Handling Missing Values

11 April 2024 16:44

	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

Xgboost

missing data
unavoidable

majority
ML

preprocess → feature imputation → additional

Xgboost → missing
Sparsity aware split finding → XGboost

LightGBM
Catboost
DT → sklearn

tree based

SASP

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

non-missing

missing

dt

20 0 -20
if f<2
20
0 -20

gain=10
20 10 -10
0, -20

gain=20
0, -20 10 -10

gain=5
20, 0, 10, -10
-20

gain=3
-20 10 -10

20, 0, -20
if f<2
20
0, -20, 10, -10
if f<4
0
-20

f<2
Yes
No, Missing
if f<4
Yes, Missing
No
0 10 -10
-20

f t
? ?
0.3

gain = 30

node missing direction

Sum of res / # of res + λ = 0.3

sparse → missing

missing = np.nan →

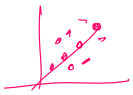
sparse → missing

missing = np.nan →
= 0
= -99999

small → default handling → xgboost → outcome
[custom handling →]

too much help is required ≷ α β λ

Hi guys
just a test → super display
new setup X = 2x+y →



≷