

# Handling Missing Values

11 April 2024 16:44

Xgboost

	Height	Weight	Country	Place	Number of days	Some column
0	12.0	35.0	India	Bengaluru	1.0	NaN
1	NaN	36.0	US	New York	2.0	NaN
2	13.0	32.0	UK	London	NaN	NaN
3	15.0	NaN	France	Paris	4.0	NaN
4	16.0	39.0	US	California	5.0	12.0
5	NaN	NaN	NaN	Mumbai	NaN	NaN
6	NaN	NaN	NaN	NaN	6.0	NaN

missing data unavailable

majority ML

pragmatic → feature imputation → additional

Xgboost → missing  
 Sparsity aware split finding → Xgboost

LightGBM  
 Catboost  
 DT → sklearn

tree based

f	t	pl	res1
1	10	30	20
?	20	30	10
3	30	30	0
?	40	30	-10
5	50	30	-20

ASP

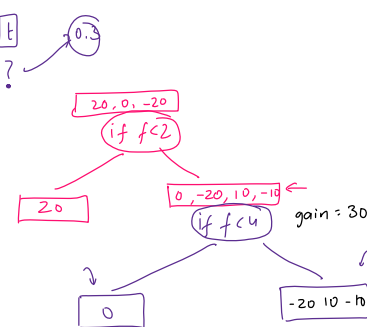
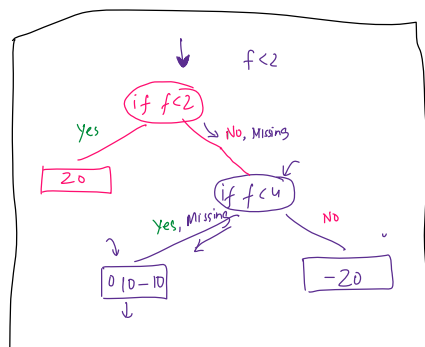
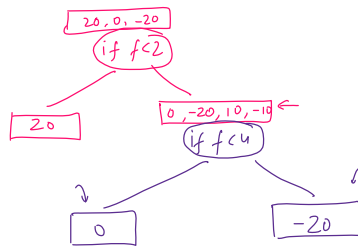
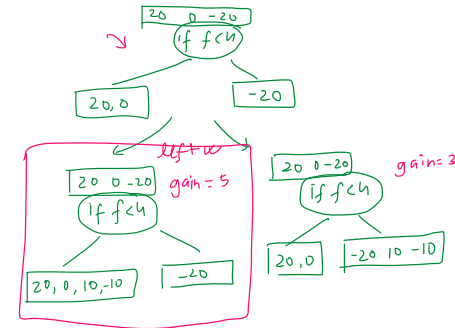
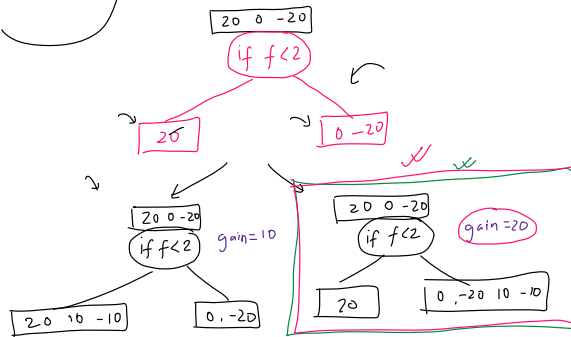
f	t	pl	res1
1	10	30	20
3	30	30	0
5	50	30	-20

non-missing

f	t	pl	res1
?	20	30	10
?	40	30	-10

missing

dt



node missing direction

$$\frac{\text{sum of res}}{\# \text{ of res} + 1} = 0.3$$

sparse → missing

missing = np.nan →

"1"

sparse → missing

missing = np.nan →

= 0

= -99999 ←

small → default handling →  $X_{gboss}$  → outcome  
[custom handling →]

Too much help is required

}

$\alpha$   $\beta$   $\lambda$

}

Hi guys  
just a test  
new setup

→ super display

$X = 2x + y$  →

