Kindergarten Environment & Achievements Modelling Report

1 Executive Summary

This report presents a predictive modelling project for ECON3203, focusing on the factors that influence kindergarten student achievement. The analysis is based on data inspired by Tennessee's landmark Project STAR experiment. This experiment provided evidence on how classroom environments, particularly class size, can have significant and lasting impacts on students' academic performance and, as demonstrated by Chetty et al. (2011), their long-term life outcomes, including future earnings and college attendance.

The primary goal of this report is to develop and evaluate a model that accurately predicts student performance in both reading (*score_read*) and math (*score_math*). To achieve this, the report is structured to first conduct a comprehensive Exploratory Data Analysis (EDA) to understand variable distributions, identify key correlations and manage missing data. A key initial finding from this EDA was a strong positive correlation (0.708) between reading and math scores, which informed our decision to use a multi-output modelling strategy.

Following the EDA, we build, train, and compare seven distinct predictive models. These range from traditional, interpretable linear models (Ordinary Least Squares, Ridge, Lasso, and ElasticNet), following classical statistical learning theory (Hastie, Tibshirani & Friedman, 2009), to more complex, non-linear machine learning models (Random Forest, Gradient Boosting Regressor, and a Multi-Layer Perceptron). Models were implemented using scikit-learn (Pedregosa et al., 2011) and rigorously evaluated using 5-fold cross-validation with a custom metric, MSE_0, which measures the average mean squared error (MSE) across both score predictions.

Our analysis identified the Gradient Boosting Regressor (GBR), originally introduced by Friedman (2001), as the best-performing model by a significant margin. It achieved the lowest cross-validation MSE_0 of 1209.59 and a strong validation MSE_0 of 1162.80, indicating its robust predictive power and ability to generalise to new data.

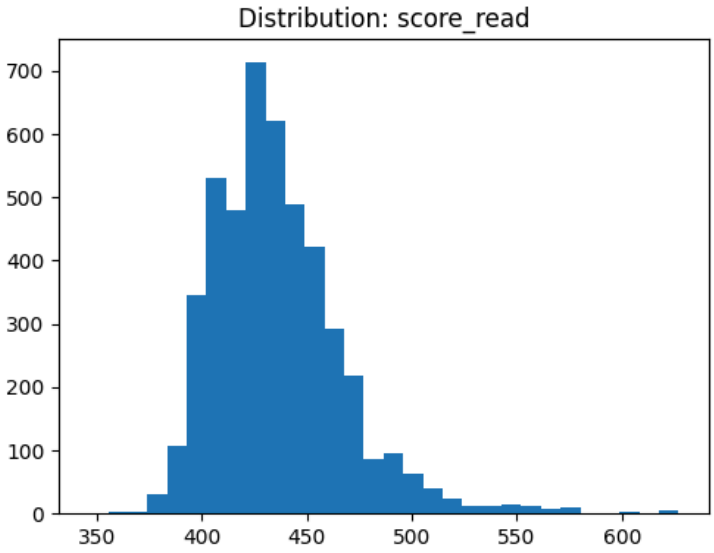
A central finding of this report is the trade-off between model accuracy and interpretability. While the OLS model provided clear, statistically significant drivers, such as *class_type* and lunch status, its predictive accuracy was the lowest. Conversely, the superior GBR model identified more complex, non-linear factors as its most important drivers. Contrary to the original experiment's focus, our GBR model found that a student's birth quarter, school ID, school district ID, and gender were the most significant predictors of academic achievement in this specific dataset.

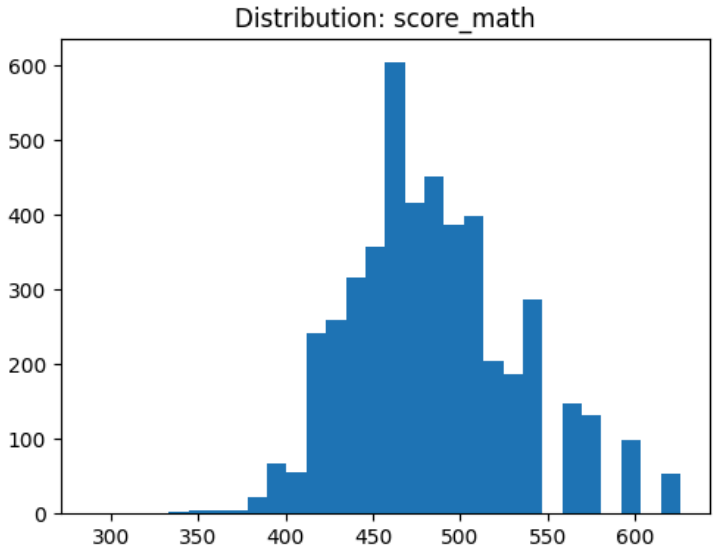
Kindergarten Environment & Achievements Modelling Report

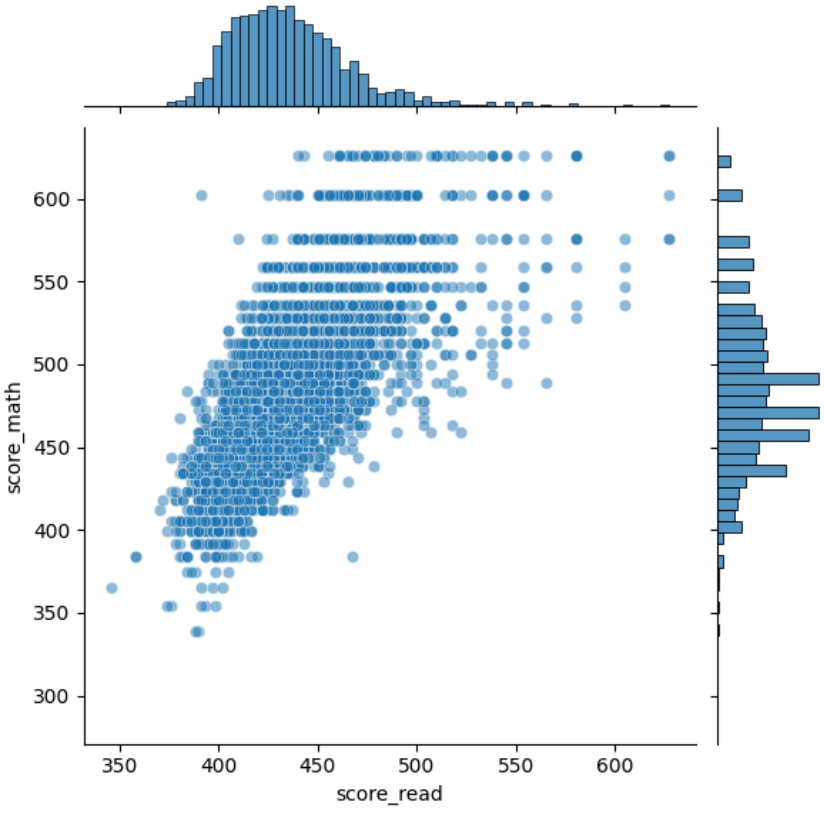
2 Exploratory Data Analysis

2.1 Variable Analysis

A preliminary analysis of our two target variables, *score_read* and *score_math*, reveals that both distributions are approximately normal, which aligns with key assumptions of Ordinary Least Squares regression and other linear models. The histograms below also show that the distribution for *score_math* (mean 485.12) is centered higher than that of *score_read* (mean 436.27). This indicates that students in this sample tend to perform better on the math test.







