XGBoost

- ## XGBoost for Regression

Problem:

cgpa	package
6.7	4.5
9.0	11.0
7.5	6.0
5.0	8.0

Predict package given cgpa.

⇒ Stage 1: Calculate mean of column 'package'
Mean(package) ≅ 7.3 → Model1

⇒

cgpa	package	model1	residual1 = package - model1
6.7	4.5	7.3	-2.8
9.0	11.0	7.3	3.7
7.5	6.0	7.3	-1.3
5.0	8.0	7.3	0.7

Similarity Score (SS) $= \dfrac{(\Sigma \text{ residuals})^2}{\# \text{residuals} + \lambda}$
(For regression)

λ → regularization parameter (assume = 0)

Node

$\boxed{-2.8, 3.7, -1.3, 0.7} \rightarrow SS_{\cancel{Root}} = \dfrac{(-2.8+3.7-1.3+0.7)^2}{4} \simeq 0.02$

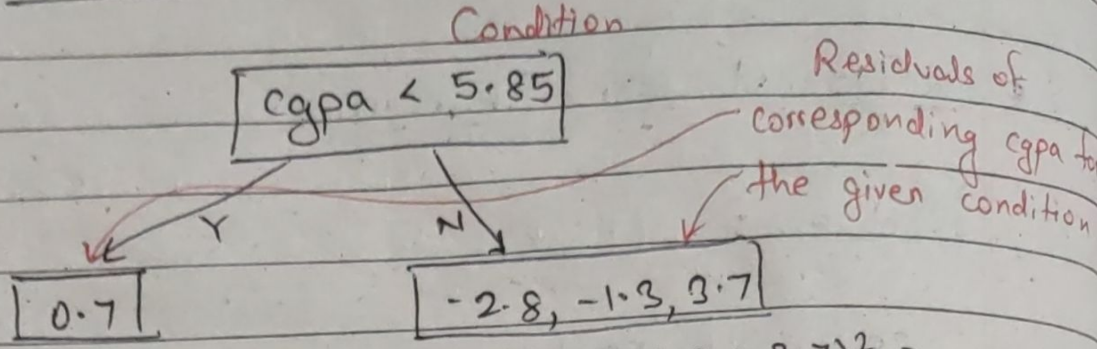
Sort cgpa column -

Avg. of two adjacent cgpa

$$\begin{tikzcd}[row sep=tiny]
5.0 \arrow[dr, dash] & \\
 & 5.85 \\
6.7 \arrow[dr, dash] & \\
 & 7.1 \\
7.5 \arrow[dr, dash] & \\
 & 8.25 \\
9.0 &
\end{tikzcd}$$

Splitting residuals on the basis of 5.85, 7.1, & 8.25 in decision tree & choose the decision tree whose leaf nodes give highest gain in SS.

Splitting Criteria - 1 : 5.85



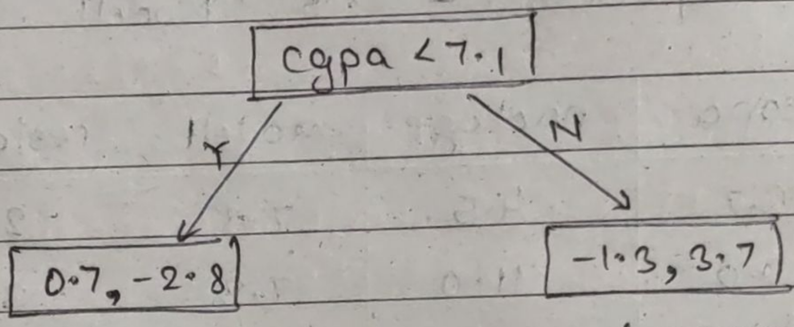
$SS_L = \frac{(0.7)^2}{1} = 0.49$ $\qquad$ $SS_R = \frac{(-2.8 - 1.3 + 3.7)^2}{3} = 0.05$

gain = $(SS_L + SS_R) - SS_{\text{~~Root~~ Parent}}$

$= (0.49 + 0.05) - 0.02$

$= 0.52$

Splitting Criteria - 2 : 7.1



$SS_L = \frac{(0.7 - 2.8)^2}{2} = 2.20$ $\qquad$ $SS_R = \frac{(-1.3 + 3.7)^2}{2} = 2.88$

