Predicting Student Performance

Linear Models (PHP2601), Prof. Ani Eloyan

Daniel Posmik, Jizhou Tian, Aristofanis Rontogiannis

2024-12-10

Table of contents I

Linear Model (OLS Regression)

Introduction

We will be analyzing educational data to understand the predictors of student performance. Specifically, we seek to **understand whether five predictors – as a subset of an exhaustive list of potential predictors – are significant predictors of student performance**.

Testing the significant of a subset of predictors is becoming increasingly important in modern statistical questions, especially with more information becoming available.

We will be using a publicly available dataset from Kaggle that contains information about students and their exam scores.

Hypothesis to be Tested

We are interested in:

- Hours Studied
- Attendance
- Sleep Hours
- Previous Scores
- Tutoring Sessions

We can formalize this question as follows:

- $H_0 : \left[1_{[0,\cdots,p+1]}, \quad 0_{[p+2,\cdots,P]}\right] \cdot \begin{bmatrix}\beta_0 & \cdots & \beta_P\end{bmatrix}^T = \beta_0 + \cdots + \beta_{p+1} = 0$
- $H_A : \{\beta_1 \neq 0\} \cap \cdots \cap \{\beta_5 \neq 0\}$

Observe the 0-indexed variables from $p+2$ to P.

Exploratory Data Analysis (EDA)

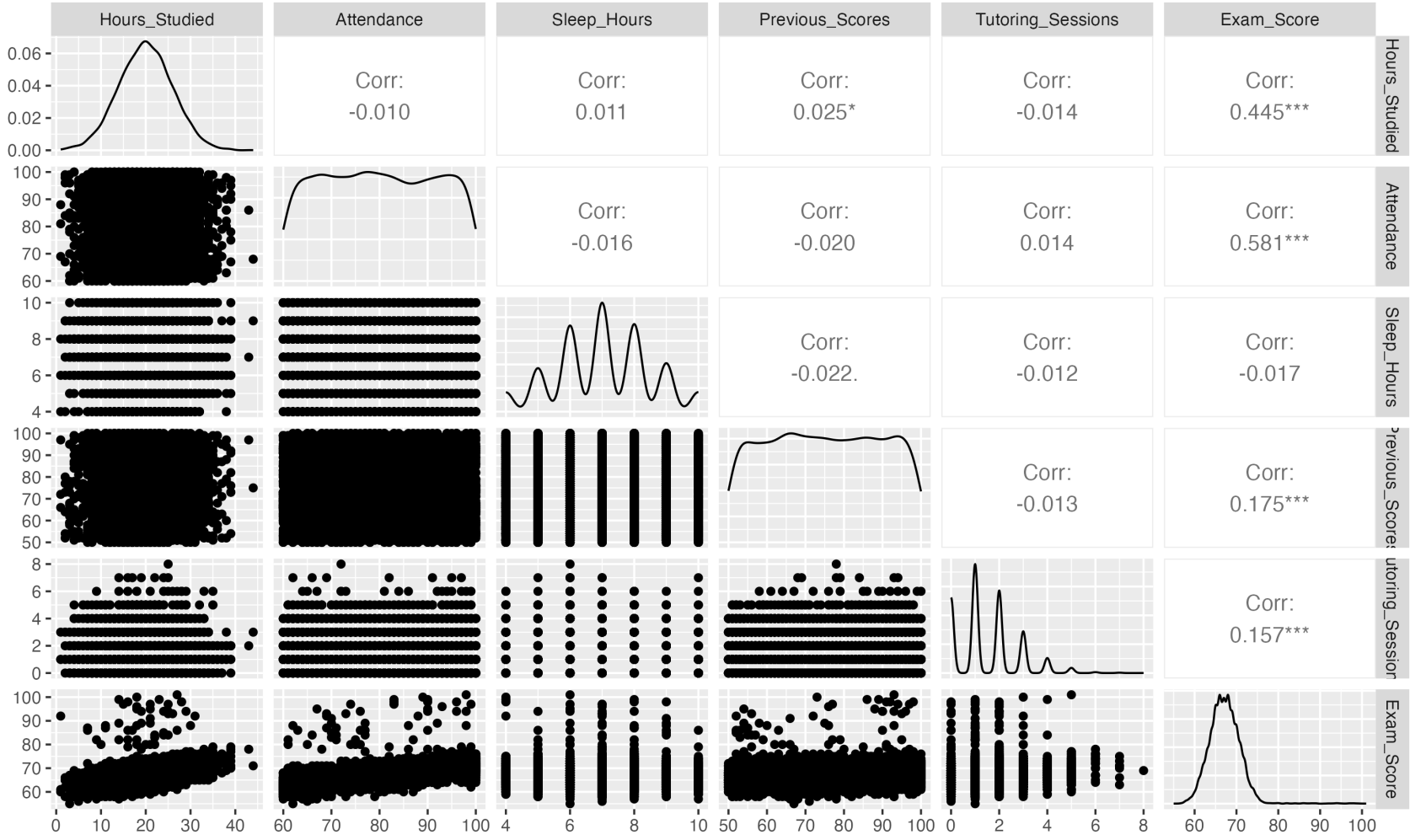


Figure 1: Correlation Matrix

Variable Transformations

We will transform the variables to ensure that the assumptions of the linear model are met.

