regions by selecting predictor variables and split points that minimise variance within each node. This allows trees to naturally model interaction and nonlinear relationships that linear models cannot express.

Decision trees were also used because they work well with mixed categorical and numerical data, especially after one-hot encoding. They also provide interpretability through visible decision rules, and they serve as a useful baseline for comparing the performance of ensemble tree methods such as random forest, XGBoost and LightGBM.

However, single decision trees are known to be high variance models, meaning small changes in the training data can produce very different trees. Hence, without regularisation through techniques such as max depth lines, they tend to overfit and learn noise rather than general patterns.

We trained separate decision tree models for reading and math scores using the scikit-learn implementation. Numerical variables were standardised, and categorical variables were one-hot encoded through the cleaning process.
We then used no aggressive hyperparameter tuning to perform for the base tree, as the purpose was to compare its natural behaviour against more advanced ensemble methods. This allows us to observe how poorly a single tree generalises relative to its ensemble counterparts.

These results demonstrate that the decision tree performs substantially worse than other regularised linear models and the ensemble tree-based method as expected.
We can interpret that the high MSE values indicate overfitting as trees tend to memorise the training data, hence they struggle to generalise when data contains noise or weak signals. The performance gap between reading and math reflects earlier findings that math scores are more variable, with weaker feature outcome relationships and more noise. The results confirm that a single tree lacks stability and fails to utilise the full structure of the dataset, particularly when many features come from one-hot encoding.
Hence, this finding justifies the use of other ensemble methods, such as random first, XGBoost and LIGHTGBM, which specifically address the instability and high variance inherent in decision trees.

5.5 LightGBM

LightGBM is a tree-based gradient boosting model designed to optimise speed, scalability, and performance. To achieve this, it uses a histogram-based algorithm that buckets continuous values into discrete bins, greatly accelerating the split-finding process. Unlike other boosting models, LightGBM grows trees leaf-wise, selecting the leaf with the highest error to expand next. This focuses model capacity on the areas with the largest mistakes and enables the algorithm to scale to large datasets while maintaining low memory usage. It also works effectively with both numerical and categorical features and is generally robust to moderate class imbalance.

LightGBM also includes several built-in regularisation mechanisms, like `min_data_in_leaf`, `num_leaves`, and `max_depth,` which can help reduce overfitting despite its aggressive tree-growing strategy. However, the model can still overfit on small datasets and may react more sensitively to noisy or poorly encoded features. As a result, it performs best on clean, well-prepared data. We trained separate models for score_read and score_math to allow direct comparison between the two targets.

The results for reading show a relatively low MSE of 478. This indicates that the errors are consistently small across observations. However, for math, the MSE is higher at 1507, possibly reflecting a noisier, more complex relationship with the data.

After establishing these baselines, we checked for any outliers or noise that could have affected the model. However, no clear data issues were identified. The outliers present fell within a normal performance range, with

score variations of roughly 100 points, which was well within expected behaviour for this dataset. Since these values reflected genuine differences rather than errors, there was no meaningful noise to remove. Therefore, we proceeded with hyperparameter tuning using a combination of grid search and random search to explore a wide range of parameter configurations.

After running grid search and cross-validation, the best results are: colsample_bytree=0.8, learning_rate=0.03, n_estimators=500, num_leaves=31, subsample=1.0.

The results show a slight reduction in both outputs, with the tuned model achieving an MSE of 467 for reading and 1514 for math. This indicates that hyperparameter tuning provided a modest improvement. Overall, reading scores were predicted with higher accuracy, likely because they have stronger relationships with the available features. In contrast, math scores remained more variable, suggesting that additional noise or unobserved factors may be influencing math performance.

5.6 Random forest

A Random Forest (RF) model was implemented to explore possible non-linear relationships across student, teacher, and school characteristics. The model works by building a larger number of decision trees, each trained on slightly different samples of the data, and then combining their predictions. By averaging across multiple trees, Random Forest reduces the impact of noise from any single tree and creates a more stable and flexible model that can capture interactions between features.

These results suggest that the Random Forest model was able to capture some of the underlying structure in the data, particularly for reading scores where prediction errors remained relatively contained. In contrast, the higher MSE for mathematics suggests that this outcome was harder for the model to learn, potentially due to weaker feature-outcome relationships or greater noise in the data. While the model could handle complex interactions and worked well with the mix of categorical and numerical variables, its overall error levels indicated that it was not making the most effective use of the available information for this dataset.

5.7 XGBoost

An XGBoost model was used to better capture the non-linear relationships and interactions present across the student, teacher, and school variables. Unlike traditional decision-tree methods, XGBoost builds its trees one at a time, with each new tree trained to correct the mistakes of the previous ones. This step-by-step learning process allows the model to focus more on the difficult cases in the data, while its built-in regularisation helps prevent overfitting. As a result, XGBoost provides a flexible yet well-controlled approach for modelling complex patterns in the dataset.

After fitting the initial base models, hyperparameter tuning (GridSearchCV) was carried out to further improve performance. GridSearchCV is a systematically evaluated parameter combination to identify the configuration that minimised cross-validated mean squared error. The grid search explored variations in tree depth, number of estimators, learning rate, subsampling parameters, and regularisation terms. This process produced the following best-performing parameter sets:

Reading (Grid Search Best Params): `n_estimators = 300, max_depth = 3, learning_rate = 0.1, subsample = 0.8, colsample_bytree = 1.0, gamma = 0.1, reg_lambda = 1, reg_alpha = 0`

Math (Grid Search Best Params): `n_estimators = 300, max_depth = 3, learning_rate = 0.1, subsample = 0.8, colsample_bytree = 1.0, gamma = 0, reg_lambda = 1, reg_alpha = 0.5`

These tuned settings provided a strong balance between flexibility and regularisation, enabling the model to learn meaningful structure while controlling overfitting. Using the tuned models, predictions were generated for reading and math scores.
The XGBoost models were effective at capturing non-linear interactions and adjusting to areas of the dataset with higher predictive difficulty. The model performed particularly well for reading, where stronger relationships in the data allowed for more accurate predictions. In contrast, math remained more variable, suggesting that additional unobserved factors may influence performance in that area. Overall, XGBoost demonstrated strong predictive ability across both outcomes and made efficient use of the available information through boosting and regularisation.

Given its combination of flexibility, strong regularisation, and the lowest overall prediction errors among the models tested, **XGBoost was selected as the final modelling approach for this project**. Its performance, particularly on the reading and math outcome and its competitive combined MSE, made it the most suitable model for capturing the underlying patterns in the dataset.

6. Results

6.1 Model Performance Overview

All models were trained separately on reading and mathematics scores and evaluated using Mean Squared Error on an unseen 20% test set.
We also computed a combined MSE (both read and math score) defined as the mean of the reading and mathematics MSEs.

6.2 Mean Squared Error

Mean Squared Error (MSE) is used to measure the average squared difference between the predicted values and the true target values,

$$MSE = \frac{1}{n}\Sigma\left(y_i - \widehat{y_i}\right)^2$$

where y_i is the true score and $\widehat{y_i}$ is the predicted score.

As the performance metric, the smaller the MSE values, it suggests that there are less errors with the predictions. MSE also penalises larger errors more heavily due to the squared term, so it provides a single iterable value for comparing models.
Since our project has two target outcomes, we also computed a combined MSE, defined as the mean of the two individual MSEs.