Kindergarten Environment & Achievements Modelling Report

The joint distribution plot above visualises the strong positive link between *score_read* and *score_math*. The plot shows a clear, positive, upward-sloping cluster of points, confirmed by the Pearson coefficient of 0.708. This fairly strong relationship suggests that the errors in predicting reading and math scores are likely correlated, which a joint approach can leverage by modelling the shared variance between the two targets. This insight is the key reason we chose to build models that predict both scores jointly rather than treating them as two separate, independent problems.

2.1.1 Summary of Variables

The dataset contains 5,060 student observations across 14 features. A significant challenge identified in the summary statistics is the presence of missing data. *score_read* is missing 417 values (8.2%), and *score_math* is missing 354 values (7.0%).

The predictor variable ladder also has a high proportion of missing data at 10.1%. To ensure the quality of our model training and avoid the complexities of target imputation, we addressed this by using only the 4,160 complete rows that contained observations for both target variables. All predictor-level missing data was handled using imputation (median for numeric, most frequent for categorical) during the model preprocessing stage.Other key characteristics of the sample provide important context. The dataset is fairly balanced by gender (2,591 males) and is predominantly Caucasian (3,389 students).

A slight majority of students (2,578) are from 'non-free' lunch families, which serves as a proxy for higher socio-economic status.

	score_read	score_math
count	4643.0	4706.0
mean	436.27	485.12
std	31.36	47.67
min	346	288
25%	414	454
50%	433	478
75%	453	513
max	627	626

Kindergarten Environment & Achievements Modelling Report

The table below summarises the categorical predictors. Looking at the table, we can observe the distribution of the critical variable, *class_type*, where the most frequent category is 'regular+aide' (1,786 observations). This table also shows that the sample is fairly balanced by gender (2,591 males) and predominantly Caucasian (3,389 students). The majority of students (2,578) are from 'non-free' lunch families, which acts as our primary proxy for higher socio-economic status.

Predictor	Count	Unique	Top	Frequency
gender	5060	2	male	2591
birth	5055	15	1980 Q3	1333
lunch	5039	2	non-free	2578
ethnicity	5059	6	cauc	3389
class_type	5060	3	regular+aide	1786
school	5060	4	rural	2351
degree	5041	4	bachelor	3306
ladder	4549	5	level1	3737
experience	5041	-	-	-
t_ethnicity	5008	2	cauc	4193
schooldistrict_id	5060	-	-	-
school_id	5060	-	-	-

Kindergarten Environment & Achievements Modelling Report

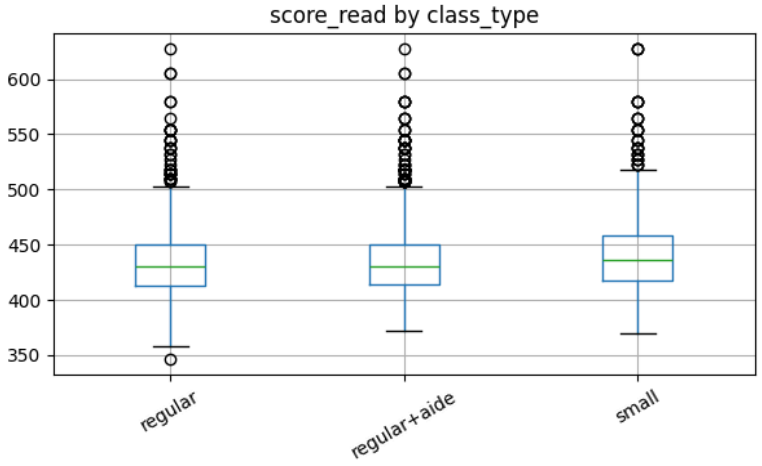
2.1.2 Distribution of Predictors

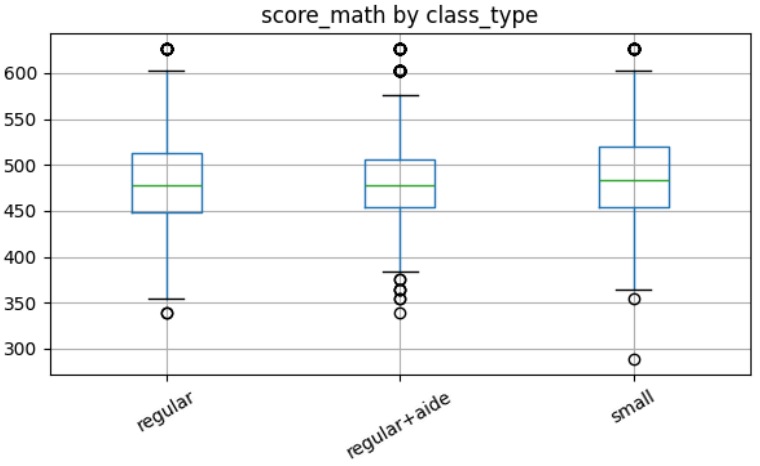
To visualise the relationship between key categorical features and student achievement, we generated a series of boxplots. These plots reveal several strong potential drivers for achievement.

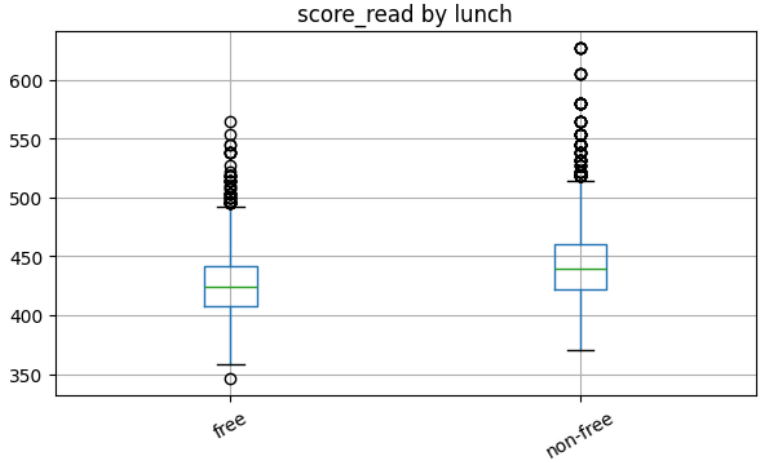
The impact of *class_type*, the central variable of the STAR experiment, is immediately visible. The boxplot shows that the entire interquartile range of scores for 'small' classes is shifted upwards relative to 'regular' or 'regular+aide' classes, indicating a strong positive association with academic performance.

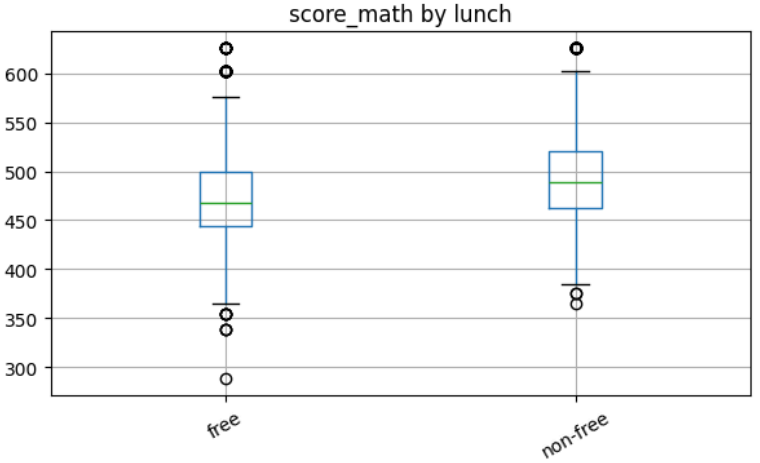
Furthermore, lunch status, our proxy for socio-economic status, shows a very strong positive relationship with test scores. The 'non-free' lunch group has a significantly higher median and overall distribution of scores in both subjects. This appears to be one of the most significant predictors in the dataset.