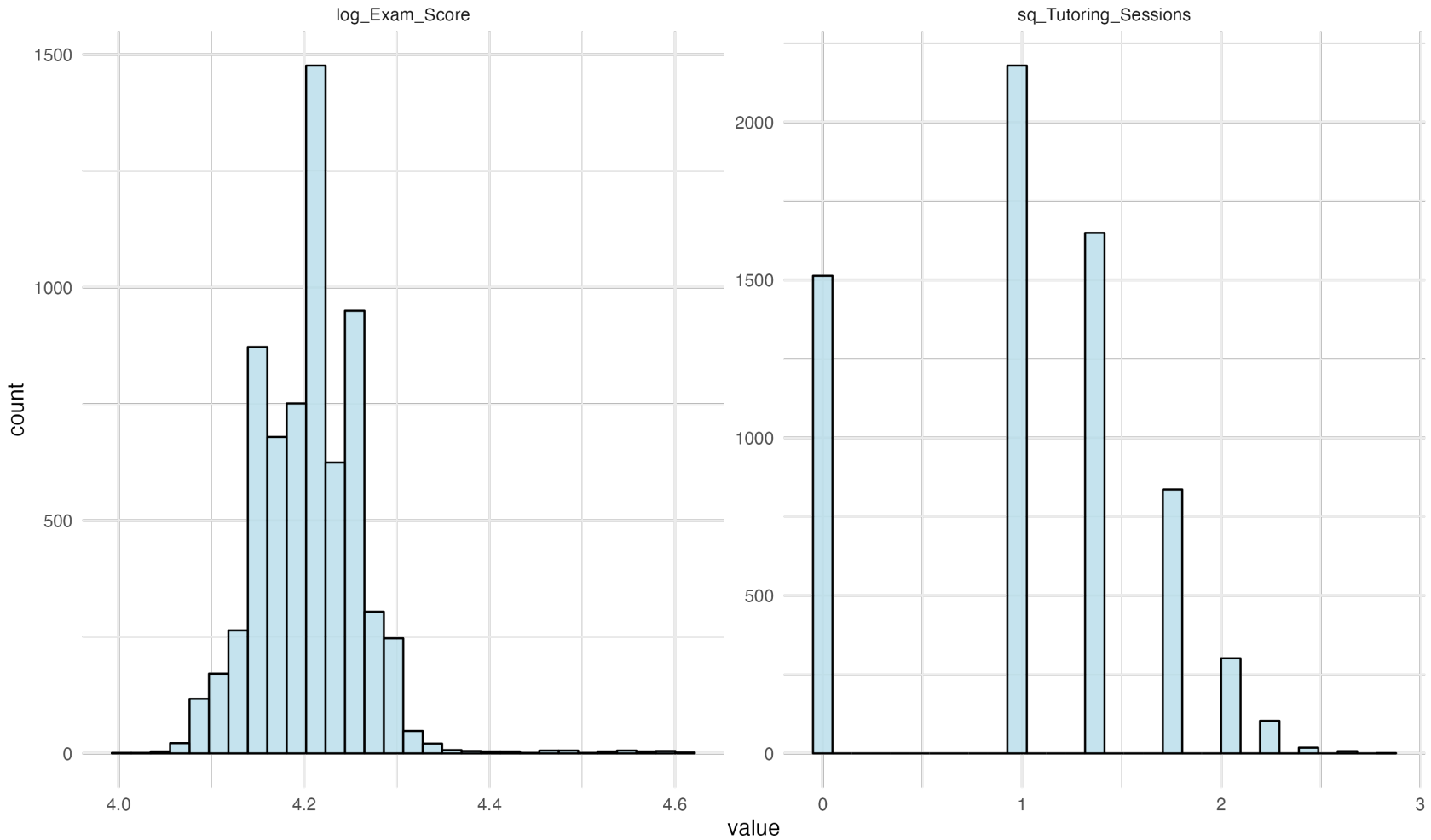


Figure 2: Variable Transformation

The Linear Model

Let us begin by discussing the assumptions of linear regression model. In a Gauss-Markov setting, we assume that our linear model is of the form:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{12} & X_{13} & \cdots & X_{1(p+1)} \\ 1 & X_{22} & X_{23} & \cdots & X_{2(p+1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & X_{n3} & \cdots & X_{n(p+1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $\mathbb{E}[\epsilon] = 0$ and $\text{Var}[\epsilon] = \sigma^2 I$ denote the zero-mean and constant variance assumptions. In our case, we begin with $p = 5$, i.e. our design matrix has $p + 1$ columns, accounting for the intercept term.

Solving for $\hat{\beta}$

We can solve for $\hat{\beta}$ via the normal equations:

$$\hat{\beta} = (X^T X)^g X^T Y$$

$$= \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1(p+1)} & X_{2(p+1)} & \cdots & X_{n(p+1)} \end{bmatrix} \begin{bmatrix} 1 & X_{12} & \cdots & X_{1(p+1)} \\ 1 & X_{22} & \cdots & X_{2(p+1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \cdots & X_{n(p+1)} \end{bmatrix} \right)^g .$$

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1(p+1)} & X_{2(p+1)} & \cdots & X_{n(p+1)} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

In our case, all predictors but Sleep Hours are significant predictors of exam scores, even at a 1% level of significance.