6.3 Final Model Results

The table below summarises the performance of all models tested:

Model:	Score_read MSE:	Score_math MSE:	Combined MSE:
Linear regression/OLS	624.7333	1841.2630	1232.9982
LASSO	627.1653	1830.8222	1228.9938
RIDGE	627.1401	1830.9069	1229.0235
ELASTIC NET	627.1653	1830.1289	1228.6471
Neural Net	694.3857	1800.4583	1247.4220
Decision Tree	801.2906	2694.7138	1748.0022
LightGBM	467.4188	1514.1838	990.8013
Random Forest	545.7758	1659.1216	1102.4487
XGBOOST	**467.4503**	**1513.2082**	**990.3292**

We see that across all measurements, including `score_read`'s MSE, `score_math`'s MSE and combined's MSE, XGBoost outperformed with the lowest MSE in comparison to any other models in order to predict kindergarten students' reading and mathematics performance.

6.4 Interpretation of Model Behaviours

The performance of each model reflects how well its underlying assumption aligns with the structure of our dataset. The linear models (OLS, Ridge, LASSO and Elastic Net) produce similar combined MSE values, which suggests that predictors between students' test scores are approximately linear, especially after one-hot encoding of categorical variables. Among the linear models, Elastic Net performed the best, likely because its combined L1 and L2 penalty stabilises coefficients in the presence of correlated variables while retaining feature selection capability.

The neural network underperformed relative to the linear baseline, which usually tends to excel with large and high-dimensional continuous datasets; however, given that our dataset is small and is dominated by sparse one-hot encoded features, it limited the benefits of a deeper nonlinear model. Hence, even with the hyperparameter during the MLP exhibited signs of overfitting and did not generalise well, it has a higher MSE.

The decision tree achieved the poorest results amongst the trees, as expected. Decision trees are high-variance models that overfit readily, especially on datasets with many categorical splits and limited data points per category. The unstable nature of single trees and their tendency to memorise training examples explains the large error observed here. Hence, Random Forest is, in comparison to the decision tree, improved substantially.

The gradient boosting models, such as lightGBM and XGBoost, on the other hand, achieve the strongest results, with XGBoost performing the best overall. Boosting sequentially learns from residual errors and enables to identification of subtle localised interactions that linear models and Random forests cannot capture, hence it was able to achieve the lowest MSE overall for all read and math scores and combined scores

6.5 Explanation Results

Since the performance difference between LightGBM and XGBoost was marginal, we next considered their model complexity. Both are boosting algorithms and therefore share similar structures, but XGBoost is known to be slightly more conservative and stable. Based on both the performance comparison and model-complexity consideration, we selected **XGBoost** as our final model.

7. Drivers of Performance

7.1 Method - XGBoost Feature Importance and Comparisons

To identify the drivers of performances, we compared feature importance across all our models. (See Appendix for feature importance values and plots for all models assessed)

Linear Models & Regularised Models
Linear models provided coefficient based importance that reflect additive effects. Regularised models stabilise these coefficients by shrinking noisy or highly correlated drivers.

Tree-based Models
Tree based models help verify variables that improve splits in non-linear interactions. Since XGBoost was selected as our final model, we decided to use SHAP as well to understand more about the direction of how each variable affects the predictions.

Shapley Additive explanations (SHAP):
SHAP is an approach that assigns an importance value to each feature, explaining how much each feature contributes to a model's prediction. Each SHAP value represents the direction and magnitude of a feature's effect on the model's output and shows whether it pushes the prediction higher or lower.

SHAP calculates each feature's contribution independently and then sums these contributions to match the model's final predictions. This method ensures local accuracy as the SHAP values add up to the exact model output for each data point.

A SHAP value of zero means that a feature is either missing or irrelevant for that specific prediction. We chose to use SHAP because it provides a dedicated tree explainer that runs efficiently on tree-based models such as XGBoost.

Advantages
- Shows whether a feature increases or decreases the prediction
- Handles interactions
- Locally accurate (SHAP values sum to model output)

Limitations
- Computationally more expensive
- Harder to interpret with many correlated or one-hot encoded features

7.2 Results

Comparing Feature Importance Linear models and Non-linear models:

Across OLS, LASSO, Ridge, Elastic Net, Random Forest, LightGBM and XGboost, we observe clear similarities and differences in how each method identifies the drivers of student reading and math scores. Although each modelling technique produces its importance measures differently, many features are shown to be consistently influential.

Linear Models (OLS, LASSO, Ridge, Elastic Net):

Reading scores:

The linear models direct interpretability because their coefficients show both direction and magnitude of each feature. For example, in OLS, the strongest reading predictors are dominated by birth year-quarter variables such as birth_1981 Q1 (50.17) and birth_1978 Q4 (-23.19). These large coefficients likely reflect cohort specific differences in age or unobserved developmental factors. Since OLS does not regularise or shrink coefficients, it can over-emphases dummy variables that have noise.

In contrast, the regularisation methods (LASSO, Ridge, Elastic Net) produce more stable and less extreme coefficient patterns because they shrink noise variables. These models have a wider range of predictors rather than overfitting to birth-quarter features. Regularisation consistently selects lunch_free/ lunch_non_free, ethnicity, class_type_small, and gender_male as the top predictors with large coefficients across reading scores. This confirms that these features have a consistent association with student reading performance.

Math scores:

Math results show a similar pattern but the magnitudes differ slightly. Math scores appear to be more strongly tied to demographics such as ethnicity and gender and lunch status compared to reading scores. Birth year-quarter effects remain present in math but become less dominant after incorporating the regularisation methods.

Tree-Based models (Random Forest and LightGBM)

Tree-based models that we have used such as Random Forest and LightGBM do not highlight any marginal effects but specifically highlights the variables that are the most useful for partitioning the data through non-linear interactions.

All three tree based models had the same handful of variables that dominated model performance. The top drivers shared across all models include: school_id, lunch status, teacher experience, ethnicity, gender, class type, and birth year-quarter indicators. The consistencies of these predictors across multiple modelling approaches suggest that underlying relationships in the data are robust.

Although general features were present, the rankings and values differ. The strongest drivers for reading scores in Random forest include school_id (0.217) and experience (0.177), and lunch status (lunch_non-free being 0.083 and lunch-free being 0.087), and ethnicity (0.0529). This could be because the model frequently splits on variables that create pure splits and a high cardinality variable like school_id are repeatedly selected.

Reading scores:

For reading scores, LightGBM shows that school_id (5408) and experience (2762) received far larger importance values compared to gender (781), class_type_small (568) and lunch_free (532). This could be due to the fact that LightGBM amplifies the importance of high-cardinality values, reflecting how school_id strongly dominated the reading outcomes.

XGBoost emphasises demographics, classroom structure and student features as more stable contributors to reading performance. Our chosen model, XGBoost produces a more balanced but consistent profile that highlights how lunch status, teacher experience, student ethnicity (being Caucasian), small class type and rural and suburban schools had the highest impact. lunch_free is the strongest predictor (0.234), far above all other features. It is followed by ethnicity_cauc (0.057), school_rural (0.054), class_type_small (0.049) and school_id (0.046).Overall, XGBoost produces a balanced profile that highlights how lunch status, ethnicity, school context, class type and teacher experience drive reading predictions.

Math Scores:
The top random forest importance ranking for math includes: school_id (0.2149), experience (0.1822), lunch_non-free (0.068), lunch_free (0.0649) and ethnicity (0.0603). Math scores here show a similar pattern to its reading scores but ethnicity and lunch status become slightly more influential relative to class type or birth cohort. The top drivers for LightGBM math scores were: school_id (5025), experience (2921), gender_male (884), class_type_regular+aide (624) and class_type_small (545). Compared to the reading performance, the math performance appears to be more sensitive to classroom structure and gender with male students showing a particularly high performance.