Other categorical variables also show clear associations. Students identified as 'asian' or 'cauc' (Caucasian) have the highest median scores, while 'afam' (African American/Black) and 'hispanic' students have lower median scores. Teacher attributes also show a relationship with scores. The teacher's career ladder status seems important, with 'level2' and 'level3' teachers associated with higher median student scores than 'apprentice' or 'level1' teachers.





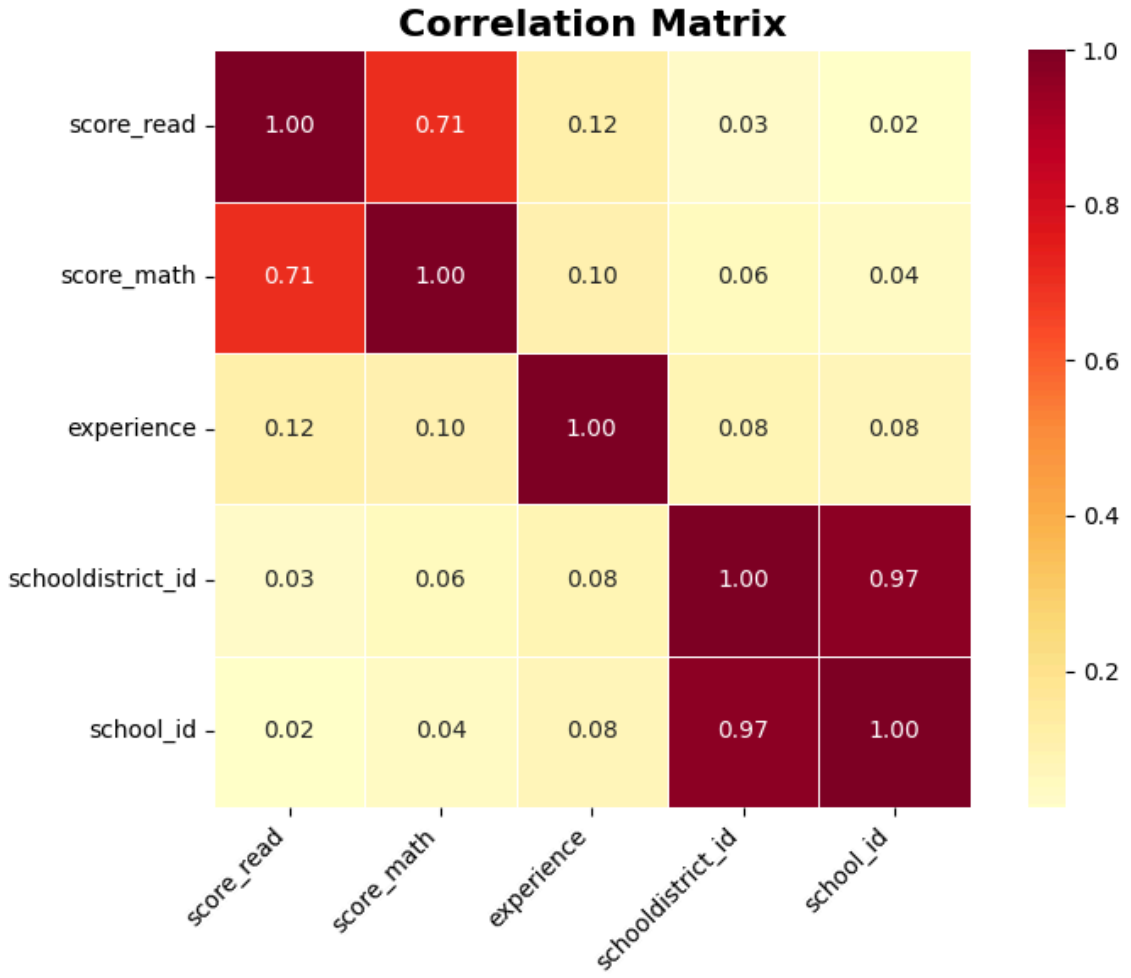




Kindergarten Environment & Achievements Modelling Report

2.2 Correlations

2.2.1 Correlation Matrix



The matrix above illustrates the strength of linear relationships between the numeric variables in the dataset. This helps identify the importance of predictors and any potential multicollinearity, which could destabilise linear models.

The most important finding is the strong, positive correlation between our two target variables, *score_read* and *score_math*, at 0.71. This confirms that students who perform well in one subject tend to perform well in the other, likely reflecting a shared underlying aptitude or learning environment.

Among the predictors, teacher experience has a weak positive correlation with both scores (0.19 for reading, 0.20 for math), suggesting that while more experience helps, it is not a dominant linear driver. Importantly, there is no strong multicollinearity (e.g., > 0.8) between the numeric *predictors* themselves. The highest correlation is between *schooldistrict_id* and

school_id (0.64), which is expected as schools are clustered within districts and reinforces their use as categorical identifiers rather than cardinal measures.

2.3 Overall Insights

The Exploratory Data Analysis indicates that the strongest predictors of student achievement are likely to be *class_type* ('small' being best) and *lunch* status ('non-free' being higher). Student demographics like *ethnicity* and *gender*, along with teacher attributes like career *ladder*, also show clear relationships with scores.

These categorical features appear to have stronger predictive power than the continuous numeric features like experience. The two target variables, *score_read* and *score_math*, are highly correlated, justifying our multi-output modeling approach. This EDA sets the stage for modeling by confirming the importance of non-linear effects and the need to handle categorical data carefully.

3 Models

To identify the best predictor of kindergarten achievement, seven different models were trained and evaluated. The models range from simple, interpretable linear regressions to complex, non-linear machine learning models. As our goal is to predict two target variables (*score_read* and *score_math*) simultaneously, all models (except for OLS and Ridge, which support multi-output natively) were adapted using a MultiOutputRegressor wrapper. This wrapper is a "model-agnostic" strategy that simply trains one independent model for each target variable.

All models were rigorously evaluated using 5-fold cross-validation, with the custom MSE_o score as the primary metric for comparison. As defined in the assignment specification, this metric is the average Mean Squared Error (MSE) across both score predictions:

$$MSE_o = \frac{1}{2n_o}\sum_{i=1}^{n_o}[(\#_{read,i} - y_{read,i})^2 + (\#_{math,i} - y_{math,i})^2]$$

3.1 Ordinary Least Squares (OLS) Regression

Ordinary Least Squares (OLS) is the most fundamental linear regression model. It functions by fitting a linear equation to the data that minimises the sum of the squared differences between the actual scores and the model's predictions (the "residuals"). The model takes the general form:

Kindergarten Environment & Achievements Modelling Report

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

where Y is the predicted score, X_p represents the various predictors, β_p are the model coefficients, and ϵ is the error term. This model assumes that the relationships are linear, the errors are independent and normally distributed, and the predictors are not multicollinear.

