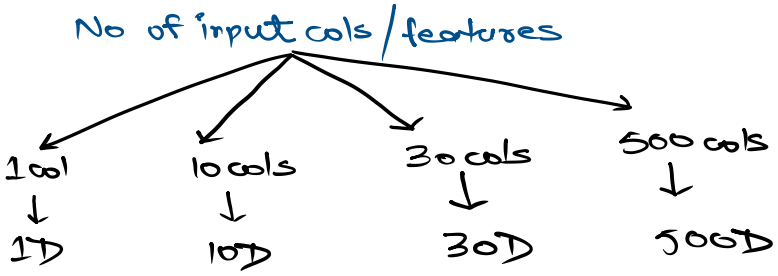
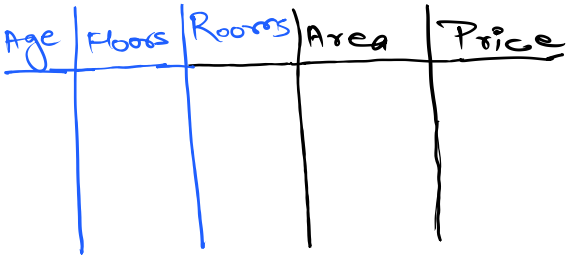
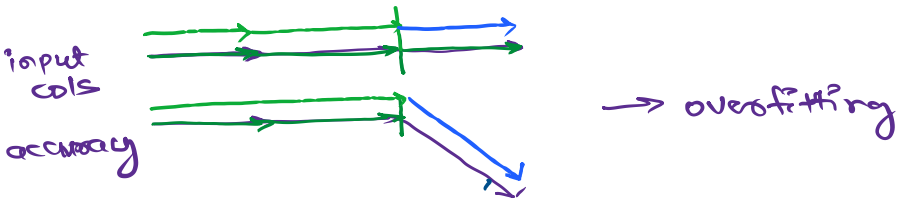
Principal Component Analysis(PCA) 20 April

17 August 2024 11:39





Curse of Dimensionality



When you have too many input cols, then there are high chances that some/many of these cols are unnecessary, because of this we start to see overfitting in the data. This problem is what we refer as Curse of Dimensionality.

↓

There are also high chances of multi-collinearity.

↓

High corr. b/w input cols

Dataset → 200 cols 200▷

↓

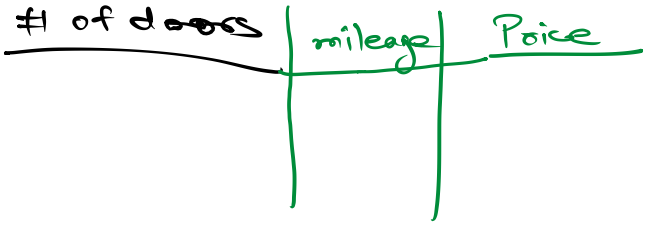
28 cols → unnecessary → Delete them

↓

172 remaining cols

↓

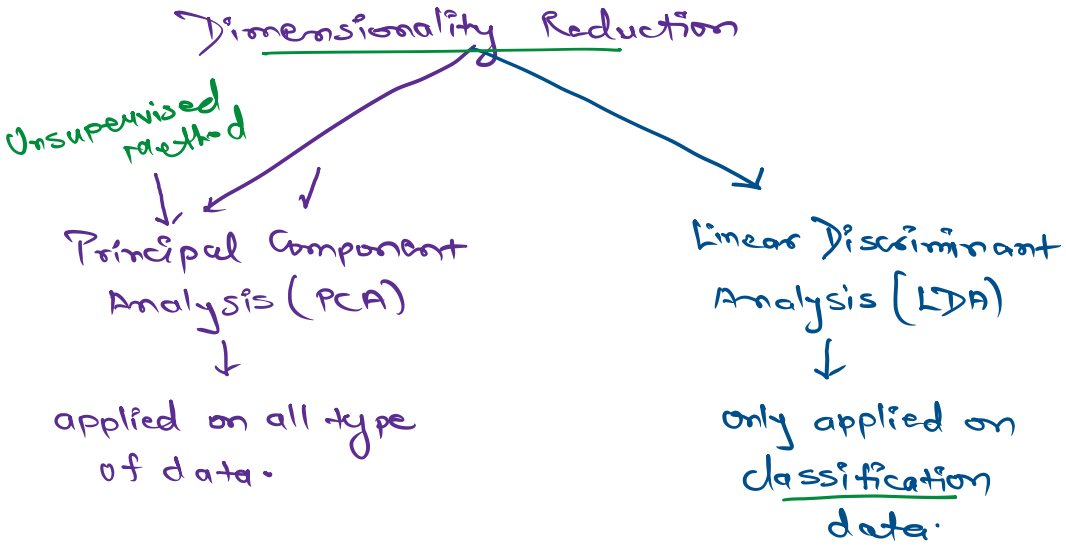
56 cols are very less important



Solution: If we have some method which can reduce the number of input cols in a smarter way, rather than directly deleting it, that would be a very good solution.