XGBoost math drivers are similar to those for reading: lunch status, ethnicity categories, class type (small or aide supported), school ID. XGBoost math drivers show lunch_free (0.135) and lunch_non_free (0.116) as the strongest predictors, followed by ethnicity_cauc (0.066), birth_1980 Q3 (0.065), school_id (0.045), class_type_small (0.044), and experience (0.033). Gender_male (0.031), school_suburban (0.039), and class_type_regular+aide (0.024) also play important roles in math predictions.

Shared consistencies across all models

Despite different modelling methods, the same features seem to dominate across every model which suggest that the relationships captured by the data are robust. The shared top drivers across OLS, regularised models, random forest and LightGBM and XGBoost include:

- School_id
- Lunch status (free, non-free)
- Teacher experience
- Ethnicity of the student
- Gender (male)
- Class type (small)

Lunch status is consistently a strong predictor of both reading and math scores for students but tends to have larger negative coefficients/importance for math which could be a reflection of stronger socioeconomic gradients in numeracy.

Although Decision Trees and Neural Networks were included in our modelling section, we did not report feature importance for these models because Decision Trees are extremely unstable and unreliable due to their sensitivity to overfitting, especially on smaller datasets like ours. In our results table, the Decision tree model showed noticeably worse MSE compared to the rest of the models, confirming that the single tree is not a stable estimator. Therefore, its feature importance would mislead interpretation unlike the rest. Additionally, since we used One-Hot-Encoded sparse inputs, Neural Network performed poorly on the sparse data which would produce unreliable or uninterpretable feature importance results. Thus, we decided to not include the feature importances for these two models that achieved the two highest MSE scores.

Drivers of Reading and Math Scores - XGBoost SHAP

Across both reading and math scores, we found that the strongest feature that affects students' scores is **Lunch**. SHAP values show that free lunch had a positive magnitude for both reading and math scores, indicating that it has a positive effect on predicting a higher score. Meanwhile, students who did not receive free lunch, presumably higher SES, showed a slight negative effect with a mean of -0.003. This result does not mean that low-income

students are naturally doing better or vice versa. Instead, it reflects the model, showing its interactions with other variables like class_type and teacher characteristics.

School ID is also one of the most important features based on our SHAP results. Some schools seem to help students do better, while others are pulling it down. However, this is not something we can directly interpret since it is just a label for each school.

Experience is also an important factor with a positive SHAP value of about 0.5. From this, we can infer that having a more experienced teacher is most likely linked to having a slightly higher predicted score. Students thought by teachers who have been in the classroom longer tend to do better which fits with what we expect.

The student's ethnic background also showed up as an important factor with the model's predictions. Students from different ethnic groups have slightly different predicted scores even when they are from the same class and taught by the same teachers. This does not mean ethnicity itself if causing the differences but more likely due to other inequalities in education such as resources and learning support. In short, ethnicity matters in the model but not directly as a student disability but more of the deeper social and educational inequalities.
Class type , school location

Differences Between Reading and Math Drivers - XGBoost SHAP

While reading and math share many drivers, there are some variables that affect differently based on the subject.

Gender had a stronger impact in reading than math. With **male** students' grades are slightly more negatively associated with reading scores. It supports existing research showing that **females** tend to develop literacy skills more quickly than males. **Birth** also had an impact, suggesting that students who are younger tend to have slightly lower reading predictions, likely reflecting maturity and developmental readiness.

In contrast, math performances were influenced by school-level factors and SES patterns. From the XGBoost findings, we can see that lunch_not_free was the top driver indicating that highSES students do better for math. In addition to that, class size and teacher experience was also consistently important in influencing math, suggesting that the predicted grades for math are highly driven by teacher's experience, and resources.

Overall, reading scores tend to be driven by socioeconomic, demographic and development factors while math scores are driven by school context and SES advantages.

8. Discussions

8.1 Model Complexity and Interpretation

Our models demonstrated a clear result to highlight a clear tension between model complexity and interpretation. On the other hand, a relatively simple regularised linear model already performs well, which tells us that a large fraction of variation in kindergarten test scores is explainable through straightforward additive relationships with observable characteristics. On the other hand, it also shows that the boosting models consistently outperformed in terms of MSE in comparison to the other models, showing that there are additional nonlinear interactions that these models captured.

Choosing XGBoost as the final model is therefore a tradeoff of gaining predictive accuracy and a better fit to the underlying data-generating process, but we lose the direct coefficient-based interpretability of OLS and ridge. Hence, we had to lean on the XGBoost feature and SHAP values instead of just finding the important feature through the XGBoost model. The need for these tools itself is information, as it confirms that important drivers of performance do not act independently but interact with context. For example, lunch status interacts with school and class type largely with the score for both reading and math.

From an econometric point of view, the takeaway is that a purely linear specification is too restrictive, but a neural network is unnecessary, making it too complicated for a smaller dataset like ours, and is unstable for this dataset. Gradient boosting with careful regularisation will achieve the middle ground where the dataset is flexible enough to model interaction and heterogeneity but structured enough to be interpretable using modern explainability tools.

This justifies our modelling choice and frames the rest of the discussion around feature importance and robustness.

8.2 Drivers of Student Performance

Across both XGBoost importance scores and SHAP results, several consistent predictors emerged.

Socioeconomic Status

Lunch status was the strongest and most consistent predictor of both reading and math scores. SES is a well-established determinant of early academic performance, and our findings align with both Project STAR research and contemporary education literature. Importantly, SHAP showed that the direction of this effect varied depending on interaction with school context, which highlights that SES does not operate in isolation.

School-Level Effect

School ID and school location of whether they're located in rural, suburban or urban areas were consistently among the top drivers. This implies that unobserved school characteristics which were not observed through the given data may have been determining factors for students' outcomes. An example of these unobserved school characteristics could be resources,extracurricular curriculums, infrastructures, etc. These school effects persisted even after adjusting for SE and demographics which showed the structural inequalities across educational environments.

Classroom Structures

Class-type variables, particularly small classes, were meaningful predictors of performance. This aligns with prior evidence demonstrating that smaller class sizes improve early learning outcomes through increased individualised attention and reduced behavioural strains.

Teacher Factors

Teacher experience had consistent positive SHAP contributions particularly for math. This suggests that more experienced teachers may be more effective instructional strategies and giving better classroom management approaches. Teacher ethnicity and qualification only had moderate effects indicating that teacher and student demographic matching or diversity might influence small engagement relevance.

Demographic Factors of Students

The student's gender or ethnicity contributed to both outcomes, with gender having a stronger effect in reading. To ensure validity we also compared XGBoost finding with OLS and regularised model variables such as SES, `school_id`, experience and class type remained influential across linear coefficients, XGBoost gain and SHAP contributions. The findings suggest that our identified drivers reflect relationships of variables rather than the model's specifications.

8.3 Limitations

Some limitations we acknowledged through these projects that would be influential, such as parental involvement, home environment, peer influence and student motivation, were absent from the dataset. The dataset consists of predictors with imbalance and missing values, which limits its reliability. Given these limitations, we also had some assumptions and modelling constraints.

Our MSE assumes systematics error costs, although under and over-prediction may not be equivalent in educational settings. The modelling strategy is predictive rather than causal, hence, no model can determine whether a certain feature causes a higher test score than others. The correlation patterns also indicate clustering across schools and demographic groups, meaning residual confounding is unavoidable.