Estimability of the Hypothesis

Question: **Can we estimate an object $K^T\beta$ with our data X?**

Formally, we say that if $\exists\ A$ s.t. $X^TA = K^T$, i.e. K^T can be expressed as a linear combination of X and some matrix A, then $K^T\beta$ is estimable.

In our case, this is straightforward to verify. Can we think of an example when this is not true? (Hint: Dimension "mismatch")

Distribution of $K^T\beta$

Since $K^T\beta$ estimable, its best linear unbiased estimator (BLUE) is given by:

$$\mathbf{K_i}^T\hat{\beta} \sim N(\mathbf{K_i}^T(X^TX)^gX^TX\beta, \sigma^2\mathbf{K_i}^T(X^TX)^g\mathbf{K_i}) \quad \text{and}$$
$$\mathbf{K}^T\hat{\beta} \sim N(\mathbf{K}^T(X^TX)^gX^TX\beta, \sigma^2\mathbf{K}^T(X^TX)^g\mathbf{K})$$

This object $K^T\beta$ may seem a bit arbitrary, even useless, at first. However, it is in fact the building block for the test statistic we will construct now!

Quadratic Form in our Joint Testing Procedure

Suppose $H := K(X^T X)^g K^T$, then

$$(K\beta)^T(\sigma^2 H)^{-1}(K\hat{\beta}) \sim \chi^2_{\text{df}=\text{rank}(H)}(\lambda)$$

where the non-centrality parameter $\lambda = \frac{1}{2}(K\beta)^T(\sigma^2 H)^{-1}(K\beta)$ is the well-known distributional result of a normal quadratic form.

Finally, our F Statistic:

$$F := \frac{\left((K\beta)^T(\sigma^2 H)^{-1}(K\beta)\right)/\text{rank}(H)}{\text{RSS}/(n-p)} \sim \frac{\chi^2(\lambda)}{\chi^2} \sim F_{\text{rank}(H),n-p}(\lambda)$$

We have successfully constructed a statistical test that allows us to test our hypothesis with a simple F-test. In `R`, we can use the `anova()` function to perform this test.

