Decision Tree Regression for Predicting House Prices

1 Introduction

This document demonstrates how a Decision Tree Regressor works by recursively splitting the data to minimize Mean Squared Error (MSE). We will walk through the process of building a regression tree manually using a small dataset.

2 Dataset

We use the following dataset of house prices with two features: `Size` (in square feet) and `Bedrooms`.

Size (sqft)	Bedrooms	Price ($1000s)
1000	2	300
1200	2	320
1500	3	400
1800	3	420
2000	3	500
2200	4	520

3 Step 1: Define Objective Function

The regression tree splits data to minimize the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

where $\bar{y}$ is the mean target value in the node.
At each split, the algorithm tries to minimize:

$$\text{Weighted MSE} = \frac{N_L}{N}\text{MSE}_L + \frac{N_R}{N}\text{MSE}_R$$

where N_L and N_R are the number of samples in the left and right nodes, respectively.
—

4 Step 2: Compute MSE for Entire Dataset

$$\bar{y}_{\text{root}} = \frac{300 + 320 + 400 + 420 + 500 + 520}{6} = 410$$

$$\text{MSE}_{\text{root}} = \frac{(300-410)^2 + (320-410)^2 + (400-410)^2 + (420-410)^2 + (500-410)^2 + (520-410)^2}{6}$$

$$= \frac{12100 + 8100 + 100 + 100 + 8100 + 12100}{6} = \frac{40500}{6} = 6750$$

So, $\text{MSE}_{\text{root}} = 6750$.

—

5 Step 3: Evaluate Possible Splits on `Size`

We test splits between unique sorted values of Size:

$$[1000, 1200, 1500, 1800, 2000, 2200]$$

Possible split thresholds: 1100, 1350, 1650, 1900, 2100.

—

5.1 Split 1: Size < 1100

- **Left node:** [1000] → Price: [300k]

$$\bar{y}_{\text{left}} = 300, \quad \text{MSE}_{\text{left}} = 0$$

- **Right node:** [1200, 1500, 1800, 2000, 2200] → Prices: [320, 400, 420, 500, 520]

$$\bar{y}_{\text{right}} = 432, \quad \text{MSE}_{\text{right}} = \frac{(320-432)^2 + (400-432)^2 + (420-432)^2 + (500-432)^2 + (520-43}{5}$$

- **Weighted MSE:**

$$\text{MSE}_{\text{after}} = \frac{1}{6} \times 0 + \frac{5}{6} \times 5216 = 4346.67$$

- **Variance Reduction:**

$$\Delta_{\text{var}} = 6750 - 4346.67 = 2403.33$$

—

5.2 Split 2: Size < 1350

- **Left node:** [1000, 1200] → Prices: [300, 320]

$$\bar{y}_{\text{left}} = 310, \quad \text{MSE}_{\text{left}} = 100$$

- **Right node:** [1500, 1800, 2000, 2200] → Prices: [400, 420, 500, 520]

$$\bar{y}_{\text{right}} = 460, \quad \text{MSE}_{\text{right}} = 2600$$

- **Weighted MSE:**

$$\text{MSE}_{\text{after}} = \frac{2}{6} \times 100 + \frac{4}{6} \times 2600 = 1766.67$$

- **Variance Reduction:**

$$\Delta_{\text{var}} = 6750 - 1766.67 = 4983.33$$

—

5.3 Split 3: Size < 1650

- **Left node:** [1000, 1200, 1500] → [300, 320, 400]

$$\bar{y}_{\text{left}} = 340, \quad \text{MSE}_{\text{left}} = 1866.67$$

- **Right node:** [1800, 2000, 2200] → [420, 500, 520]

$$\bar{y}_{\text{right}} = 480, \quad \text{MSE}_{\text{right}} = 1866.67$$

- **Weighted MSE:**

$$\text{MSE}_{\text{after}} = \frac{3}{6} \times 1866.67 + \frac{3}{6} \times 1866.67 = 1866.67$$

- **Variance Reduction:**

$$\Delta_{\text{var}} = 6750 - 1866.67 = 4883.33$$

—

5.4 Split 4: Size < 1900

- **Left node:** [1000, 1200, 1500, 1800] → [300, 320, 400, 420]

$$\bar{y}_{\text{left}} = 360, \quad \text{MSE}_{\text{left}} = 2200$$

- **Right node:** [2000, 2200] → [500, 520]

$$\bar{y}_{\text{right}} = 510, \quad \text{MSE}_{\text{right}} = 100$$

- **Weighted MSE:**

$$\text{MSE}_{\text{after}} = \frac{4}{6} \times 2200 + \frac{2}{6} \times 100 = 1500$$

- **Variance Reduction:**

$$\Delta_{\text{var}} = 6750 - 1500 = 5250$$

—

5.5 Split 5: Size < 2100

- **Left node:** [1000, 1200, 1500, 1800, 2000] $\rightarrow$ [300, 320, 400, 420, 500]

$$\bar{y}_{\text{left}} = 388, \quad \text{MSE}_{\text{left}} = 5168$$

- **Right node:** [2200] $\rightarrow$ [520]

$$\text{MSE}_{\text{right}} = 0$$

- **Weighted MSE:**

$$\text{MSE}_{\text{after}} = \frac{5}{6} \times 5168 + \frac{1}{6} \times 0 = 4306.67$$

- **Variance Reduction:**

$$\Delta_{\text{var}} = 6750 - 4306.67 = 2443.33$$

—

Best Split: Size < 1900 gives the highest variance reduction (5250).

—

6 Step 4: Recursive Splitting

After splitting at Size < 1900, we recursively apply the same logic to each node until stopping conditions are met (e.g., minimum samples per leaf or zero MSE).

6.1 Left Node (Size < 1900)

Contains data points: [1000, 1200, 1500, 1800].

Mean = 360, MSE = 2200. We can try splitting further by Bedrooms or Size.

—

7 Step 5: Pseudocode for Decision Tree Regression

```
function build_tree(data, depth=0):
    if stopping_condition(data):
        return LeafNode(mean(data.y))

    best_split = None
    best_gain = -inf
    for feature in features:
        for threshold in unique_values(feature):
            left, right = split(data, feature, threshold)
            gain = variance(data.y) - weighted_mse(left, right)
            if gain > best_gain:
                best_gain = gain
                best_split = (feature, threshold)
    left_branch = build_tree(left)
    right_branch = build_tree(right)
    return DecisionNode(best_split, left_branch, right_branch)
```

—

8 Conclusion

We manually computed the MSE for multiple splits and found that the optimal first split for this dataset is:

$$\boxed{\text{Size} < 1900}$$

This split gives the highest variance reduction and will be chosen as the root node. The tree then continues recursively to model non-linear relationships between house features and prices.