Hence, XGBoost's strong performance confirms that these influences combine through non-linear, interactive, and context-dependent effects rather than simple linear relationships.

8.4 Random Assignment

Random assignment was beneficial to this study as students from different classes, backgrounds, and achievement levels were randomly selected. This process helps eliminate selection bias that may arise from teacher preference or performance-based sorting, which strengthens the internal validity of our analysis.

Because the assignment process is not influenced by prior academic performance, the resulting student groups are more comparable. This reduces the likelihood that observed associations between variables and outcomes are driven by underlying structural or demographic differences across classrooms. Random assignment ensures that:

- Scores are not artificially inflated or deflated by classroom sorting,
- Teacher effects are distributed more evenly across the dataset, and
- The model captures genuine student-level variation rather than school- or classroom-level clustering.

This project aimed to understand which student, teacher and school-level factors reflect on the scores of kindergarten students' reading, mathematics scores and to evaluate which predictive modelling approach most effectively captures these relationships. The results highlight several important results around interpretability, model complexity, data limitation and broader implications for early childhood education.

9. Conclusion

This project set out to identify the key drivers of early reading and mathematics performance and to determine the most effective predictive model for explaining students' achievement. Through our workflow, ranging from data cleaning, exploratory analysis, to the evaluation of modelling approaches, we found that XGBoost provides the best predictive performance both separate read MSE score of 467.4503 and mathematical MSE score of 1513.2082 and with the lowest combined MSE of 990.33 .

Our finding shows that socioeconomic status, school environment and classroom structure are the strongest determinants for both reading and mathematics scores. In particular, the lunch status of `free-lunch` emerged as the single most influential variable. Students receiving free lunch scored higher in models not because of low SES improved outcome but also because they were placed in `small-classes` and `high-performing-school`. This highlights how structural interaction, such as a target class size reduction can offset socioeconomic barriers.

Another factor was the `school,` where `rural` schools showed a stronger reading performance while `suburban` schools were more predictive for math. This suggests that the learning environment interacts differently with reading and mathematical developments.

At the classroom level, `small-class` types had clear benefits for both subjects, strengthening the evidence that reduced student-teacher ratios enhance early achievements. Teacher's `experience` further contributed positively, especially for mathematics, indicating that numeracy skills may depend more on family experience.
For demographics, the `caucasian ethnicity` and `male` gender had slight advantages in mathematics and disadvantages in reading.

Overall, our results support the view that early academic performance is not driven by any single factor but emerges from interactions between socioeconomic background, school-level quality and classroom structure. The study reinforces well - established evidence from Project STAR - school assignment, class size and teachers, characteristics materially shaping children's academic trajectories and these structures can either magnitude or amplify demographic inequalities.

Future work should incorporate a broader dataset with richer household behaviours and use causal designs to isolate the treatment effect more precisely. However, this analysis provides a detailed explanation of the evidence of factors and variables that affect kindergarten students' learning ability of reading and mathematics.

Acknowledgement of AI Assistance

Some parts of the code development process involved limited assistance from AI for support. Particularly, AI assistance was used to clarify the implementation details of models that were not covered in the course. It includes XGBoost, LightGBM, and methods for computing feature importance. Assistance was used to improve understanding of certain programming concepts and help debug code.

References

References: including referencing AI and statement on AI use. See https://www.unsw.edu.au/student/managing-your-studies/academic-skills-support/ toolkit/referencing/ai-referencing

1. Analytics Vidhya. (2018) *An end-to-end guide to understand the math behind XGBoost.* Available at: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/ (Accessed 10 November 2025)

2. DataCamp (n.d.) Ordinary Least Squares (OLS) Regression in Python: A Complete Tutorial. Available at: https://www.datacamp.com/tutorial/ols-regression (Accessed 5 November 2025).

3. DataCamp (n.d.) Introduction to SHAP Values for Machine Learning Interpretability. Available at: https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability (Accessed 8 November 2025).

4. Domino Data Lab (2020) SHAP and LIME: A Comparison of Explainability Techniques in Python. Available at: https://domino.ai/blog/shap-lime-python-libraries-part-1-great-explainers-pros-cons (Accessed 8 November 2025).

5. GeeksforGeeks. (n.d.) *LightGBM - Light Gradient Boosting Machine.* Available at: https://www.geeksforgeeks.org/machine-learning/lightgbm-light-gradient-boosting-machine/ (Accessed 8 November 2025)

6. IBM. (n.d.) *What is a neural network?* Available at: https://www.ibm.com/think/topics/neural-networks (Accessed: 5 November 2025).

7. IBM. (n.d) *What are decision trees?* Available at: https://www.ibm.com/think/topics/decision-trees (Accessed 6 November 2025)

8. IBM. (n.d) *What is a random forest?* Available at: https://www.ibm.com/think/topics/random-forest (Accessed 10 November 2025)

9. LightGBM Developers. (n.d) *LGMClassifier - LightGBM documentation*. Available at: https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html (Accessed 8 November 2025)

10. Machine Learning No Jutsu (2019) Linear Regression and Ordinary Least Squares. Medium. Available at: https://medium.com/@machinelearningnojutsu/linear-regression-ordinary-least-square-cabe0d97117 (Accessed 5 November 2025).

11. Muller, J. (n.d.) *Random Forest Algorithm explained.* BuiltIn. Available at: https://builtin.com/data-science/random-forest-algorithm (Accessed 10 November 2025)

12. SHAP Documentation (n.d.) An Introduction to Explainable AI with Shapley Values. Available at: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html (Accessed 8 November 2025).

13. Turkish Technology. (2020) *LightGBM: Light and powerful gradient boost algorithm.* Medium. Available at:

https://medium.com/@turkishtechnology/light-gbm-light-and-powerful-gradient-boost-algorithm-eaa1e804eca8 (Accesed 8 November 2025)

14. UCLA Institute for Digital Research and Education (n.d.) Linear Regression: R Data Analysis Examples. Available at: https://stats.oarc.ucla.edu/r/dae/linear-regression-r-data-analysis-examples/ (Accessed 10 November 2025).

15. XGBoost Developers. (n.d.) *XGBoost model tutorial.* Available at: https://xgboost.readthedocs.io/en/stable/tutorials/model.html (Accessed 9 November 2025)

16. XGBoost Developers. (n.d.) *XGBoost Documentation*. Available at: https://xgboost.readthedocs.io/en/stable/ (Accessed: 9 November 2025).