Results

```
Analysis of Variance Table

Model 1: log_Exam_Score ~ 1
Model 2: log_Exam_Score ~ Hours_Studied + Attendance + Sleep_Hours + Previous_Scores +
    sq_Tutoring_Sessions
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   6606 20.6969
2   6601  7.4771  5     13.22 2334.2 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: F-Test Results

The result shows that under the null hypothesis, the probability of getting a more extreme result than our calculate F-test statistics $\Pr(> F)$ is $2.2e-16$.

This evidence would lead us to reject the null hypothesis and conclude that our subset of predictors is indeed a significant predictor of exam scores

Linear Model (Elastic Net)

Why Use Elastic Net?

- **Limitations of Lasso**: May select only one variable from a group of highly correlated predictors.
- **Limitations of Ridge**: Cannot produce sparse models (i.e., no feature selection).
- **Elastic Net Advantage**:
  - Encourages group selection.
  - Balances sparsity and multicollinearity handling.

Elastic Net Formula

Elastic Net adds two penalty terms:

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right)$$

- $\|\beta\|_1$: Lasso penalty (L1).
- $\|\beta\|_2^2$: Ridge penalty (L2).
- λ_1, λ_2: Regularization parameters.

Tuning Parameters in Elastic Net

1. α: Controls the mix between Ridge and Lasso.
   - $\alpha = 0$: Ridge.
   - $\alpha = 1$: Lasso.
   - $0 < \alpha < 1$: Elastic Net.
