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.

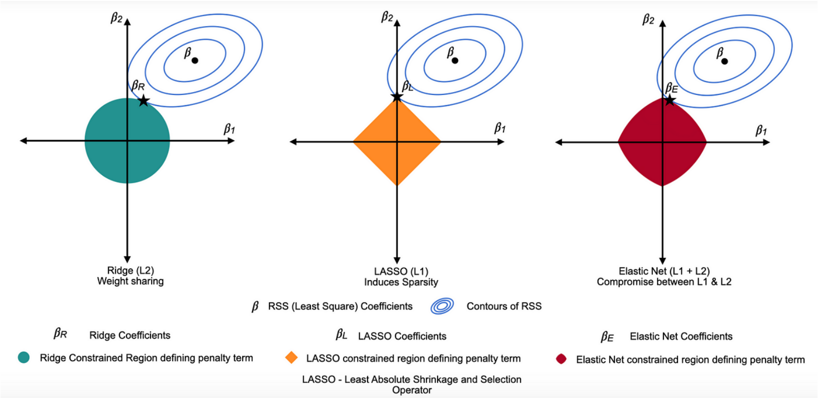


Figure 1: Elastic Net compared to Lasso and Ridge Regression

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

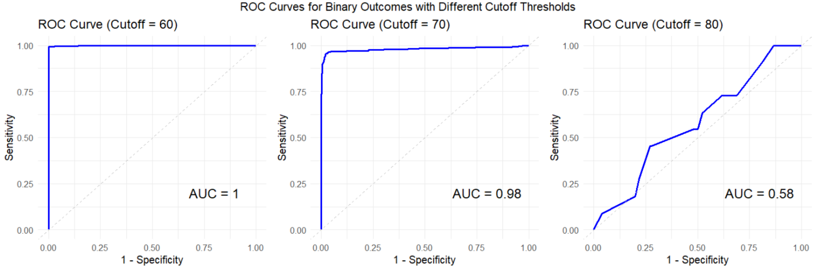
Elastic Net Coefficients

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 2: 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.

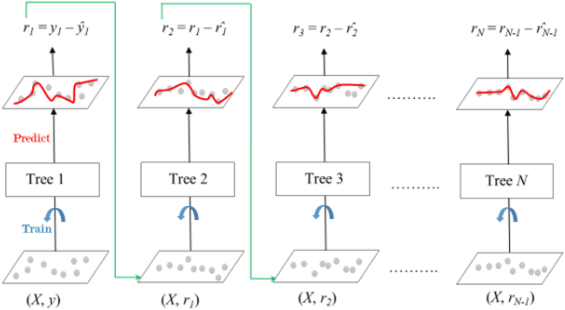
Stochastic Gradient Boosting Machine (GBM) - Gradient Boosting Machine Algorithm

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 3: 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.