Appendix

3.3 Distributions of Predictor Variables

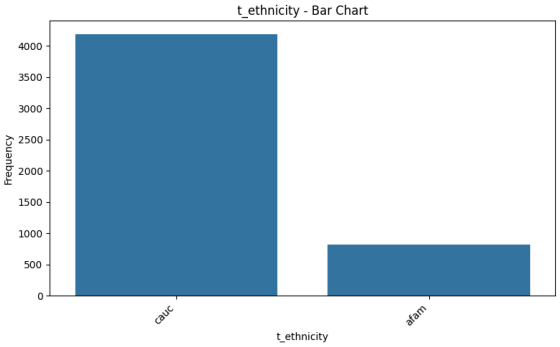


Figure 4: `t_ethnity` Bar Graph

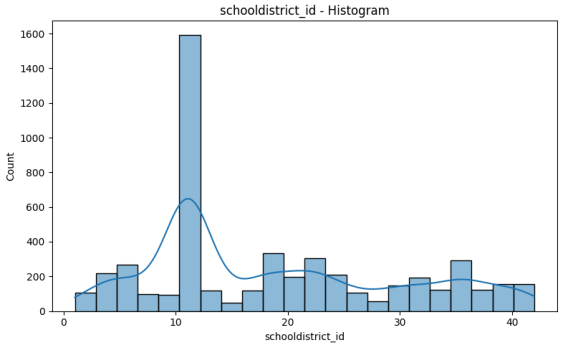


Figure 5: `schooldistrict_id` Histogram

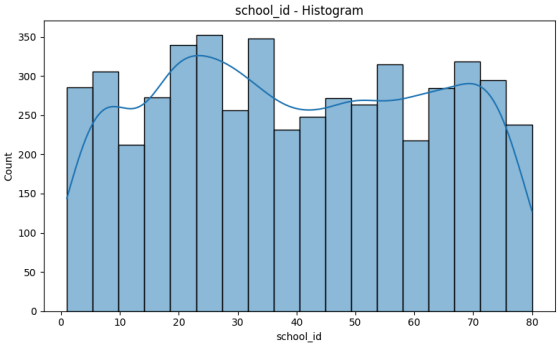


Figure 6: `school_id` Histogram

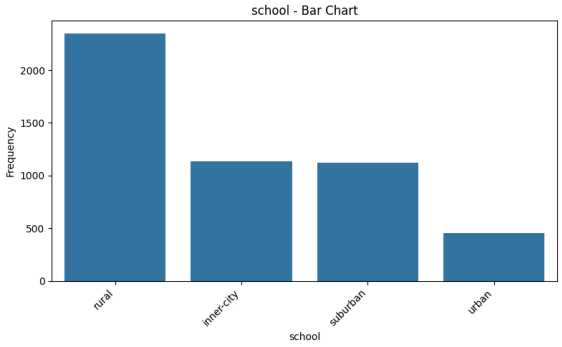


Figure 7: `school` Bar Graph

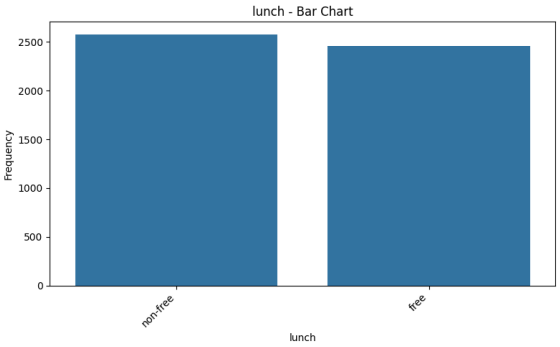


Figure 8: `lunch` Bar Graph

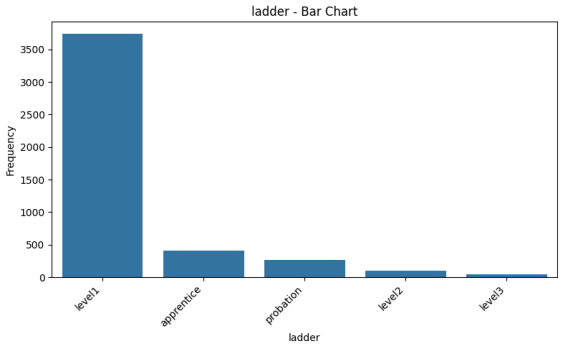


Figure 9: `ladder` Bar Graph

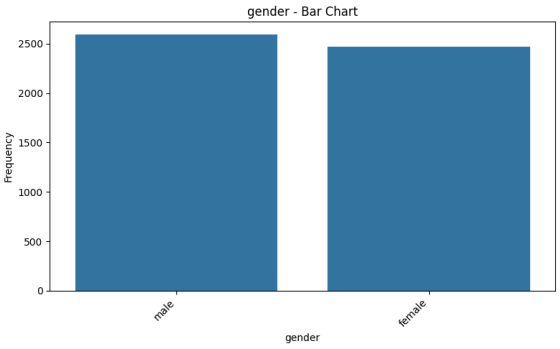


Figure 10: `Gender` Bar Graph

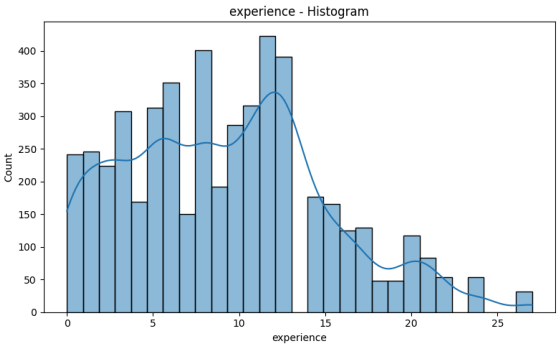


Figure 11: `experience` histogram

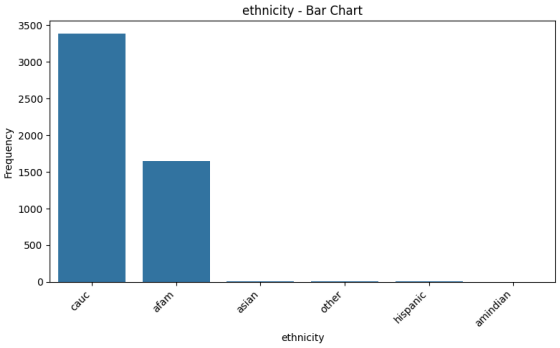