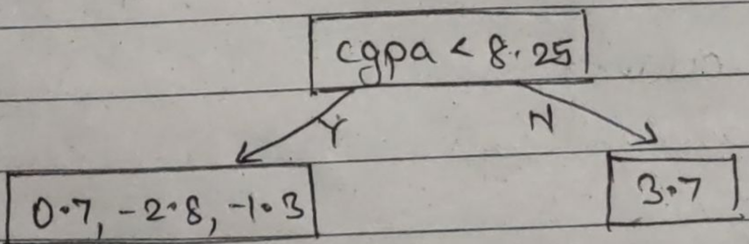
gain = $SS_L + SS_R - SS_{\text{~~Root~~ Parent}}$

$= 2.2 + 2.88 - 0.02$

$= 5.06$

SS increases when similar points go to one side

Splitting Criteria - 3 : 8.25



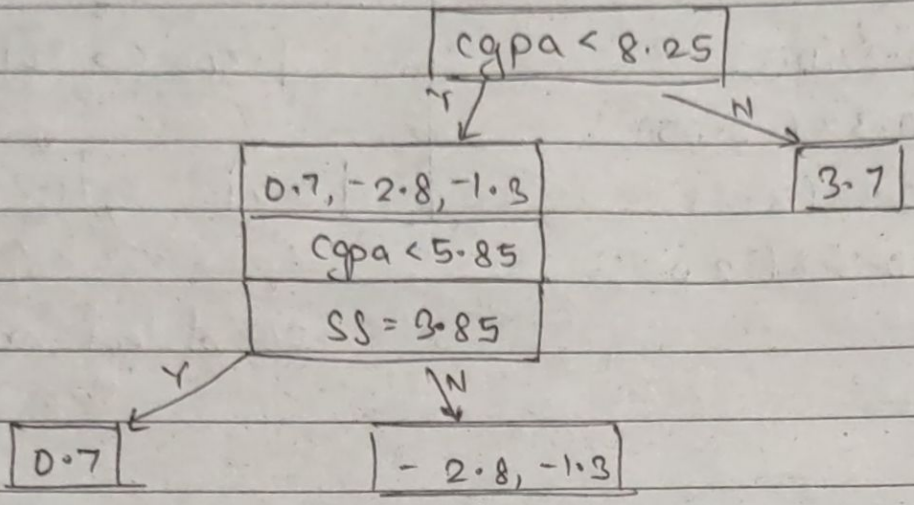
$SS_L = \frac{(0.7 - 2.8 - 1.3)^2}{3} = 3.85$ $\qquad$ $SS_R = \frac{(3.7)^2}{1} = 13.69$

gain = $3.85 + 13.69 - 0.02$

$= 17.52$

Splitting Criteria - 3 is selected. Continuing...

