07-09-2025

Agenda:—

FE-IV

Data Encoding

label encoding, Ordinal Encoding

One hot encoding

(1)	Red	400 – 480 Thz
(2)	orange	480 – 510 Thz
(3)	Yellow	510 – 530 Thz
(4)	Green	530 – 600 Thz
(5)	Blue	600 – 670 Thz
(6)	Indigo	670 – 700 Thz
(7)	violet	700 – 750 Thz

10 20 30 40 50 60 70

(sort alpha betic) → sort by order

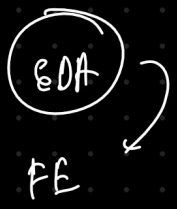
Color-freq	category of color	LE	LE(Ordinal)	B_OHE	R_OHE	I_OHE	V_OHE	O_OHE
625	Blue	0	2	1	0	0	0	0
420	Red	3	0	0	1	0	0	0
680	Indigo	1	3	0	0	1	0	0
690	Indigo	1	3	0	0	1	0	0
710	violet	4	4	0	0	0	1	0
490	orange	2	1	0	0	0	0	1

Color-freq	category of color	Target_Encoding
625	Blue	value_3
420	Red	value_2
680	Indigo	value_1
690	Indigo	value_1
710	violet	value_4
490	orange	value_5

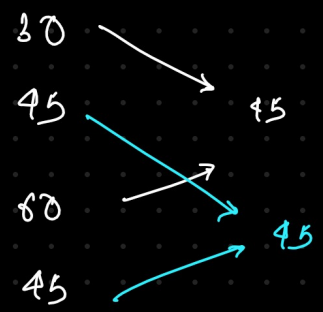
Indigo → 680, 690 → mean → value_1

orange → 490 → mean → value_2

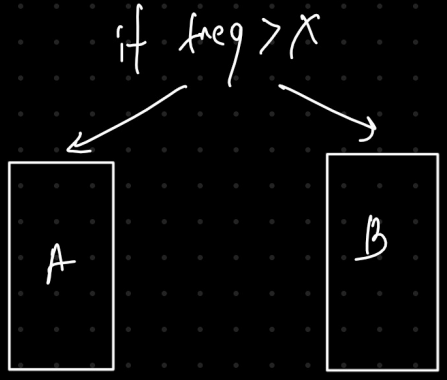
⋮

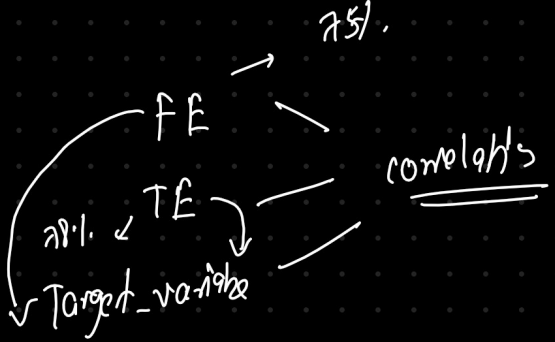


Grade	Marks
A	—
B	—
A	—
B	—
B	—
C	



Color-freq	category of color	Frequency_Encoding
625	Blue	1
420	Red	1
680	Indigo	2
690	Indigo	2
710	violet	1
490	orange	1