Figure 12: `ethnicity` Bar Chart

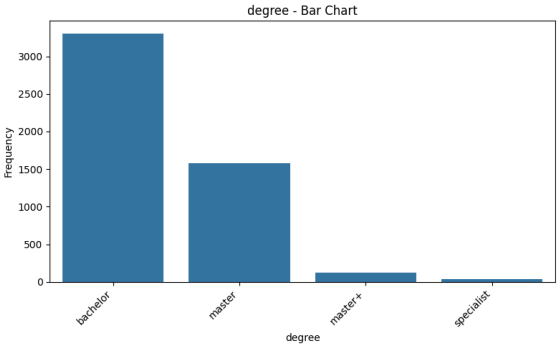


Figure 13: `degree` Bar Chart

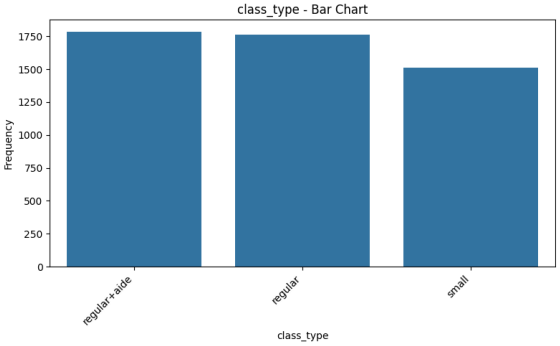


Figure 14: `class_type` Bar Chart

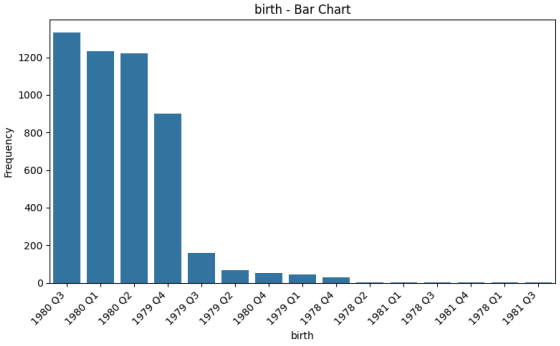


Figure 15: `birth` Bar Chart

3.4 Correlation Analysis

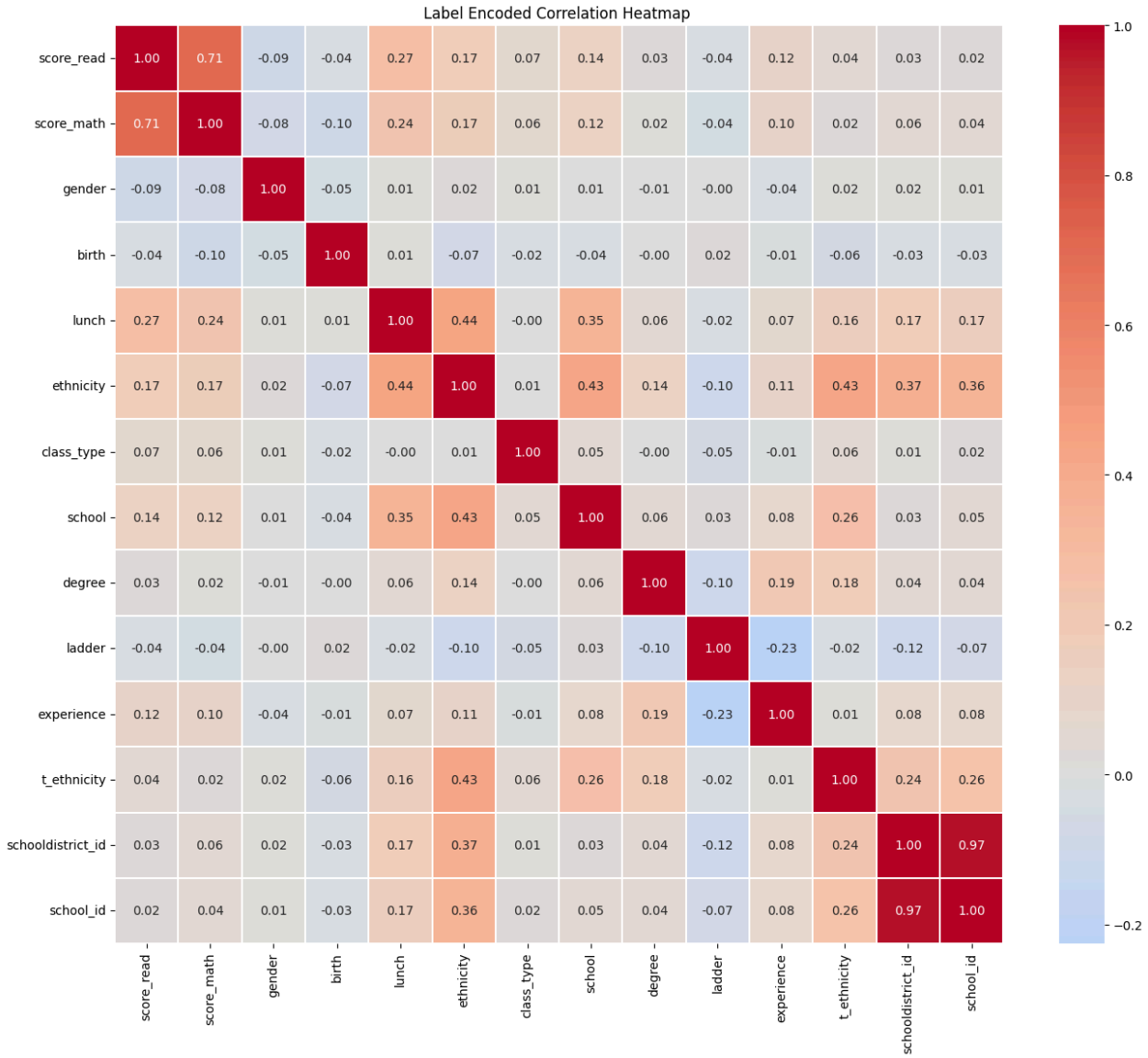


Figure 16: Variables Correlation Heatmap

3.5 Outlier Detection:

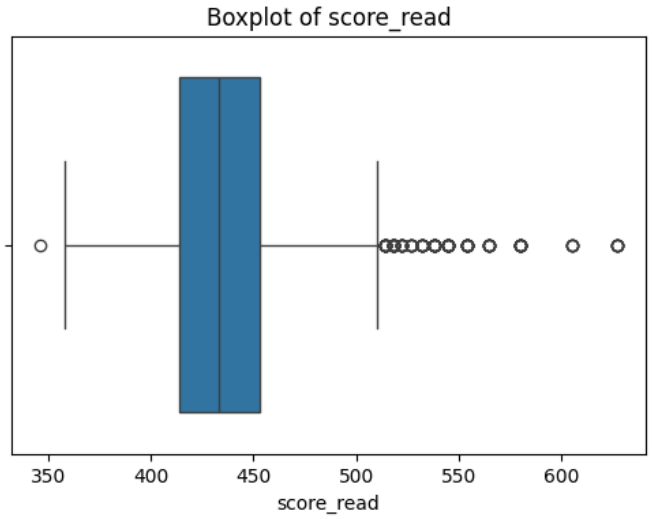


Figure 19: Boxplot of Outliers for `score_read`

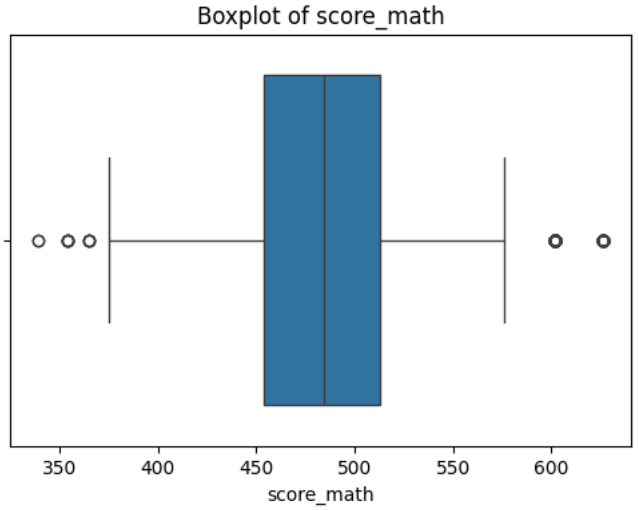


Figure 20: Boxplot of Outliers for `score_math`

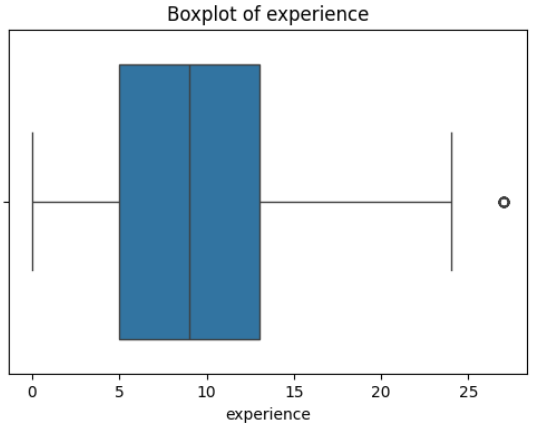
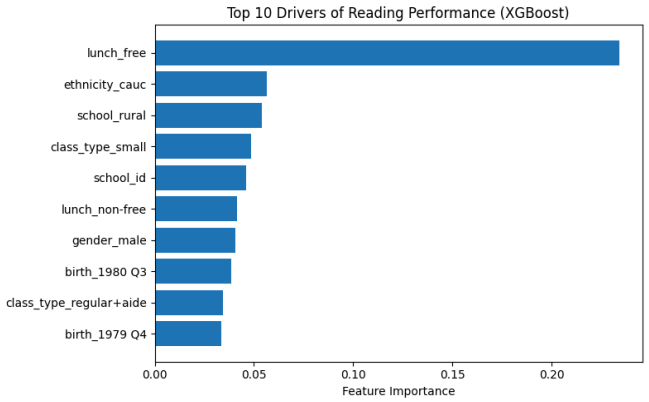