2. λ: Controls the overall strength of regularization.

- Grid Search: Perform cross-validation to find optimal values of α and λ.

Elastic Net: Geometric Interpretation

- Elastic Net creates a penalty region combining L1 (diamond) and L2 (circle).
- Encourages sparsity while handling correlated features.

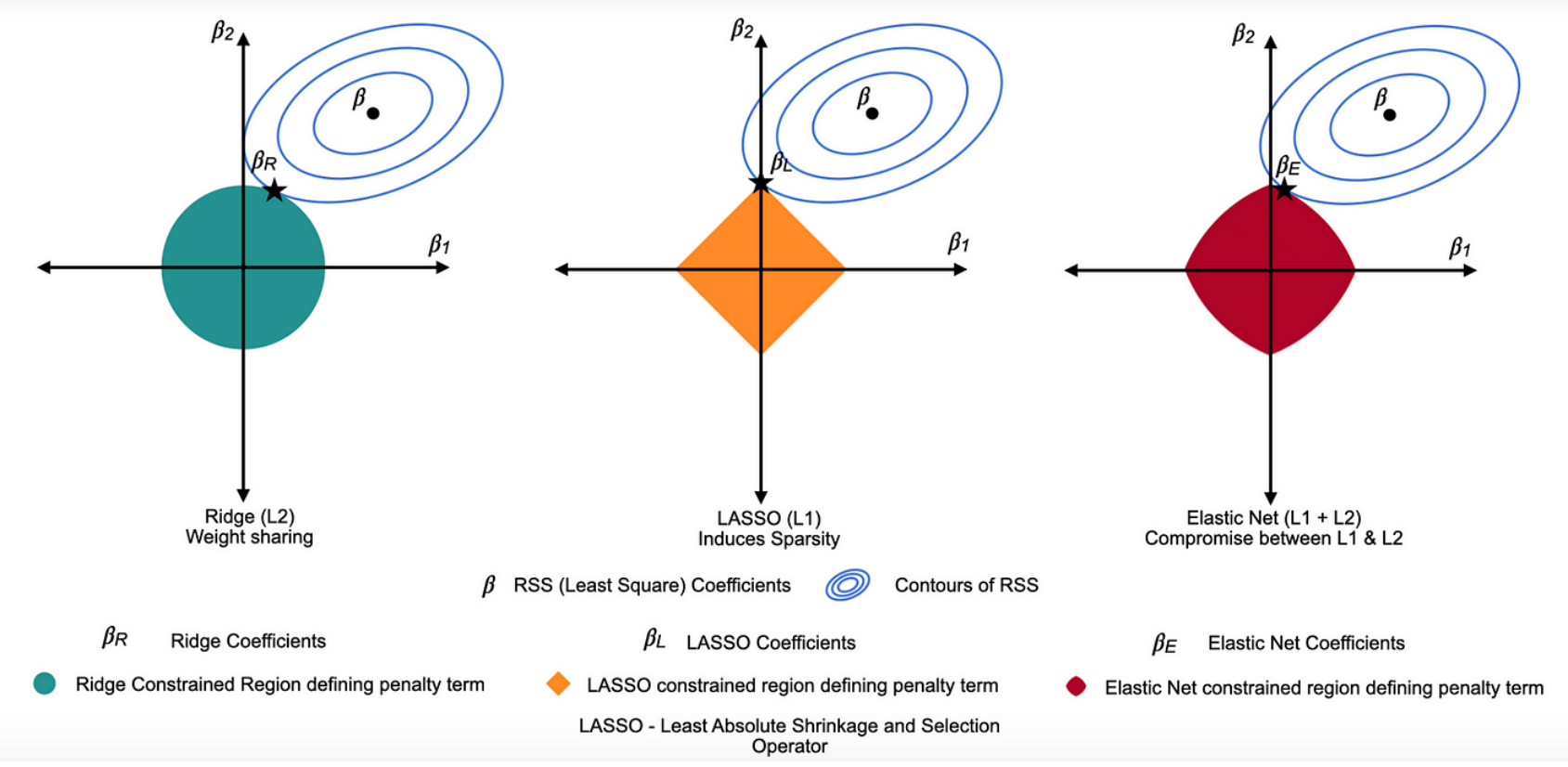


Figure 4: Elastic Net compared to Lasso and Ridge Regression

Elastic Net with Continuous Outcome

The R^2 on the test data is calculated to be approximately 76%, meaning our model is able to explain 76% of the variance in Exam Score in the test dataset

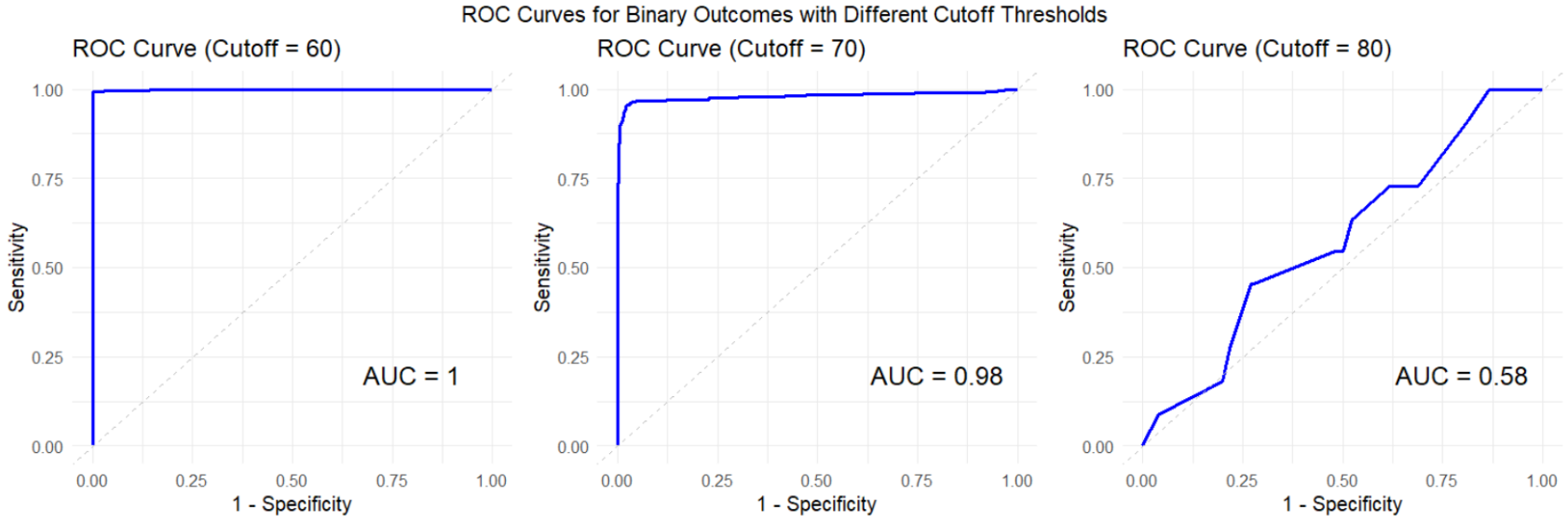
Elastic Net Coefficients

Predictor	Elastic Net Coefficient
(Intercept)	40.7199
Access_to_ResourcesLow	-2.0123
Teacher_QualityLow	-1.1046
Family_IncomeLow	-1.0943
Parental_InvolvementHigh	1.0873
Motivation_LevelLow	-1.0276
Peer_InfluencePositive	1.0049
Internet_AccessYes	1.0020
Access_to_ResourcesMedium	-0.9558
Distance_from_HomeNear	0.8904
Parental_InvolvementLow	-0.8389
Learning_DisabilitiesYes	-0.7820
Family_IncomeMedium	-0.6332
Extracurricular_ActivitiesYes	0.6087

Figure 5: The first 14 rows of the Elastic Net coefficients table

Note: Elastic Net (or Lasso) did not drop any variables as all predictors contribute to reducing the loss function, even with regularization applied.

Elastic Net with Binary Outcome