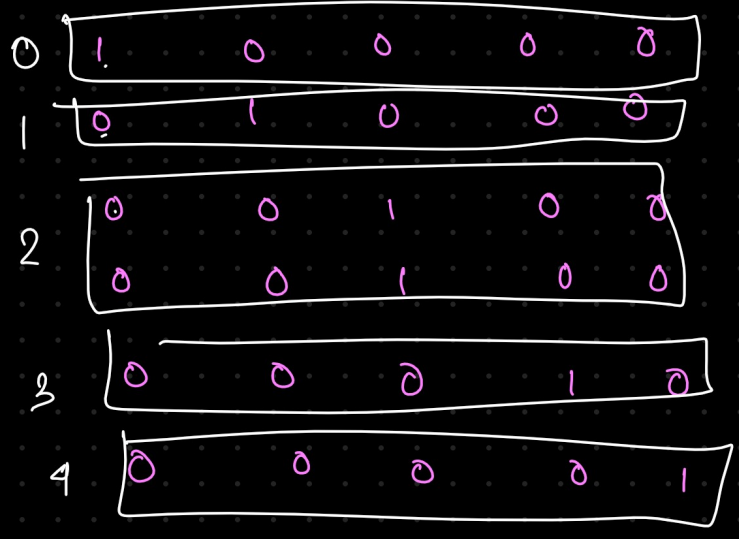




OHE → sklearn → sparse matrix

B_OHE, R_OHE, I_OHE, V_OHE, O_OHE

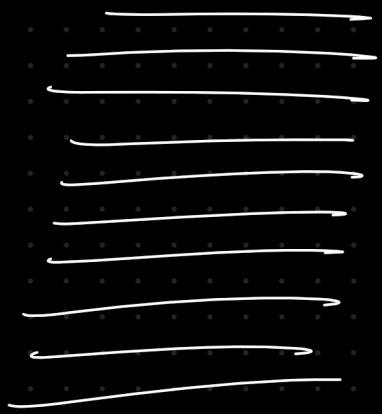
coords	values
0, 0	1.0
1, 1	1.0
2, 2	1.0
3, 3	1.0
4, 4	1.0