This model was used as a baseline for performance. Its primary value is not in its predictive power but in providing the most basic, interpretable benchmark. Every other, more complex model must demonstrate a clear improvement over the OLS score to justify its added complexity. Furthermore, its coefficients provide the basis for the statistical inference in our "Drivers of Performance" analysis.

As expected, the OLS model was the weakest performer in our analysis. It achieved a 5-fold cross-validation MSE_0 of 1457.48. This high error score strongly suggests that the relationships between student, teacher, and school characteristics and student achievement are not purely linear and that the strict assumptions of OLS are likely violated.

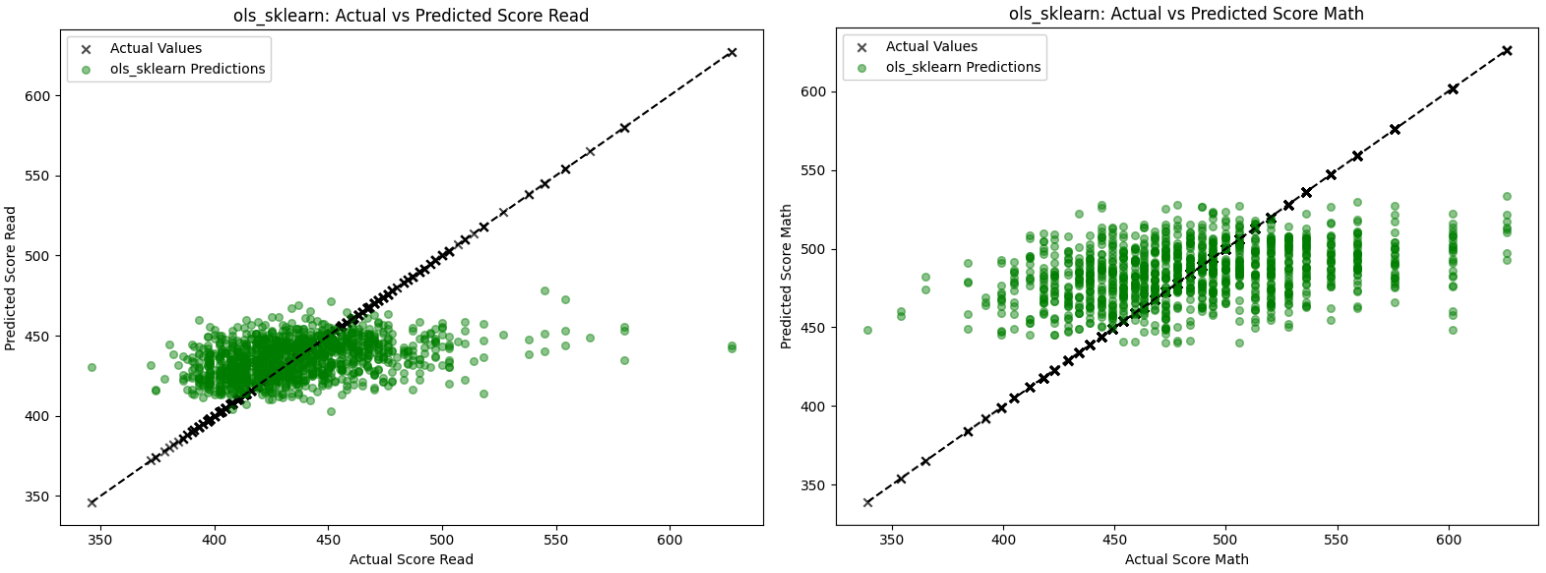


Figure: Scatter plots of actual vs. predicted Reading and Math scores for the OLS Regression model.

3.2 Penalised OLS

The next three models are advanced variations of OLS known as "penalised" or "regularised" regressions. They address the common weaknesses of OLS, such as overfitting and multicollinearity, by adding a penalty term to the loss function. This penalty discourages the model from assigning excessively large coefficients to any feature,

especially in a model with many predictors. This results in a more stable model by reducing its variance.

3.2.1 Ridge Regression

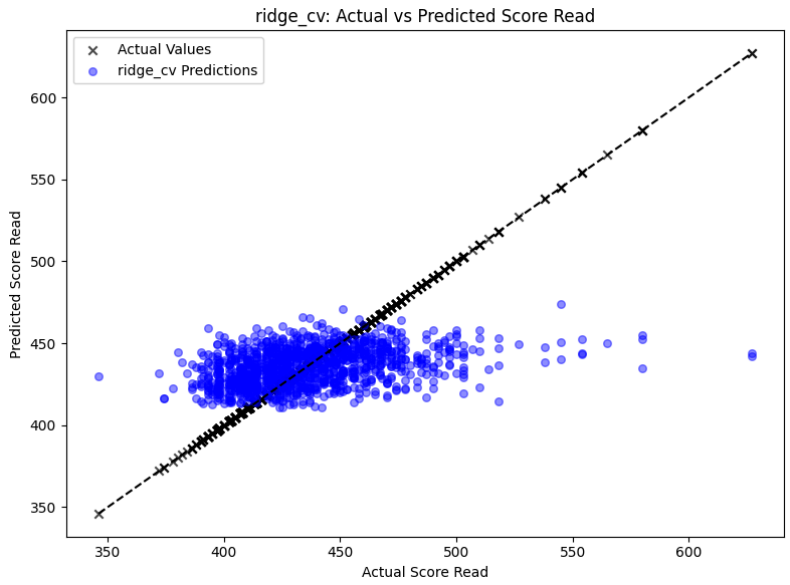
Ridge Regression uses an "L2 penalty," which adds the sum of the *squared* coefficients to the loss function, controlled by a tuning parameter, λ. The objective is to minimise:

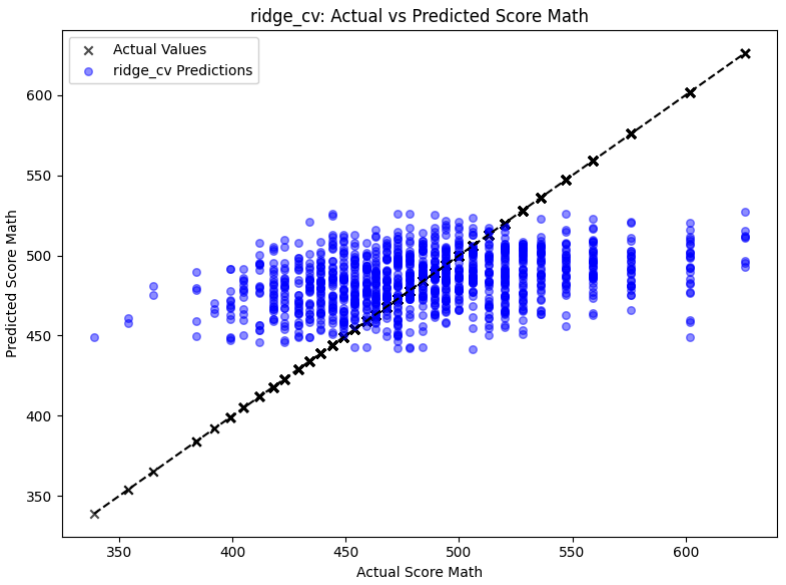
$$Loss_{Ridge} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}(\beta_j)^2$$

This penalty shrinks all model coefficients towards zero but never sets them to zero.