Note: This is an imbalanced dataset as most of the students have scored more than 60 in the exams. The median (and the mean) of the dataset is very close to the 3rd quantile (69). For instance, if we use threshold $= 70$, we can predict the probability a student's score is within/out the top 25% of the scores almost perfectly.

Non-Linear Model (Random Forest)

Random Forest

Random Forest is based on the idea of combining multiple decision trees to improve predictive performance and robustness.

Each decision tree in the Random Forest is trained on a bootstrap sample of the original data. At each split in a tree, a subset of the predictor variables is selected **randomly** to determine the split. This process ensures that the individual trees in the ensemble are decorrelated.

- Mitigate the over-fitting problem of individual decision tree.
- Handle complex non-linear relationships between variables.

Variable Importance Plot

0ut-of-bag (OOB) MSE is 6.638.

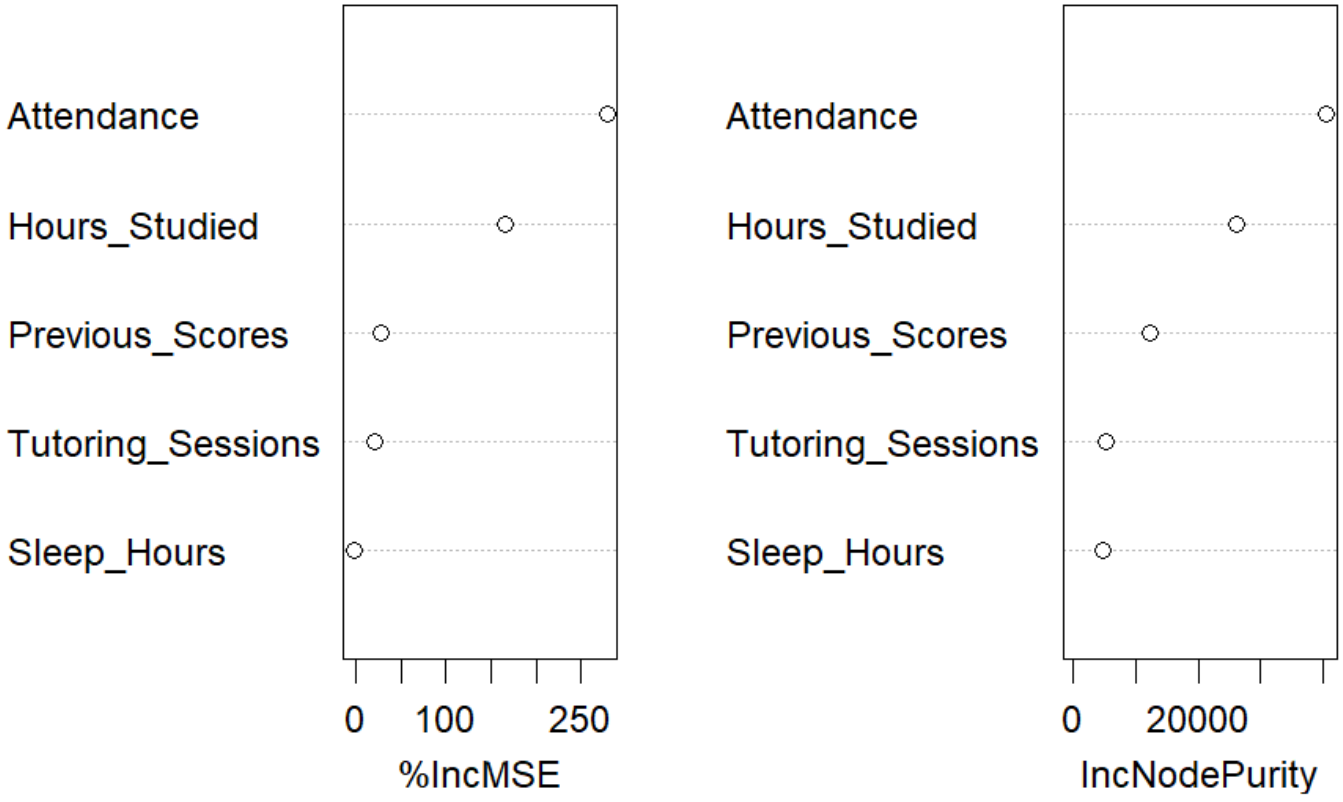


Figure 6: Variable Importance Plot

Attendance is the most important variable, while Sleep_Hours is the least important.

Approximation by a Single Regression Tree

Random Forest:

- Computationally intensive
- Less interpretable compared to simpler models

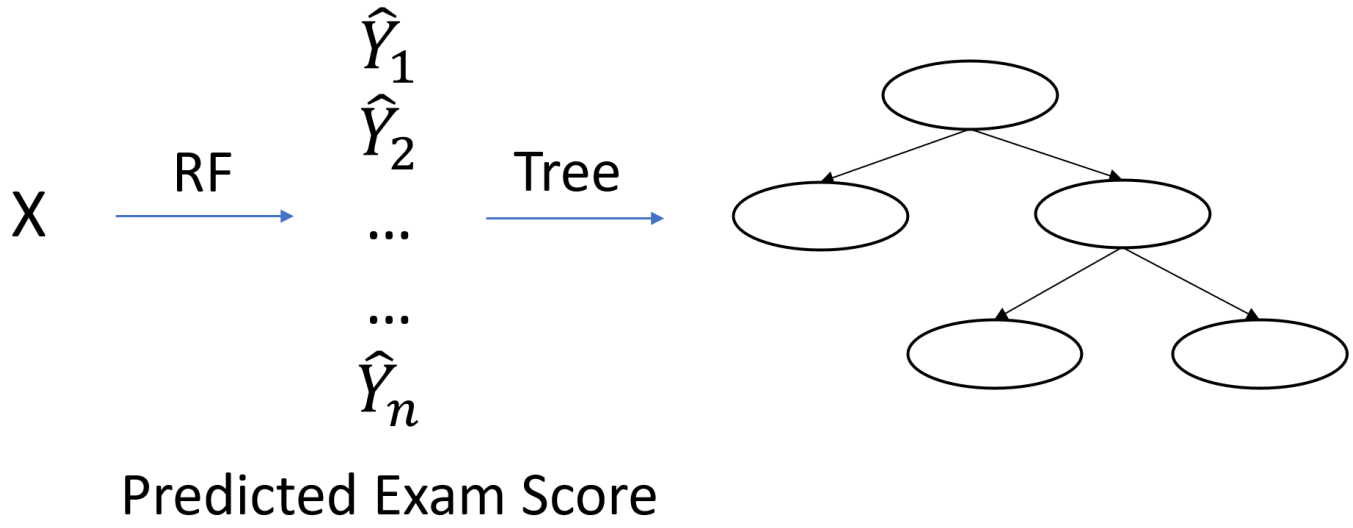
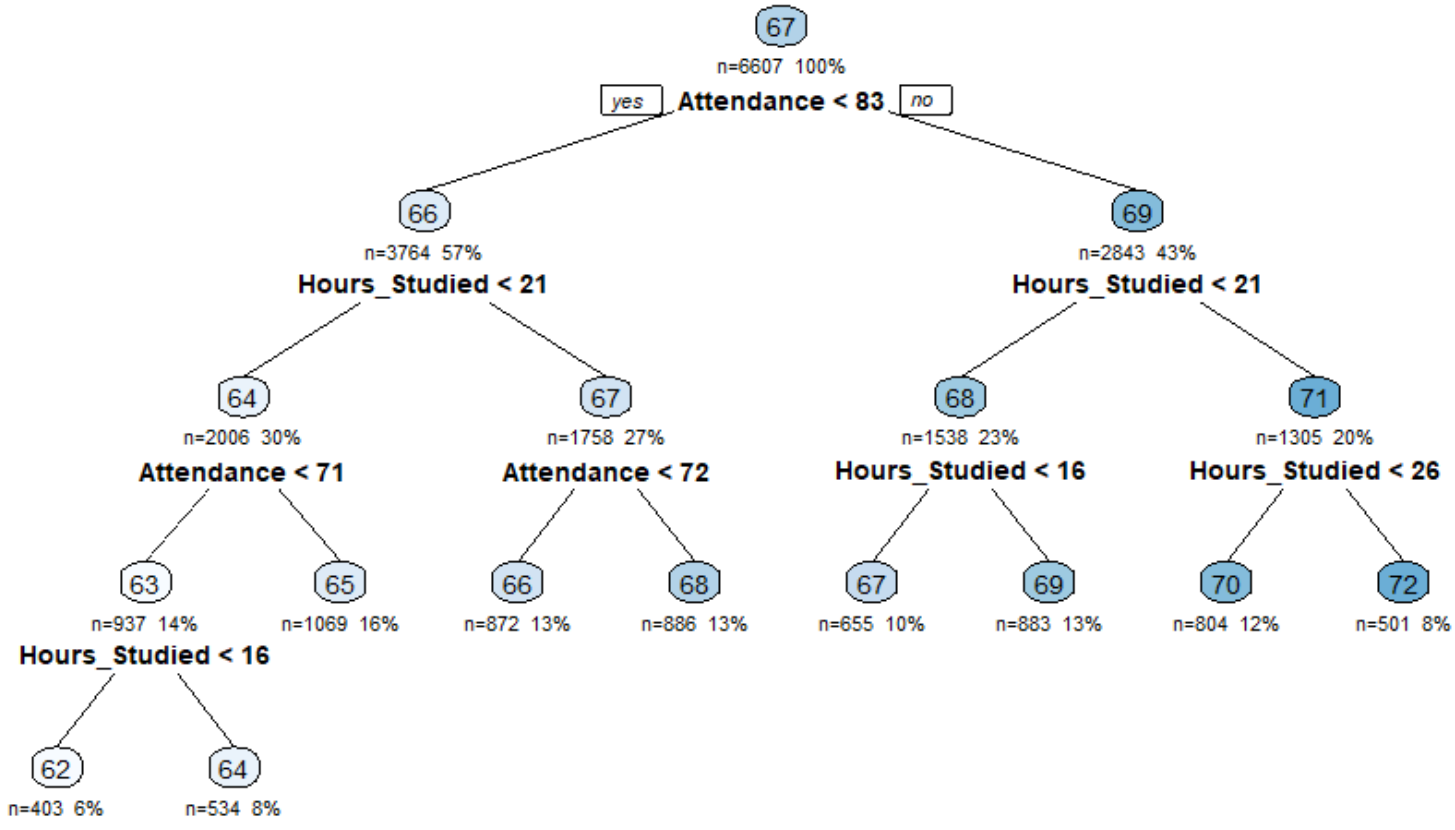


Figure 7: RF and Single Tree

Pearson correlation $= 0.898$.

Approximation by a Single Regression Tree



Results

Attendance emerges as the most critical variable.

