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, learning rate, and maximum depth 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 using boosting techniques, 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 criteria used in the implementation include the minimum number of samples required to split a node (`min_samples_split`) and the maximum depth of the tree (`max_depth`). These criteria help prevent overfitting by controlling the growth of the tree and 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.
- **learning_rate**: Contribution of each tree's prediction.
- **max_depth**: Maximum depth of each tree.

Regularization Technique: The complexity of the trees was constrained using regularization techniques such as limiting the maximum depth (`max_depth`) of each tree and setting a minimum number of samples required to split a node (`min_samples_split`). By controlling these parameters, we reduced the risk of overfitting, ensuring that the model generalizes well to unseen data. Specifically, after evaluating the performance for different depths, we found that increasing `max_depth` from 10 to 15 provided only a marginal improvement in accuracy. Therefore, a depth of 10 was chosen to maintain a balance between model performance and complexity.

4 Hyperparameter Tuning

We performed manual grid search with the following values:

- **n_trees**: [5]
- **learning_rate**: [0.1, 0.2, 0.3, 0.5]
- **max_depth**: [5, 10, 15]

The optimal parameters are:

- **n_trees**: 5

- **learning_rate**: 0.5
- **max_depth**: 10

Increasing `max_depth` from 10 to 15 did not yield significant improvement, so we chose 10 to avoid unnecessary complexity.

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, `learning_rate`=0.3, `max_depth`=10): 0.4417771598466974

The ensemble approach significantly outperformed individual decision tree models.

6 Conclusion

The Ensemble Regression Tree, with optimized hyperparameters (`n_trees`=5, `learning_rate`=0.5, `max_depth`=10), proved to be effective in reducing prediction error. The use of variance reduction as the metric and setting appropriate stopping criteria (`min_samples_split` and `max_depth`) 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.