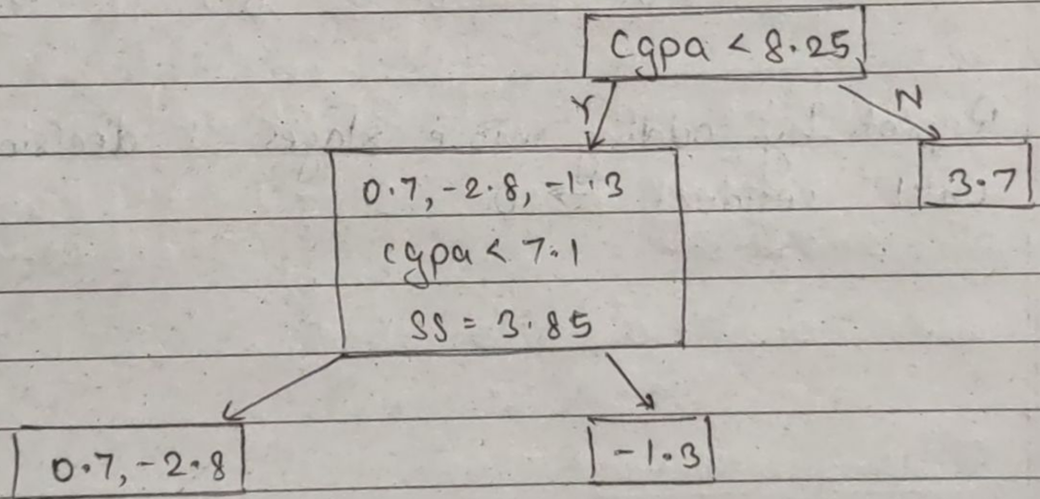
①



$SS_L = \frac{(0.7)^2}{1} = 0.49$ $SS_R = \frac{(-2.8-1.3)^2}{2} = 8.40$

gain = 0.49 + 8.40 - 3.85 = 5.04

②

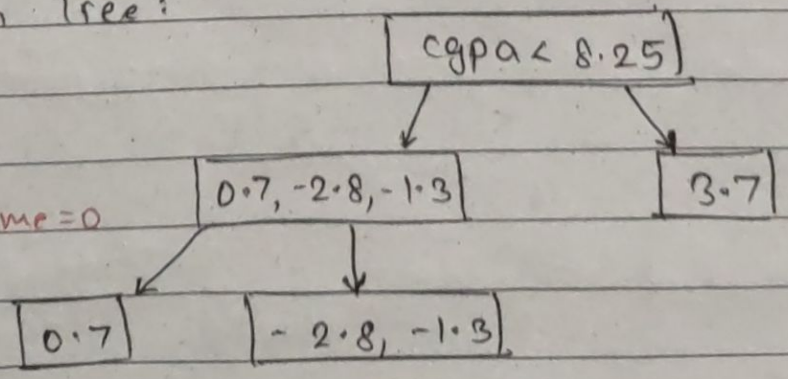


$SS_L = \frac{(0.7-2.8)^2}{2} = 2.20$ $SS_R = \frac{(-1.3)^2}{1} = 1.69$

gain = 2.20 + 1.69 - 3.85 = 0.04

∵ ① > ② ⇒ ① is selected

Final Decision Tree:



$\frac{3.7}{1} = 3.7$

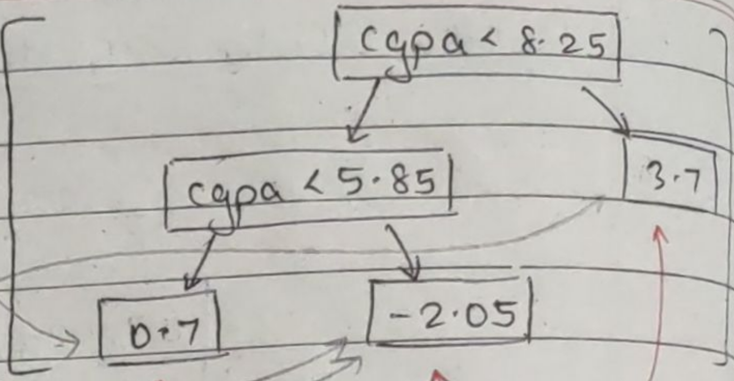
Leaf node o/p:

$\frac{\Sigma \text{ residuals}}{\# \text{residuals} + \lambda}$ assume = 0

⇒ $\frac{0.7}{1} = 0.7$ $\frac{-2.8-1.3}{2} = -2.05$

learning rate
by default = 0.3

⇒ Stage 2: $7.3 + \underset{\text{eta}}{\eta} \times$



$7.3 + 0.3 \times (0.7)$

$7.3 + 0.3 \times (-2.05)$

$7.3 + 0.3 \times (3.7)$

$7.3 + 0.3 \times (-2.05)$

o/p of leaf nodes from last step

⇒ cgpa	package	model1	residual1	model2	residual2
6.7	4.5	7.3	-2.8	6.69	-2.19
9.0	11.0	7.3	3.7	8.41	2.59
7.5	6.0	7.3	-1.3	6.69	-0.64
5.0	8.0	7.3	0.7	7.51	0.49

Repeat by adding more stages & decision trees until residual → 0.