Variance Reduction for Regression Decision Tree: Splitting on Age = 35

Tran Quoc Thai
Student ID: 2370759

March 22, 2025

1 Introduction

In regression decision trees, the optimal feature split is determined by maximizing variance reduction. Variance quantifies the spread of numerical values in a dataset, and splitting a dataset on a chosen threshold should minimize the variance within each subset. This report presents the calculation of variance reduction when splitting on `Age` at 35 for predicting `CreditScore` and compares this criterion with information gain used in classification trees.

2 Variance Calculation Before Splitting

The variance of a dataset is given by:

$$\text{Var}(S) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{1}$$

where x_i are individual values, $\bar{x}$ is the mean, and N is the number of samples.

From the dataset:

$$\bar{x} = \frac{720 + 650 + 750 + 600 + 780 + 630 + 710 + 640}{8} = 685$$

Variance before the split:

$$\begin{aligned}\text{Var}(S) &= \frac{(720-685)^2 + (650-685)^2 + (750-685)^2 + (600-685)^2}{8} \\ &+ \frac{(780-685)^2 + (630-685)^2 + (710-685)^2 + (640-685)^2}{8} \\ &= \frac{1225 + 1225 + 4225 + 7225 + 9025 + 3025 + 625 + 2025}{8} \\ &= \frac{28600}{8} = 3575\end{aligned}$$

3 Variance Calculation After Splitting on `Age` = 35

Splitting on `Age` at 35 results in:

- **Left subset** (`Age` < 35): CreditScores = $\{650, 600, 630, 640\}$
- **Right subset** (`Age` ≥ 35): CreditScores = $\{720, 750, 780, 710\}$

3.1 Left Subset Variance

$$\bar{x}_{\text{left}} = \frac{650 + 600 + 630 + 640}{4} = 630$$

Variance:

$$\begin{aligned}\text{Var}(S_{\text{left}}) &= \frac{(650-630)^2 + (600-630)^2 + (630-630)^2 + (640-630)^2}{4} \\ &= \frac{400 + 900 + 0 + 100}{4} = \frac{1400}{4} = 350\end{aligned}$$

3.2 Right Subset Variance

$$\bar{x}_{\text{right}} = \frac{720 + 750 + 780 + 710}{4} = 740$$

Variance:

$$\begin{aligned}\text{Var}(S_{\text{right}}) &= \frac{(720-740)^2 + (750-740)^2 + (780-740)^2 + (710-740)^2}{4} \\ &= \frac{400 + 100 + 1600 + 900}{4} = \frac{3000}{4} = 750\end{aligned}$$

4 Variance Reduction Calculation

Variance reduction is calculated as:

$$\text{Reduction} = \text{Var}(S) - \left(\frac{|S_{\text{left}}|}{|S|} \text{Var}(S_{\text{left}}) + \frac{|S_{\text{right}}|}{|S|} \text{Var}(S_{\text{right}}) \right) \tag{2}$$

Substituting values:

$$\begin{aligned}\text{Weighted Var}(S) &= \frac{4}{8} \times 350 + \frac{4}{8} \times 750 \\ &= 0.5 \times (350 + 750) = 550\end{aligned}$$

Variance reduction:

$$\text{Reduction} = 3575 - 550 = 3025$$

5 Conclusion

The variance reduction when splitting on `Age` $= 35$ for predicting `CreditScore` is computed as 3025, indicating a significant improvement in prediction accuracy. This splitting criterion differs fundamentally from information gain in classification trees as it optimizes numerical variance rather than class purity. Understanding this distinction is crucial in selecting the appropriate tree-building strategy for different machine learning problems.