The tree reveals **interaction effects** between Attendance and Hours_Studied, as their thresholds creating different groups with varying predicted values.

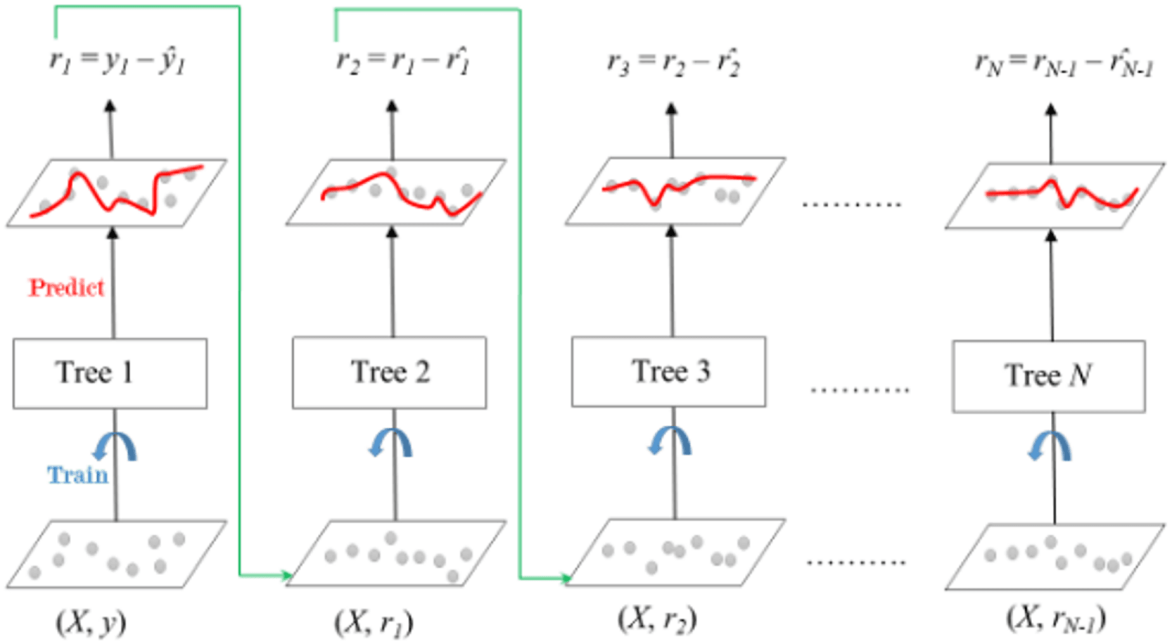
Non-Linear Model (Gradient Boosting)

What is Gradient Boosting?

- **Definition**: Gradient boosting is a machine learning ensemble technique (ensemble models combine predictions from multiple base models to enhance overall performance) that sequentially combines the predictions of multiple weak learners, typically decision trees.
- **Purpose**: It aims to improve overall predictive performance by optimizing the model's weights based on the errors of previous iterations, gradually reducing prediction errors and enhancing the model's accuracy.

How It Works

- **Step 1**: Start with a baseline model (e.g., mean prediction for regression)
- **Step 2**: Compute residuals or errors from the current model
- **Step 3**: Fit a new model to the residuals (weak learners like decision trees)
- **Step 4**: Update the overall model by adding the new learner
- **Step 5**: Repeat until convergence or a predefined number of iterations



In our dataset

Feature <chr>	Gain <dbl>	Cover <dbl>	Frequency <dbl>
Attendance	0.423404987	0.149385059	0.126120858
Hours_Studied	0.242200075	0.154748062	0.148343080
Previous_Scores	0.078287699	0.087571347	0.142300195
Tutoring_Sessions	0.037361117	0.066218270	0.073099415