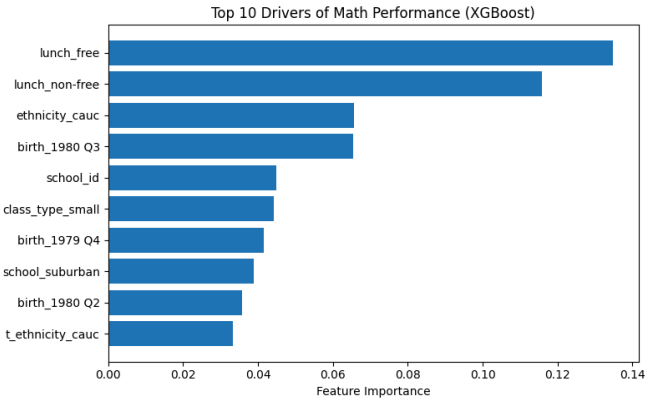
Figure 21: Boxplot of Outliers for `experience`

7.1 Method - XGBoost Feature Importance and Comparisons

7.1 XGBoost Feature Importance and Comparisons

XGBoost feature importance findings:





Drivers for Reading Performance (using feature_importances_):	**Drivers for Math Performance (using feature_importances_):**
feature importance	feature importance
lunch_free 0.234177	lunch_free 0.134849
ethnicity_cauc 0.056597	lunch_non-free 0.115900
school_rural 0.053816	ethnicity_cauc 0.065724
class_type_small 0.048660	birth_1980 Q3 0.065418
school_id 0.046046	school_id 0.044873
lunch_non-free 0.041492	class_type_small 0.044253
gender_male 0.040642	birth_1979 Q4 0.041583
birth_1980 Q3 0.038670	school_suburban 0.038829
class_type_regular+aide 0.034338	birth_1980 Q2 0.035689
birth_1979 Q4 0.033602	t_ethnicity_cauc 0.033282
birth_1979 Q2 0.033112	t_ethnicity_afam 0.033088
experience 0.032971	experience 0.033046
school_suburban 0.032315	gender_male 0.031083
school_urban 0.030634	birth_1979 Q3 0.028938
degree_master 0.027961	birth_1980 Q4 0.027028
t_ethnicity_afam 0.026199	school_urban 0.026464
birth_1978 Q4 0.023513	degree_master 0.026369
birth_1980 Q1 0.020082	birth_1980 Q1 0.025697
degree_other 0.018922	school_rural 0.024551
ethnicity_other 0.018431	class_type_regular+aide 0.023845
birth_1979 Q3 0.017736	degree_other 0.019312
birth_1980 Q2 0.016948	birth_1978 Q4 0.018414
birth_1979 Q1 0.016133	birth_1979 Q2 0.017583
t_ethnicity_cauc 0.015024	birth_1978 Q2 0.013353
birth_1980 Q4 0.013630	birth_1979 Q1 0.011737
birth_1981 Q1 0.010304	ethnicity_other 0.011659
birth_Unknown 0.008704	birth_1978 Q3 0.007434
birth_1978 Q2 0.004899	birth_1981 Q1 0.000000
birth_1981 Q4 0.004440	birth_1981 Q4 0.000000
birth_1978 Q3 0.000000	birth_Unknown 0.000000
birth_1981 Q3 0.000000	birth_1981 Q3 0.000000

SHAP feature importance findings: XGBoost

Top Grouped SHAP Drivers – READING:

	feature	mean_shap	mean_abs_shap
8	school_id	-0.28657603	6.6712685
0	lunch	0.09845901	2.8081238
7	experience	0.049035057	2.6800334
9	gender_male	-0.001838262	2.2417858
1	ethnicity	0.110081695	1.3481685
3	class_type	0.003209642	1.3144138
6	school_location	-0.024028355	0.6264678
2	t_ethnicity	-0.05731968	0.2941521
5	degree	0.0035203758	0.26233467
4	birth_quarter	-0.0005686489	0.24888757

Top Grouped SHAP Drivers – MATH:

	feature	mean_shap	mean_abs_shap
8	school_id	-0.49595	11.196949
0	lunch	0.17491703	4.233727
7	experience	-0.06821132	3.8971775
1	ethnicity	0.22468853	3.233497
9	gender_male	-0.0074543073	2.6745749
3	class_type	0.0015142288	2.366667
6	school_location	0.06247865	0.92036664
4	birth_quarter	-0.0032662156	0.6383791
5	degree	-0.04551419	0.60170186
2	t_ethnicity	-0.12992962	0.36054426

Top 20 SHAP importances – READING:

	feature	mean_shap
20	onehotencoder__ethnicity_cauc	0.21331187
18	onehotencoder__lunch_free	0.19955121
0	standardscaler__experience	0.049034994
25	onehotencoder__school_suburban	0.0077928584
23	onehotencoder__class_type_small	0.007330002
9	onehotencoder__birth_1979 Q4	0.00710012
21	onehotencoder__ethnicity_other	0.0068515986
28	onehotencoder__degree_other	0.0044850935
17	onehotencoder__birth_Unknown	0.0032564956
8	onehotencoder__birth_1979 Q3	0.0031557442
27	onehotencoder__degree_master	0.0025556616
11	onehotencoder__birth_1980 Q2	0.001298086
6	onehotencoder__birth_1979 Q1	0.0012139428
3	onehotencoder__birth_1978 Q2	0.00027519144
4	onehotencoder__birth_1978 Q3	0.0
15	onehotencoder__birth_1981 Q3	0.0
16	onehotencoder__birth_1981 Q4	-8.337639e-05
22	onehotencoder__class_type_regular+	-0.00091075135
13	onehotencoder__birth_1980 Q4	-0.0011667985
2	onehotencoder__gender_male	-0.0018382337
14	onehotencoder__birth_1981 Q1	-0.0018893775
5	onehotencoder__birth_1978 Q4	-0.0025786334
19	onehotencoder__lunch_non-free	-0.0026332652
10	onehotencoder__birth_1980 Q1	-0.004647708
12	onehotencoder__birth_1980 Q3	-0.0063664494