Ridge is effective at handling multicollinearity. By shrinking the coefficients, it stabilises the model and reduces the influence of less important features, rather than just selecting one and discarding the other. We used RidgeCV, which automatically finds the optimal λ using cross-validation.

The Ridge model performed slightly better than standard OLS, achieving a cross-validation MSE_0 of 1451.34. This small improvement confirms that some regularisation is beneficial for this dataset, likely by mitigating the multicollinearity identified in the EDA.





Kindergarten Environment & Achievements Modelling Report

Figure: Scatter plots of actual vs. predicted Reading and Math scores for the Ridge Regression model.

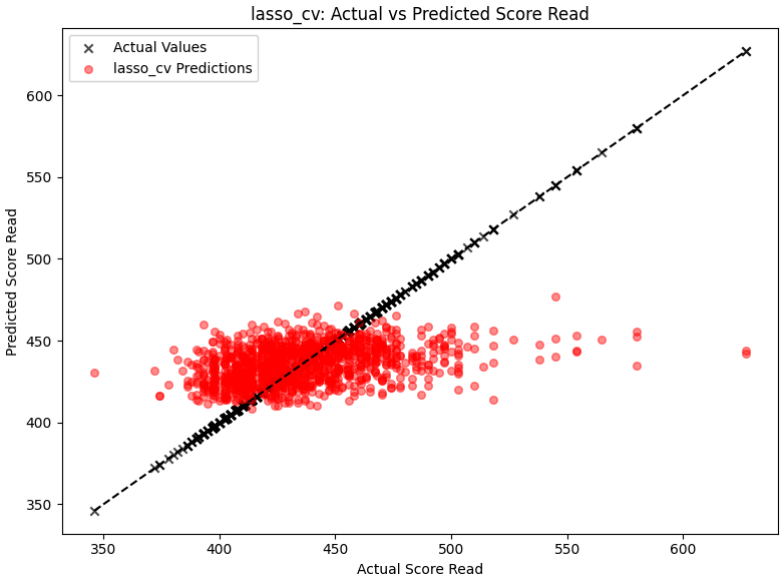
3.2.2 LASSO Regression

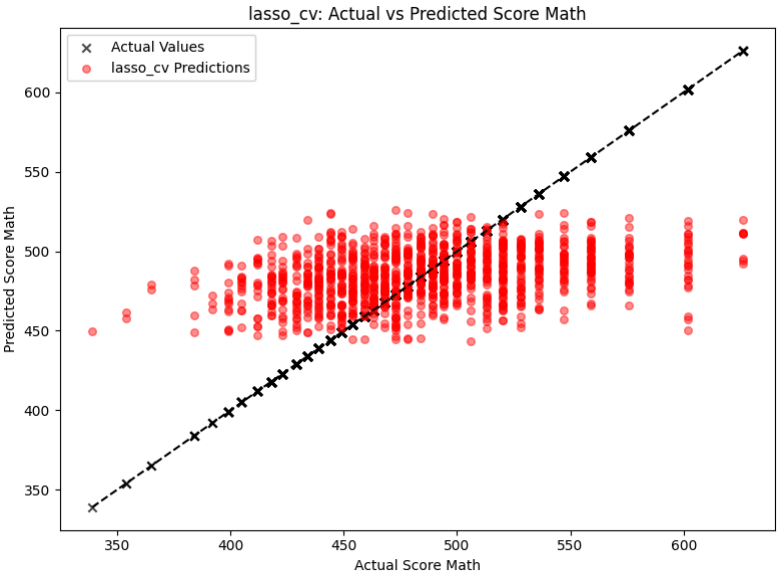
LASSO (Least Absolute Shrinkage and Selection Operator) Regression uses an "L1 penalty," which adds the sum of the *absolute values* of the coefficients, controlled by a parameter λ. The objective is to minimise:

$$Loss_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

This model was chosen because its L1 penalty performs automatic "feature selection." Its unique property is that it can shrink the coefficients of the least important features all the way to exactly zero. This creates a more simpler model, helping us identify which predictors are just noise and can be discarded.

The LASSO model's performance was similar to Ridge, with a cross-validation MSE_0 of 1452.58. This result suggests that while many features may have a small impact, few are completely irrelevant (i.e., deserving of a zero coefficient), so the model's feature-selection benefit did not translate into superior accuracy in this case.





Kindergarten Environment & Achievements Modelling Report

Figure: Scatter plots of actual vs. predicted Reading and Math scores for the LASSO Regression model.

3.2.3 Elastic Net Regression

Elastic Net is a hybrid model that combines both the L1 penalty of LASSO and the L2 penalty of Ridge, controlled by two parameters, λ_1 and λ_2. The objective is to minimise:

$$Loss_{ElasticNet} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}(\beta_j)^2$$

This model was seen as a best of both worlds approach. It is useful when there are many predictors, some of which are correlated. It can perform feature selection like LASSO while also handling correlated predictors effectively like Ridge which LASSO struggles with.

As expected, its performance landed directly between the other two penalised models, with a cross-validation MSE_0 of 1451.89. Ultimately, all three regularised linear models performed similarly and offered only a marginal improvement over OLS, confirming that a linear approach is not sufficient for this dataset.

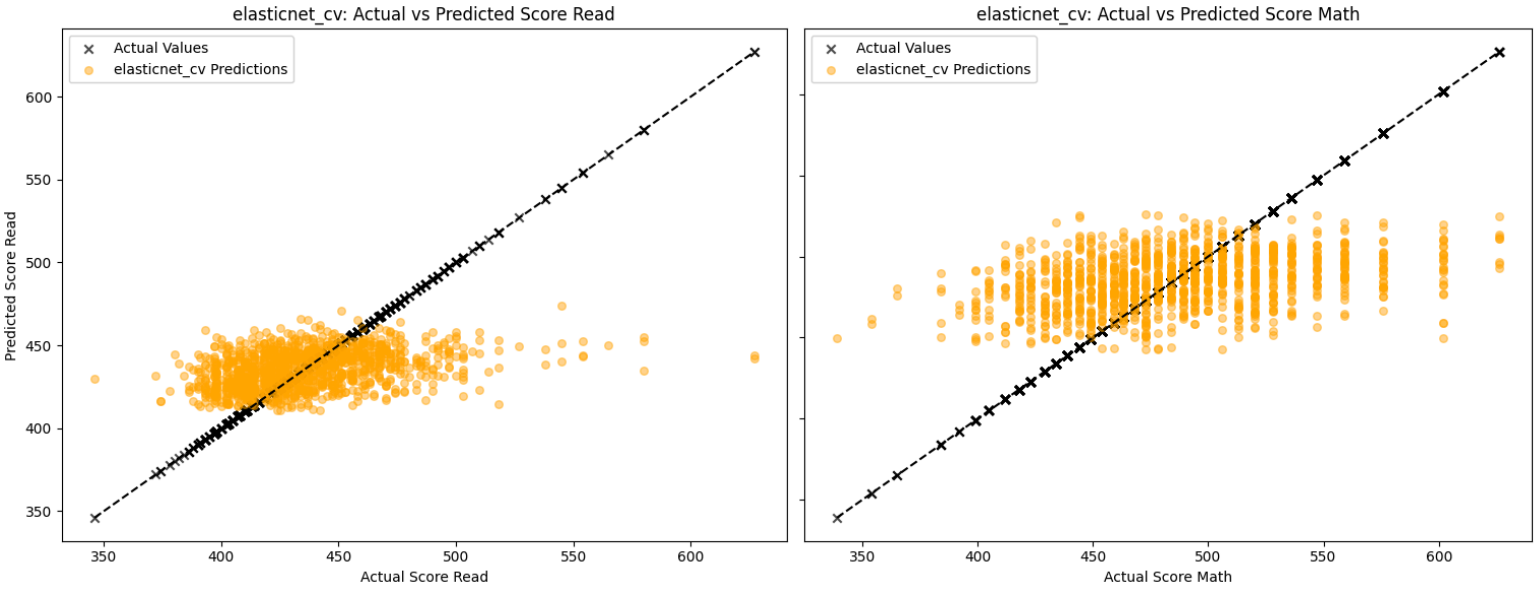


Figure: Scatter plots of actual vs. predicted Reading and Math scores for the Elastic Net model.

