Ensemble Regression Tree Analysis

1 Introduction

The purpose of this project is to develop a regression model using an ensemble of decision trees to improve prediction accuracy. The aim is to understand how hyperparameters such as the number of trees and minimum samples required to split a node (`min_samples_split`) influence model performance.

2 Problem Statement

The objective is to solve a regression problem by minimizing the mean squared error (MSE) through optimizing an ensemble of decision trees.

3 Methodology

Model Selection: The base model is a decision tree regressor. An ensemble of these decision trees is created iteratively, where each subsequent tree learns from the residuals of the previous iteration.

Implementation Details: The regression tree was implemented using variance reduction as the metric to determine the best split at each node. The stopping criterion used in the implementation is the minimum number of samples required to split a node (`min_samples_split`). This criterion helps prevent overfitting by ensuring that splits occur only when there is enough data to justify them.

Hyperparameters: The key hyperparameters include:

- **n_trees**: Number of decision trees in the ensemble (set to 5 as per requirements).
- **min_samples_split**: Minimum number of samples required to split a node.

Stopping Criterion: The regression tree stops growing when the number of samples at a node is less than `min_samples_split`. At each split, the variance reduction is calculated as:

$$\text{Variance Reduction} = \text{Var(parent)} - w_l \cdot \text{Var(left child)} - w_r \cdot \text{Var(right child)},$$

where w_l and w_r are the proportions of samples in the left and right children, respectively.

4 Hyperparameter Tuning

We implemented an automatic tuning process for `min_samples_split` using K-Fold cross-validation. The tuning process began with a default value and incremented stepwise until performance improvements fell below a predefined threshold (`epsilon`) for a certain number of iterations (`patience`). This approach avoids overfitting and identifies the optimal parameter.

Optimal Parameters:

- **n_trees**: 5 (fixed)
- **min_samples_split**: Determined automatically through K-Fold validation.

5 Results

The Mean Squared Error (MSE) for each model is as follows:

- **MSE (Sklearn Decision Tree)**: 0.6206453857772722

- **MSE (Custom Decision Tree)**: 0.6206453857772722
- **MSE (Ensemble Regression Tree)** with parameters (`n_trees`=5, `min_samples_split`=optimal): **0.47322202060136365**

The ensemble approach significantly outperformed individual decision tree models.

6 Conclusion

The Ensemble Regression Tree, with optimized hyperparameters (`n_trees`=5, `min_samples_split`=optimal), proved to be effective in reducing prediction error. The use of variance reduction as the metric and setting appropriate stopping criteria (`min_samples_split`) helped ensure that the regression tree model was both accurate and generalized well to new data. Future work could explore other ensemble methods such as Random Forests or Gradient Boosted Trees, and investigate automated hyperparameter tuning.