1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

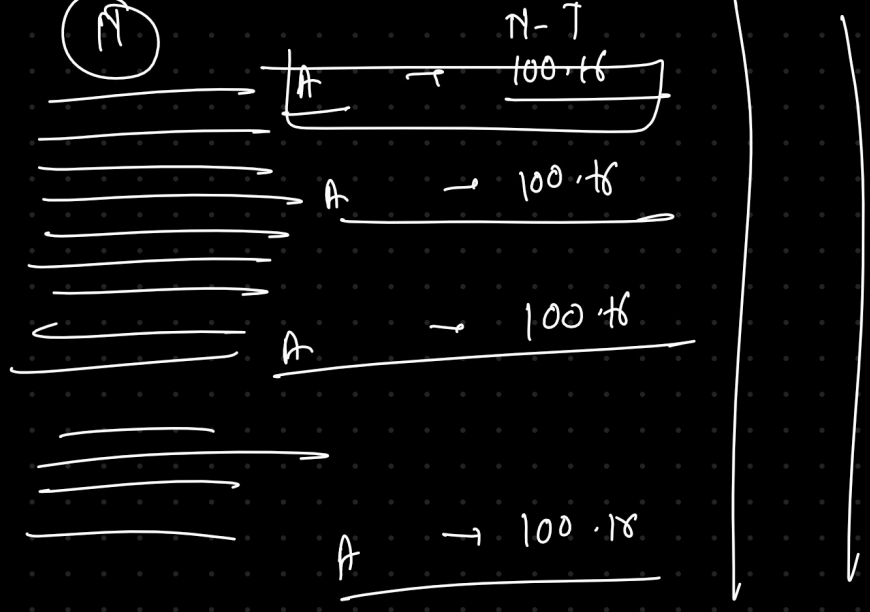
sparse matrix → OHE

A B C



→ unique

Drop all rows



Feature selection :—

df → 80 columns
↓
100 colum

ML

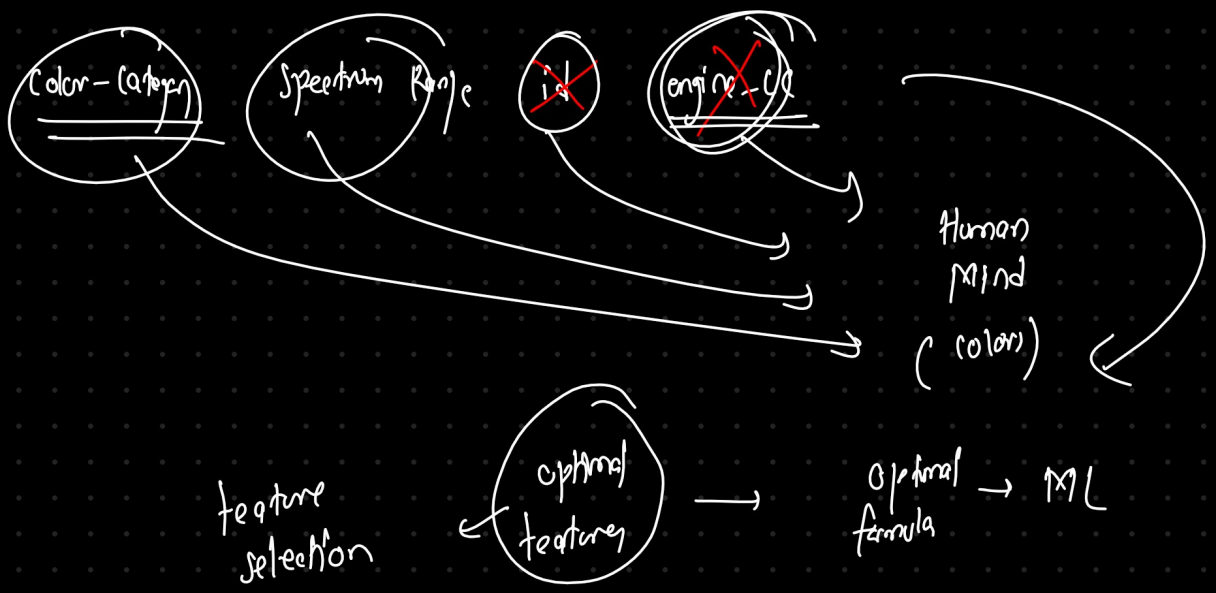
$m_0 + m_1 x_1$ → linear regression

↗
1 feature

$m_0 + m_1 x_1 + m_2 x_2 \dots + m_{100} x_{100}$

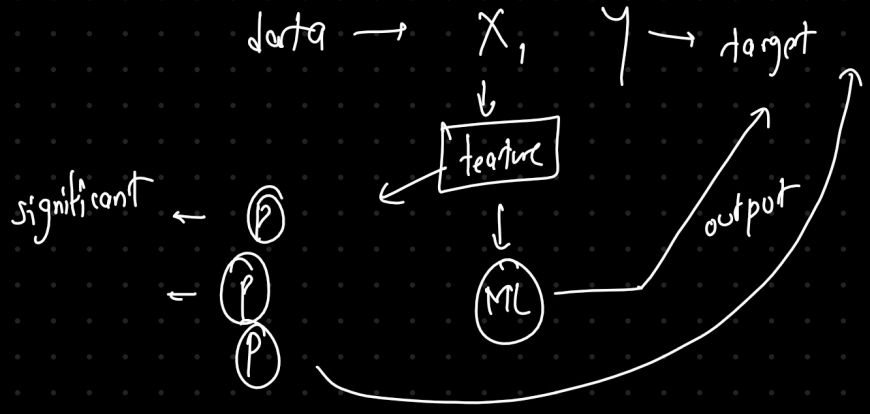
no feature column
↓
Constant
↓ Intercept

$y = mx + C$



Feature selection :—

(1) Backward Elimination (based p-values in regression)



(1) data

(2) data → X, y
add constant

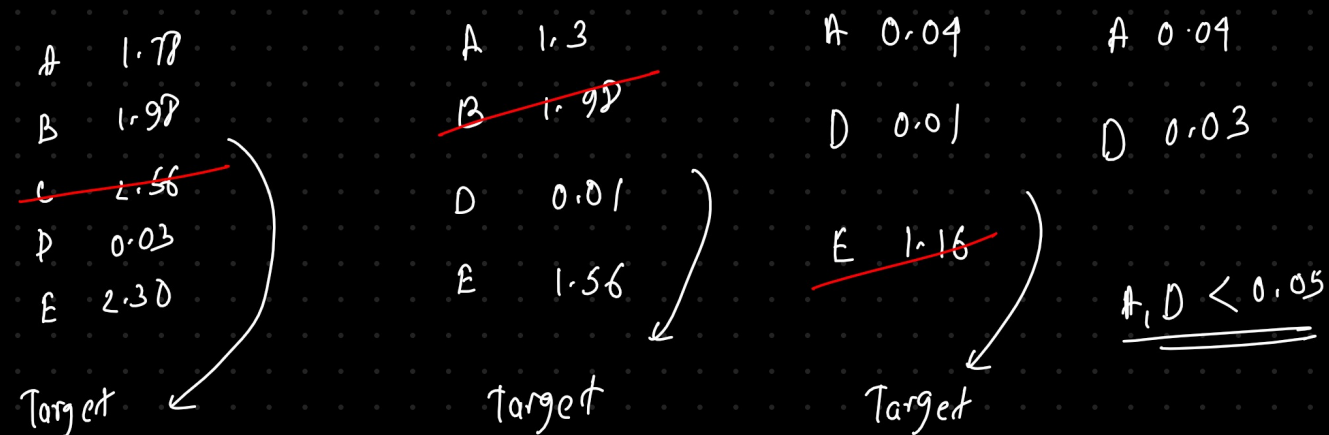
(3) (X) → ML ← y (learn)