3.3 Random Forest

A Random Forest is a powerful non-linear "ensemble" model that uses a technique called "bagging" (bootstrap aggregating). It works by building hundreds of individual "decision

trees" on random subsets of the data. To make a final prediction, it averages the predictions of all the individual trees:

$$\#_{RF} = \frac{1}{B}\sum_{b=1}^{B}[f_b(x)]$$

We used this model because it is highly effective at capturing complex, non-linear relationships (e.g., "teacher experience only matters if the class is small") that linear models cannot. It also naturally prevents overfitting by averaging many different, high-variance, and uncorrelated trees, which results in a low-variance ensemble.

This model provided a significant leap in performance over the linear methods. It achieved a cross-validation MSE_0 of 1362.82, which is approximately 6.5% better than OLS. This score, substantially lower than any linear model, confirms that non-linear interactions are critical to accurately predicting student scores.

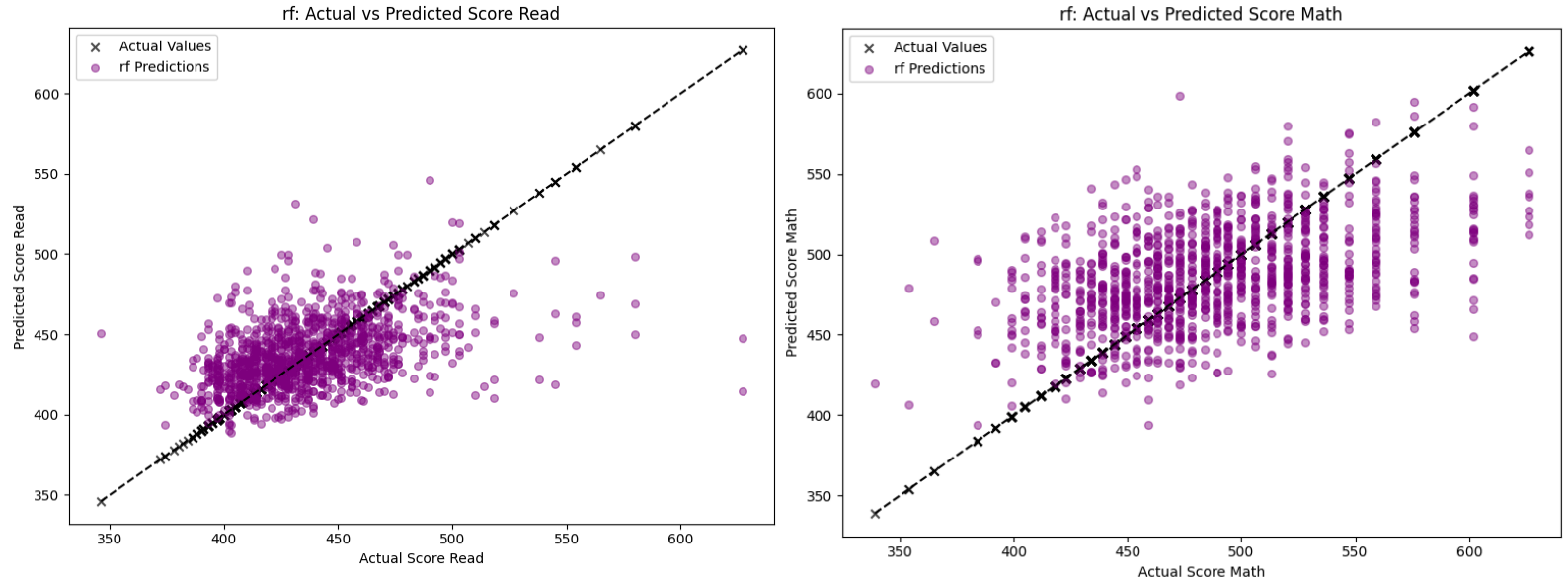


Figure: Scatter plots of actual vs. predicted Reading and Math scores for the Random forest model.

3.4 Gradient Boosting Regressor (GBR)

Like Random Forest, the GBR is another ensemble model that uses decision trees, but it uses a technique called "boosting." Instead of building and averaging trees independently, it builds them sequentially. Each new tree is specifically trained to identify and correct the errors (residuals) made by the previous tree. The model is built iteratively:

Kindergarten Environment & Achievements Modelling Report

$$F_m(x) = F_{m-1}(x) + h_m(x)$$

where $F_m(x)$ is the new model, $F_{m-1}(x)$ is the previous model, and $h_m(x)$ is the new tree that fits the "gradient" of the loss function.

This model was chosen because it is widely regarded as one of the most powerful and accurate models for data. Its ability to learn from its mistakes allows it to create a single, highly accurate predictor. The GBR model was the clear winner of our analysis, achieving the lowest (best) cross-validation MSE_0 by a large margin, with a score of 1209.59. This represents an 11% improvement over Random Forest and a 17% improvement over OLS. This performance is why it was selected as our final model.

