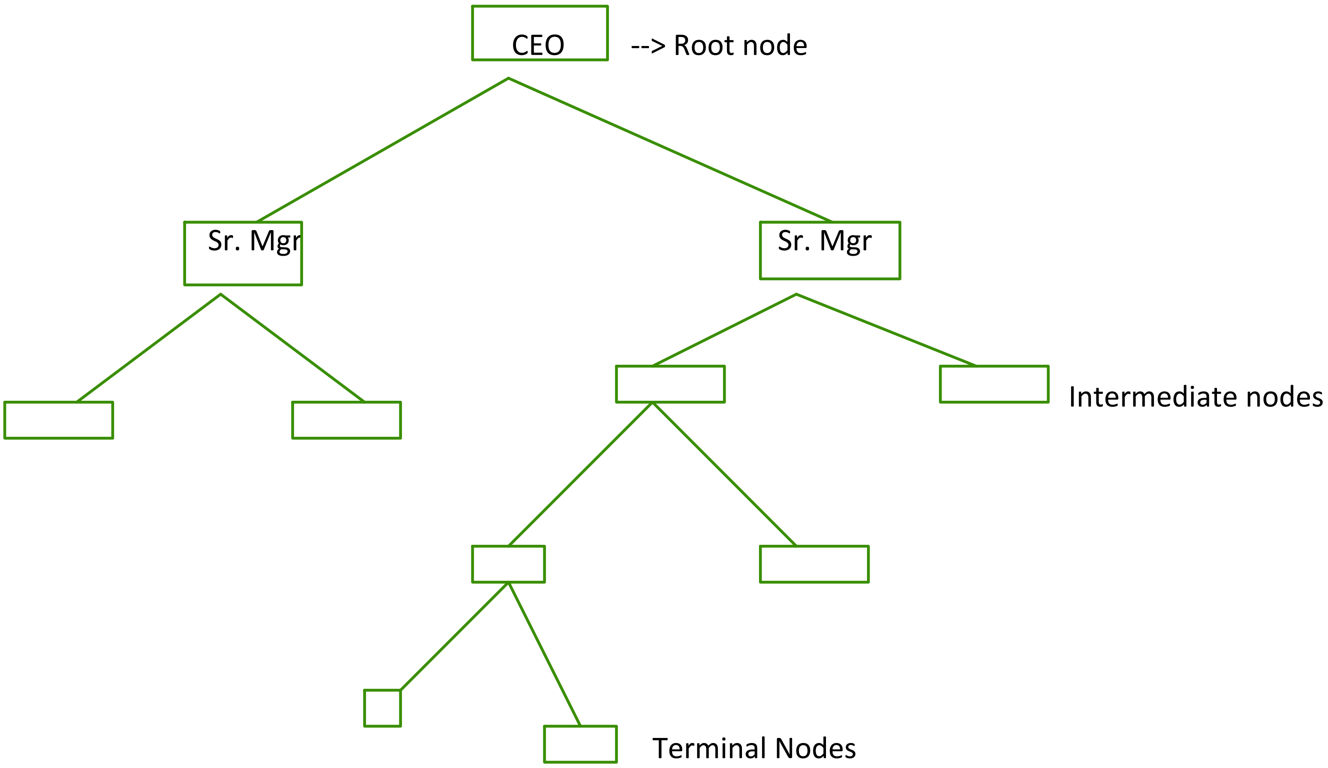
Decision Tree

18 April 2022 11:13

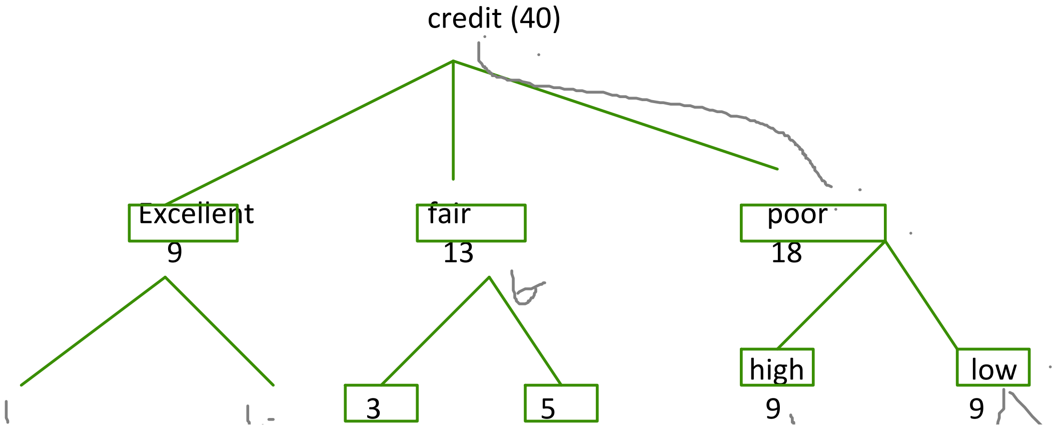
It can be applied for both kind of problems