(4) ML → p-value for all feature-column of X

(5) check highest p-value column (least significant feature)

(6) remove that column from X

(7) repeat step 3-6 until all high p-value column are removed.



(2) Forward Selection

(1) data → data.columns (collect column names)

(2) X, Y (X → 1000 rows, 12 columns)

(3) (only constant) $\xrightarrow{X}$ ML ← Y
(1000 rows, 1 column)

(4) ML → R^2 of all columns
(bigger is better)

(5) get the name of a column having highest R^2 value

(6) add any 1 column from step 1 to X

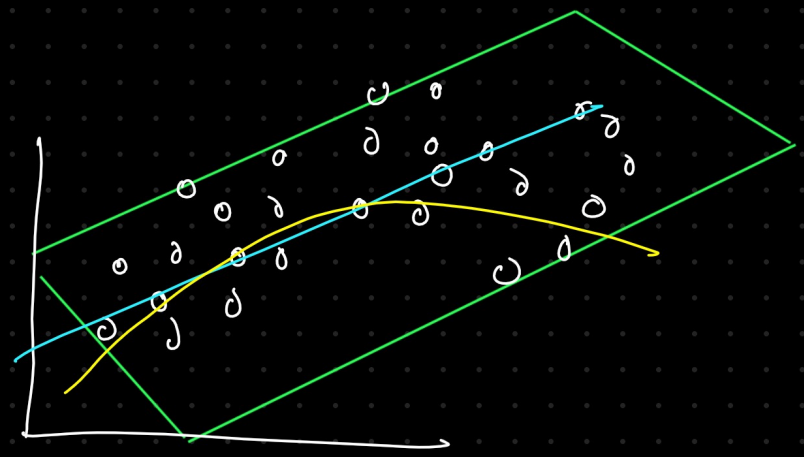
(7) repeat step 3 & 6 until all columns are done

x_0 (constant)	A	B	C	D
1	Blue	128	11.12	23
1	Blue	162	11.34	22
1	Red	3000	36.86	21

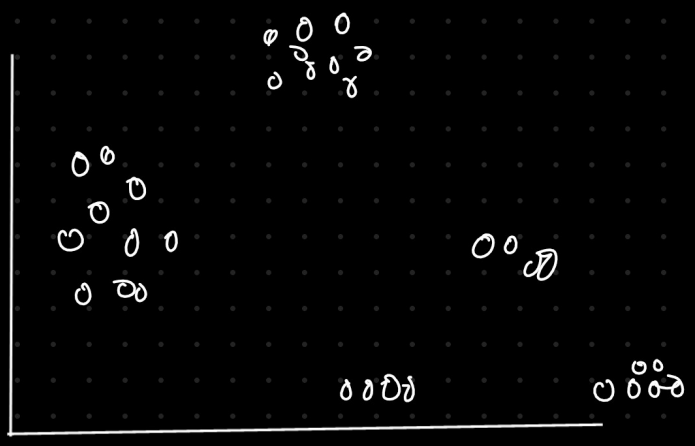
X → ML

$m_0 + m_1x_1 + m_2x_2 + m_3x_3$

data



linear data



Non linear data

Tree based model

Random forest → important

ML

D.I

ML