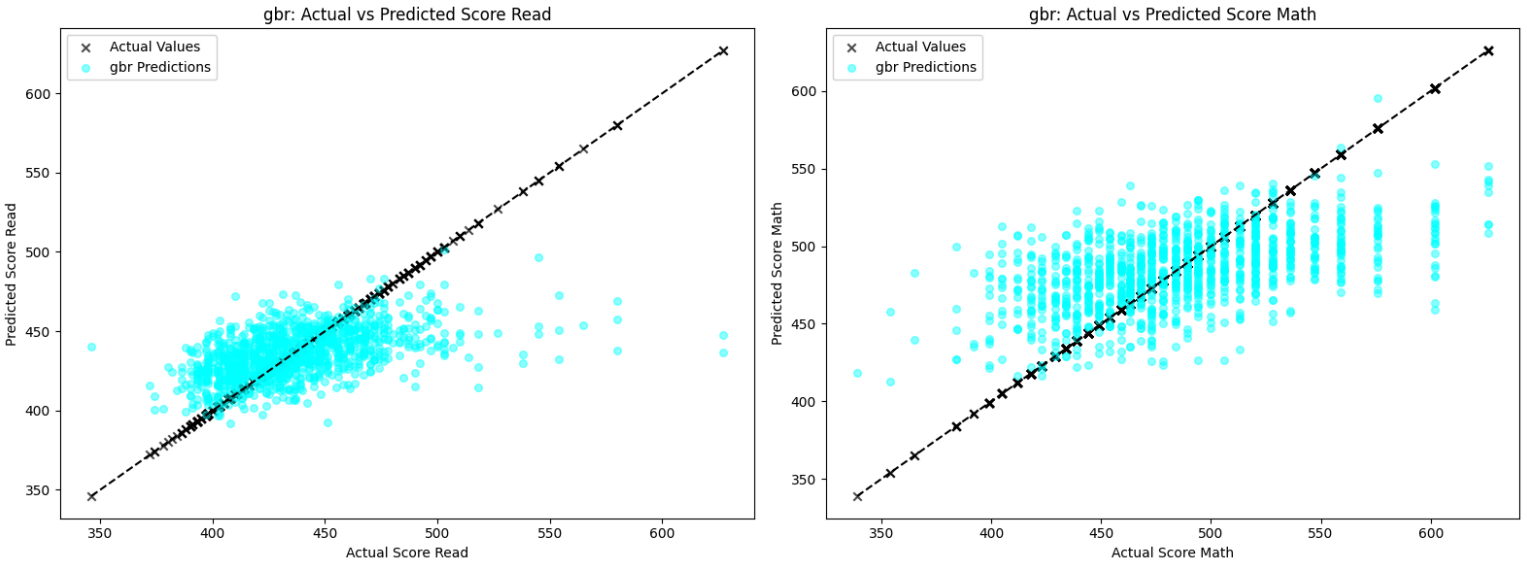


Figure: Scatter plots of actual vs. predicted Reading and Math scores for the GBR model.

3.5 Multi-Layer Perceptron (MLP)

The MLP is a type of "neural network". It passes data through several interconnected layers of nodes (we used two hidden layers of 128 and 64 nodes). Each node applies a non-linear activation function to a weighted sum of inputs from the previous layer. This architecture allows it to learn highly abstract and complex non-linear patterns.

As a key requirement of the assignment, we included this advanced machine learning model to see if it could capture even more intricate patterns than the tree-based ensembles.

The MLP performed well, achieving a cross-validation MSE_0 of 1425.04. This was a significant improvement over the OLS models, but was not as effective as the tree-based ensembles. This is likely because neural networks typically require very large datasets and

extensive hyperparameter tuning to perform at their best. The GBR, being a tree-based method, often achieves superior performance on structured, tabular data of this size.

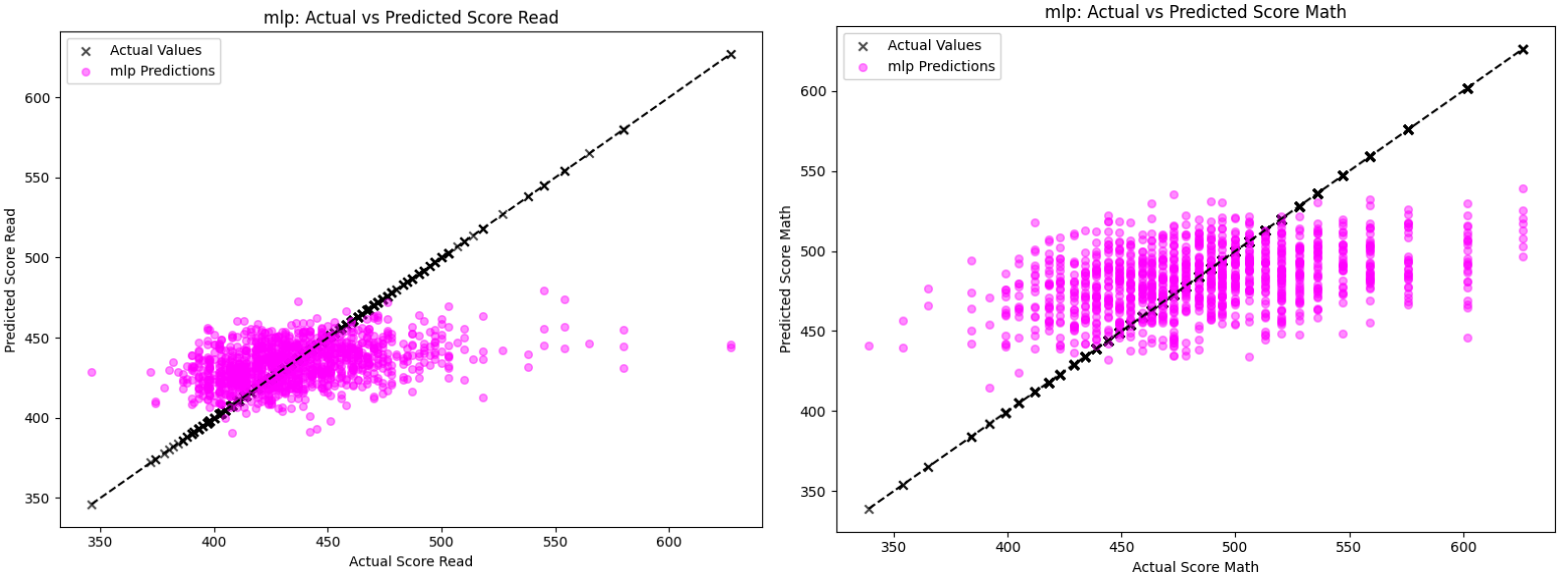


Figure: Scatter plots of actual vs. predicted Reading and Math scores for the MLP model.

3.6 Model Comparison Summary

Below is a summary of the 5-fold cross-validation results for all models, which clearly shows the superior performance of the Gradient Boosting Regressor.

Model	CV Mean MSE_o	CV Std. Dev. MSE_o
Gradient Boosting (GBR)	**1209.59**	24.81
Random Forest (RF)	1362.82	44.87
Multi-Layer Perceptron (MLP)	1425.04	56.25
Ridge Regression	1451.34	37.22
Elastic Net	1451.89	37.72
LASSO Regression	1452.58	37.40
Ordinary Least Squares	1457.48	38.25

Kindergarten Environment & Achievements Modelling Report

(OLS)		

The table above confirms the significant performance disparity between the linear and non-linear models. The ensemble methods (GBR and RF) achieved substantially lower MSE_0 scores, demonstrating a superior ability to capture the complex, non-linear relationships driving student achievement.

Specifically, the Gradient Boosting Regressor (GBR) secured the best score of 1209.59, which is over 200 points lower than the worst-performing linear model, the Ordinary Least Squares (OLS) regression. This performance made the GBR the choice for the final predictive model.

Kindergarten Environment & Achievements Modelling Report

4 Drivers of Performance

To understand what factors, or covariates, drive student achievement, we used two different methods. First, we used a simple Ordinary Least Squares (OLS) model to get statistically significant, interpretable coefficients. Second, we used a permutation importance test on our best model, the Gradient Boosting Regressor (GBR), to see which features the most accurate model relied on most.

4.1 Findings from Linear Models (OLS)

The OLS model provides standard statistical tests (like p-values) that tell us if a feature's relationship with scores is statistically significant and not just due to random chance. Based on the OLS output from our notebook, we found several key drivers.

As the original Project STAR experiment hypothesised, *class_type* was highly significant ($p<0.0001$), being in a 'small' class was associated with an increase of 6.25 points for reading and 9.76 points for math compared to a regular class. The strongest predictor, however, was *lunch* status ($p<0.0001$). Students with 'non-free' lunch, a proxy for higher socio-economic status, scored, on average, 14.57 points higher in reading and 18.21 points higher in math.