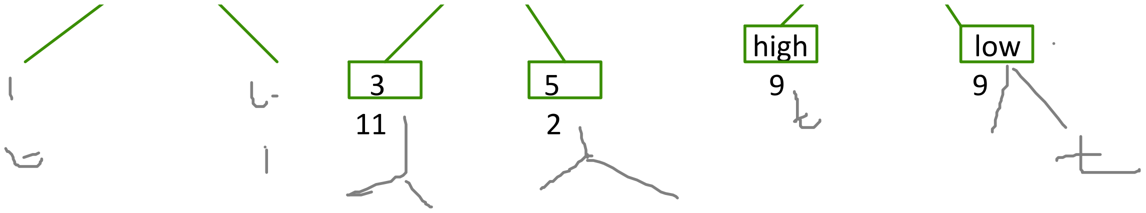
1. Regression technique
2. Classification technique



To identify the Root variables, we have many techniques

1. Miss classification Error (R only)
2. Gini Index (R and Python) --> Classification
3. Information gain / Entropy (Python) --> Classification
4. Mean square error (Python) --> Regression
5. ID3
6. CHAID
7. Variance reduction technique
8. C4.5





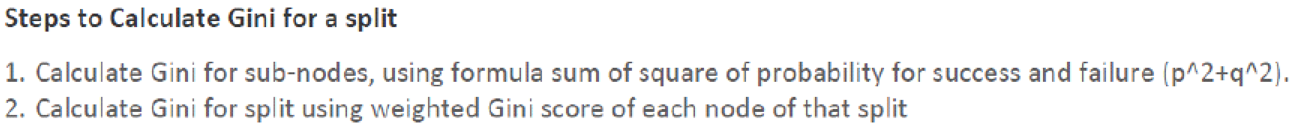
Gini Index

Gini =

$$\sum_{i=1}^{C} (p_i)^2$$

$\Sigma\ Wi(p2 + q2)$

Gini index will be calculated for each of the X variable and will see which X variable contains highest Gini index that variable will become Root Variable.

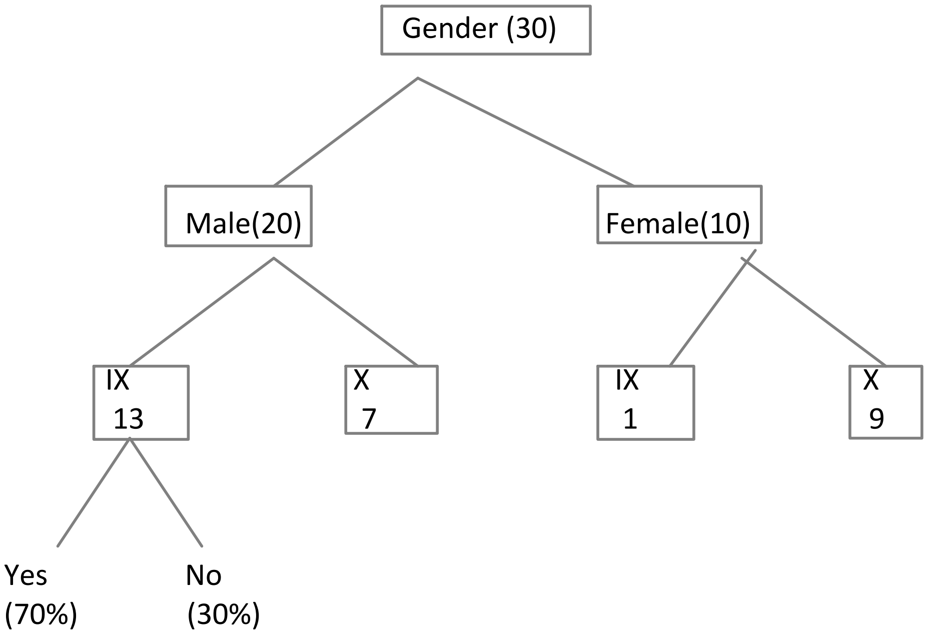


Steps to Calculate Gini for a split

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node of that split

3. To calculate the exact Gini score by multiplying Total(step 1 * step 2)

We have to verify same calculative method for each of the x variable and finalized with one of the X variable.


Gender (30)
Male(20)
Female(10)
IX
13
X
7
IX
1
X
9
Yes
(70%)
No
(30%)