Parental_InvolvementHigh	0.025577492	0.033937397	0.030994152
Access_to_ResourcesLow	0.024440631	0.048652850	0.031968811
Parental_InvolvementLow	0.015517165	0.027940427	0.030994152
Access_to_ResourcesMedium	0.014754528	0.028019759	0.027290448
Distance_from_HomeNear	0.011689131	0.032677304	0.025536062
Peer_InfluencePositive	0.011325110	0.034127656	0.023976608
Physical_Activity	0.010671582	0.027324448	0.046003899
Sleep_Hours	0.010637743	0.028012841	0.041520468
Family_IncomeLow	0.010449042	0.031686573	0.020662768
Motivation_LevelLow	0.010428008	0.031833015	0.023196881

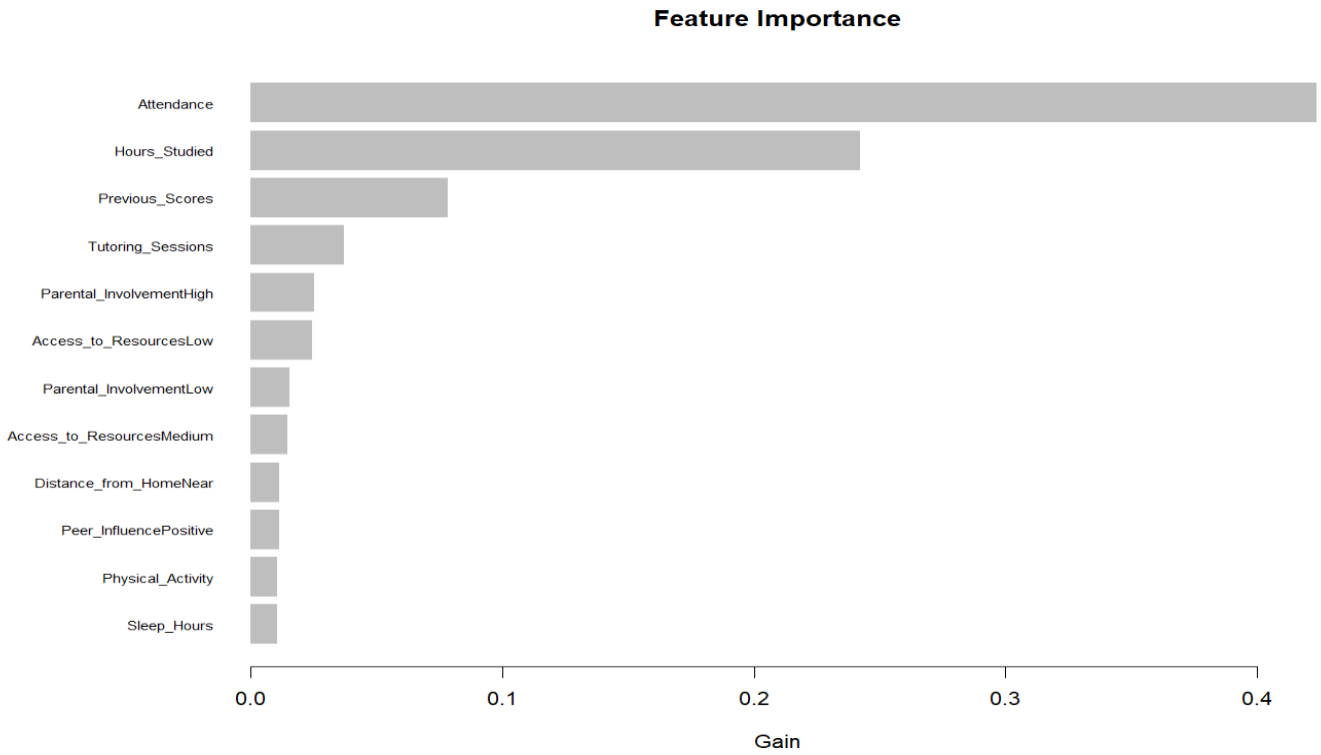
Figure 9: Feature Importance Summary

- **Feature**: Lists the features (variables) in the dataset.
- **Gain**: Contribution of the feature to the model's accuracy. Higher values indicate greater importance. **Attendance** contributes the most (0.4234).
- **Cover**: Proportion of samples impacted by the feature during splits. Higher values mean broader impact. **Attendance** has the highest Cover (0.1493).
- **Frequency**: How often the feature is used in tree splits. Higher values suggest frequent use. **Hours_Studied** is split most often (0.1484).

In our dataset (cont.)

Key Insights:

- **Top Features**: “Attendance” and “Hours_Studied” are the most impactful features.
- **Low-Impact Features**: Features like “Motivation_LevelLow” contribute minimally and may be less relevant.



Thank You

Thank you for your attention!