The variable *gender* was also highly significant ($p<0.0001$), with male students scoring 5.76 points lower in reading and 8.13 points lower in math. Finally, teacher factors were also important, teacher experience had a small but significant positive effect ($p<0.0001$), and the teacher's career ladder level was a strong driver, with 'level3' teachers associated with a score increase of over 20 points in both subjects.

4.2 Findings from the Gradient Boosting (GBR) Model

While OLS shows us simple linear relationships, our best-performing GBR model can find more complex, non-linear patterns. We used a technique called permutation importance, which measures how much our model's error increases when a feature is shuffled. For categorical variables, the total permutation importance was found by taking the sum of all permutation importances across their one-hot encoded representatives. A higher error increase means the model relies on that feature heavily.

The results from the GBR were comparatively different. The most important features for this model were

1. Student Birth Quarter
2. School ID
3. School District ID
4. Gender
5. Teacher Experience

This discrepancy is a key finding. The GBR model found that complex, non-linear interactions between a student's birth quarter and their specific school had the most predictive power. In contrast, *class_type* and lunch, which were dominant in the OLS model, were less important for the GBR. This suggests the GBR captured the linear effects of class size and lunch status easily, but had to work harder (place more importance) on the non-linear patterns of when a student was born and where they went to school. THe difference likely reflects the predictive performance benefit in using a complex machine learning tree model like the GBR as opposed to the interpretable OLS model.

4.3 Omitted Variables and Random Assignment

A major challenge in this type of analysis is omitted variable bias, with key data missing, such as individual student aptitude, parental income and education, or specific details on teacher quality. Under normal circumstances, this bias would make any inferences and findings unusable. For example, parents with higher education (an omitted variable) might be more likely to enrol their children in schools with small classes. In that case, we wouldn't know if the higher scores were due to the small class or the parents' education.

This highlights the importance of random assignment for Project STAR. As noted in the assignment background, students were randomly assigned to one of the three class types within their school, and this randomisation breaks the link between variables like parental background and the student's class type. On average, it ensures that all class types (small, regular, regular+aide) have the same mix of students from different backgrounds. Due to this random assignment, the effect we measured for *class_type* is likely a credible and unbiased estimate of the causal effect of small classes on achievement. It must be noted that this special causal interpretation only applies to the *class_type* variable.

5 Discussion

5.1 Model Interpretability

A major theme in our modelling process was the fundamental trade-off between interpretability and accuracy. The linear models (OLS, Ridge, etc.) have the benefit of being highly interpretable, relative to more complex models. For example, it can be easily determined that being in a small class is associated with a 6.25-point increase in reading scores, holding all other factors constant. The drawback is that these models are limited to simple linear relationships, and therefore can lack predictive performance in comparison with other models, as substantiated by our OLS model achieving an MSE_0 of 1457.48.

On the other hand, our chosen Gradient Boosting Regressor (GBR) is far more accurate but is an opaque black-box model. The model learns thousands of complex branching rules and interactions that cannot be distilled into a simple equation. Consequently, we cannot isolate the effect of a single variable with the same ease as OLS. While it may be known what features are most important to the GBR (such as birth quarter), we cannot easily

characterise the functional form or interaction patterns through which they influence predictions.

For this assignment, where the objective was to build "well-validated predictive models" and performance was evaluated on the MSE_0 metric, the GBR was the clear choice. Its MSE_0 1209.59 was significantly better than any other model, making it the most accurate and best model according to the project's criteria.

5.2 Caveats and Limitations

Several limitations must be considered when interpreting these results. First, our model was trained on a specific sample of data, and the OLS diagnostics revealed potential stability issues. In particular, the OLS model exhibited signs of high multicollinearity. While some of this collinearity may be explained by variables that are expected to be correlated (e.g., *school_id* and *schooldistrict_id)*, the specific coefficients from the linear models may be unstable and might not be reliable if applied to a new set of data. Although the coefficients may broadly reflect the general influence of each variable's contribution to the model, they may not provide reliable quantitative estimates.

A key limitation is how we handled missing scores, having to drop over 800 rows (16%) where either a reading or math score was missing. This method assumes the missing data is random, although it is reasonable to suspect that lower-performing students may be more likely to have missing scores. This resulted in our model being trained on an overly optimistic sample, meaning its actual performance on the full student population is likely lower than reported.

Finally, the MSE_0 metric was predetermined by the assignment. This metric not only gives equal importance to reading and math errors but, as an MSE, it also penalises large errors quadratically. This means a model that is generally accurate but makes a few wrong predictions will be penalised more heavily than one that is consistently slightly inaccurate. In a real-world policy scenario, a school district may decide that improving reading scores is twice as important as improving math, or that they prefer a model with no extreme errors. This would require a custom-weighted metric, which could, in turn, lead to a different best model.

6 Conclusion

This report successfully developed and evaluated seven different statistical models to predict kindergarten student achievement in reading and math. The analysis began with an Exploratory Data Analysis (EDA), which identified strong correlations between the target variables (*score_read* and *score_math*) and highlighted key potential predictors, including *class_type*, *lunch status*, and *gender*.

Kindergarten Environment & Achievements Modelling Report

Our comparative analysis of the seven models, using the MSE_0 metric, demonstrated a clear performance hierarchy. Standard linear models (OLS, Ridge, Lasso, Elastic Net) were poor predictors (OLS MSE_0: 1457.48), confirming that simple linear relationships are insufficient to capture the complexities of student achievement. In contrast, non-linear ensemble models were far superior.

The Gradient Boosting Regressor (GBR) was identified as the best-performing model, achieving the lowest (best) cross-validation MSE_0 of 1209.59. This model was therefore selected as our final predictive model.

While the GBR model is the most accurate for prediction, it is a black box" that is difficult to interpret. Conversely, the OLS model, while less accurate, provided clear, interpretable drivers. Most notably, due to the experiment's random assignment, the OLS model gives a credible causal estimate of the impact of class size, showing that 'small' classes are associated with an average score increase of 6.25 points in reading and 9.76 points in math.

Ultimately, this report highlights the trade-off between accuracy and interpretability. We recommend the GBR model for tasks where predictive accuracy is the key. For understanding policy implications and finding causal drivers, however, the OLS model provides the most valuable and unbiased insights into the specific, positive causal effect of reducing class sizes.

Kindergarten Environment & Achievements Modelling Report

7 References

Breiman, L. 2001, ‘Random Forests’, *Machine Learning*, vol. 45, no. 1, pp. 5–32.

Chetty, R., Friedman, J.N., Hilger, N., Schanzenbach, D.W., Saez, E. and Yagan, J. 2011, ‘How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR’, *The Quarterly Journal of Economics*, vol. 126, no. 4, pp. 1593–1660.

Friedman, J.H. 2001, ‘Greedy Function Approximation: A Gradient Boosting Machine’, *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232.

Hastie, T., Tibshirani, R. and Friedman, J.H. 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn, Springer.

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2023, *An Introduction to Statistical Learning: With Applications in Python*, Springer.

Pedregosa, F. et al. 2011, ‘Scikit-learn: Machine Learning in Python’, *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830.

Scikit-learn Developers. 2025, *GradientBoostingRegressor — scikit-learn 1.5.0 Documentation*, viewed 6 November 2025, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.

StatQuest with Josh Starmer. 2025, *StatQuest Video Index*, viewed 8 November 2025, https://statquest.org/video_index.html.

Tibshirani, R. 1996, ‘Regression Shrinkage and Selection via the Lasso’, *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288.