Top 20 SHAP importances – MATH:

	feature	mean_shap
20	onehotencoder__ethnicity_cauc	0.21331187
18	onehotencoder__lunch_free	0.19955121
0	standardscaler__experience	0.049034994
25	onehotencoder__school_suburban	0.0077928584
23	onehotencoder__class_type_small	0.007330002
9	onehotencoder__birth_1979 Q4	0.00710012
21	onehotencoder__ethnicity_other	0.0068515986
28	onehotencoder__degree_other	0.0044850935
17	onehotencoder__birth_Unknown	0.0032564956
8	onehotencoder__birth_1979 Q3	0.0031557442
27	onehotencoder__degree_master	0.0025556616
11	onehotencoder__birth_1980 Q2	0.001298086
6	onehotencoder__birth_1979 Q1	0.0012139428
3	onehotencoder__birth_1978 Q2	0.00027519144
4	onehotencoder__birth_1978 Q3	0.0
15	onehotencoder__birth_1981 Q3	0.0
16	onehotencoder__birth_1981 Q4	-8.337639e-05
22	onehotencoder__class_type_regular+	-0.00091075135
13	onehotencoder__birth_1980 Q4	-0.0011667985
2	onehotencoder__gender_male	-0.0018382337
14	onehotencoder__birth_1981 Q1	-0.0018893775
5	onehotencoder__birth_1978 Q4	-0.0025786334
19	onehotencoder__lunch_non-free	-0.0026332652
10	onehotencoder__birth_1980 Q1	-0.004647708
12	onehotencoder__birth_1980 Q3	-0.0063664494

Feature importance of linear models:

Coefficients for OLS:

Top 10 Feature Importances – OLS_READ:

	feature	coef	abs_coef
14	birth_1981 Q1	50.168529	50.168529
5	birth_1978 Q4	-23.190509	23.190509
16	birth_1981 Q4	22.254661	22.254661
18	lunch_free	-21.855164	21.855164
3	birth_1978 Q2	-17.945609	17.945609
7	birth_1979 Q2	-17.227762	17.227762
29	t_ethnicity_afam	12.205296	12.205296
6	birth_1979 Q1	-11.791650	11.791650
12	birth_1980 Q3	-11.257415	11.257415
19	lunch_non-free	-11.020172	11.020172

Top 10 Feature Importances – OLS_MATH:

	feature	coef	abs_coef
4	birth_1978 Q3	65.963994	65.963994
15	birth_1981 Q3	49.134613	49.134613
16	birth_1981 Q4	46.037092	46.037092
3	birth_1978 Q2	-44.853778	44.853778
9	birth_1979 Q4	35.804262	35.804262
13	birth_1980 Q4	33.221290	33.221290
10	birth_1980 Q1	31.178940	31.178940
8	birth_1979 Q3	27.524925	27.524925
18	lunch_free	-26.705916	26.705916
11	birth_1980 Q2	22.821107	22.821107

Coefficients for Lasso:

Top 10 Feature Importances – LASSO_READ:

	feature	coef	abs_coef
5	birth_1978 Q4	-11.252605	11.252605
18	lunch_free	-11.091005	11.091005
7	birth_1979 Q2	-7.846880	7.846880
20	ethnicity_cauc	7.127103	7.127103
23	class_type_small	5.549912	5.549912
29	t_ethnicity_afam	4.801666	4.801666
2	gender_male	-4.630284	4.630284
9	birth_1979 Q4	4.256628	4.256628
30	t_ethnicity_cauc	3.719046	3.719046
12	birth_1980 Q3	-3.161511	3.161511

Top 10 Feature Importances – LASSO_MATH:

	feature	coef	abs_coef
18	lunch_free	-13.854188	13.854188
20	ethnicity_cauc	12.022484	12.022484
9	birth_1979 Q4	11.327636	11.327636
23	class_type_small	8.269577	8.269577
12	birth_1980 Q3	-7.734846	7.734846
10	birth_1980 Q1	6.906501	6.906501
2	gender_male	-5.820128	5.820128
29	t_ethnicity_afam	4.739906	4.739906
0	experience	2.823014	2.823014
25	school_suburban	2.706501	2.706501

Coefficients for Ridge:

Top 10 Feature Importances – RIDGE_READ:

	feature	coef	abs_coef
5	birth_1978 Q4	-9.433929	9.433929
18	lunch_free	-9.396829	9.396829
7	birth_1979 Q2	-7.338437	7.338437
20	ethnicity_cauc	7.124884	7.124884
23	class_type_small	5.522892	5.522892
14	birth_1981 Q1	5.289305	5.289305
29	t_ethnicity_afam	5.253211	5.253211
9	birth_1979 Q4	4.743896	4.743896
2	gender_male	-4.640722	4.640722
30	t_ethnicity_cauc	4.127267	4.127267

Top 10 Feature Importances – RIDGE_MATH:

	feature	coef	abs_coef
20	ethnicity_cauc	11.487581	11.487581
9	birth_1979 Q4	10.477096	10.477096
12	birth_1980 Q3	-8.537310	8.537310
18	lunch_free	-8.038589	8.038589
23	class_type_small	8.023956	8.023956
19	lunch_non-free	6.498866	6.498866
10	birth_1980 Q1	6.304974	6.304974
2	gender_male	-6.042137	6.042137
29	t_ethnicity_afam	5.443392	5.443392
5	birth_1978 Q4	-5.341915	5.341915

Coefficients for Elastic Net:

Top 10 Feature Importances – ELASTIC NET_READ:

	feature	coef	abs_coef
5	birth_1978 Q4	-11.252605	11.252605
18	lunch_free	-11.091005	11.091005
7	birth_1979 Q2	-7.846880	7.846880
20	ethnicity_cauc	7.127103	7.127103
23	class_type_small	5.549912	5.549912
29	t_ethnicity_afam	4.801666	4.801666
2	gender_male	-4.630284	4.630284
9	birth_1979 Q4	4.256628	4.256628
30	t_ethnicity_cauc	3.719046	3.719046
12	birth_1980 Q3	-3.161511	3.161511

Top 10 Feature Importances – ELASTIC NET_MATH:

	feature	coef	abs_coef
20	ethnicity_cauc	11.655194	11.655194
9	birth_1979 Q4	10.417476	10.417476
18	lunch_free	-8.383010	8.383010
12	birth_1980 Q3	-8.260278	8.260278
23	class_type_small	7.949987	7.949987
10	birth_1980 Q1	6.211627	6.211627
19	lunch_non-free	6.179271	6.179271
2	gender_male	-5.819069	5.819069
29	t_ethnicity_afam	4.470262	4.470262
0	experience	2.866231	2.866231

Feature importance for Random Forest:

Top 10 Drivers for Reading Performance (Random Forest):

	feature	importance
1	school_id	0.217326
0	experience	0.177020
18	lunch_free	0.087506
19	lunch_non-free	0.083370
20	ethnicity_cauc	0.052923
2	gender_male	0.045745
23	class_type_small	0.040517
27	degree_master	0.031015
22	class_type_regular+aide	0.028294
25	school_suburban	0.027069

Top 10 Drivers for Math Performance (Random Forest):

	feature	importance
1	school_id	0.214867
0	experience	0.182169
19	lunch_non-free	0.068277
18	lunch_free	0.064918
20	ethnicity_cauc	0.060312
2	gender_male	0.044754
12	birth_1980 Q3	0.037700
23	class_type_small	0.036551
9	birth_1979 Q4	0.033895
27	degree_master	0.033464

Feature importance for LightGBM:

Top 10 Drivers for Reading Performance (LightGBM):

	feature	importance
1	school_id	5408
0	experience	2762
2	gender_male	781
23	class_type_small	568
22	class_type_regular+aide	561
18	lunch_free	532
12	birth_1980 Q3	513
27	degree_master	485
11	birth_1980 Q2	451
20	ethnicity_cauc	380

Top 10 Drivers for Math Performance (LightGBM):

	feature	importance
1	school_id	5025
0	experience	2921
2	gender_male	884
22	class_type_regular+aide	624
23	class_type_small	545
27	degree_master	511
18	lunch_free	499
11	birth_1980 Q2	469
12	birth_1980 Q3	465
10	birth_1980 Q1	415