↓

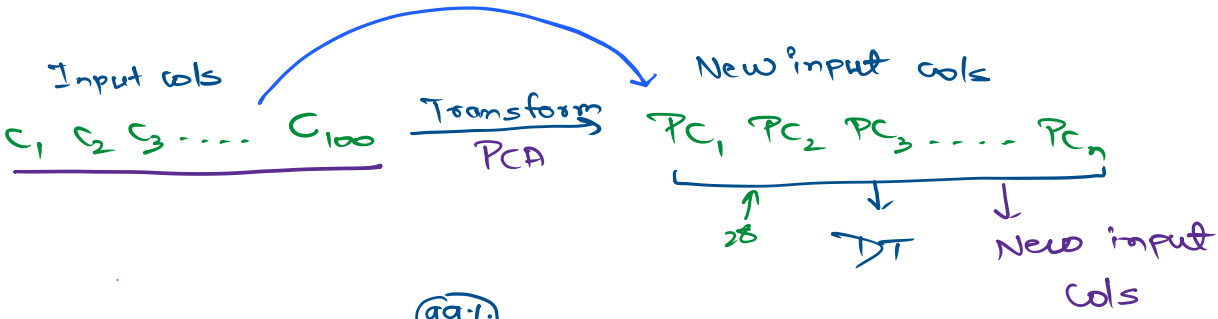
This is where 'Dimensionality Reduction' technique come into picture.

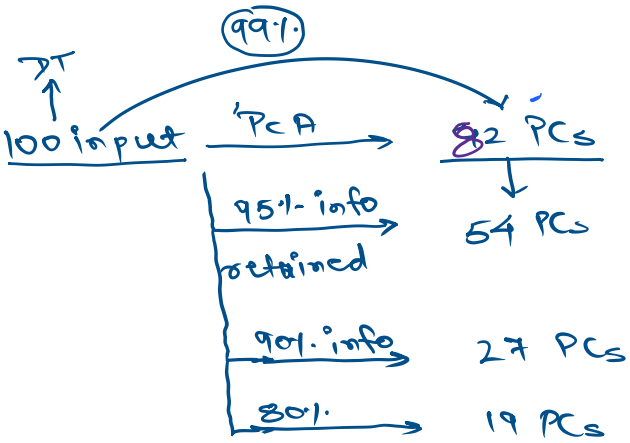


Principal Component Analysis →

PCA helps reduce the dimension of the data in such a way that we loose very little information while reducing the no of input cols.

$$2.5C_1 + \frac{1}{17}C_{22} + 3C_3$$



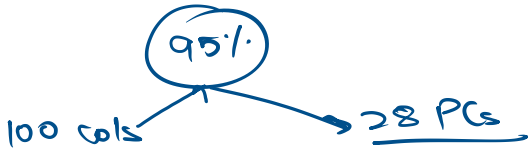


'information contained in PCs decrease from PC_1 to PC_{28} →

PC_1 PC_2 PC_3 PC_4 . . . . . PC_{28}

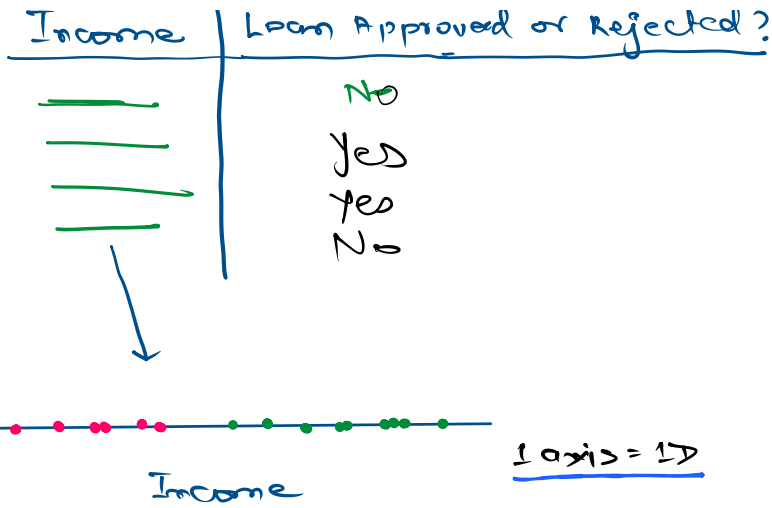
combination of two most imp. input cols.

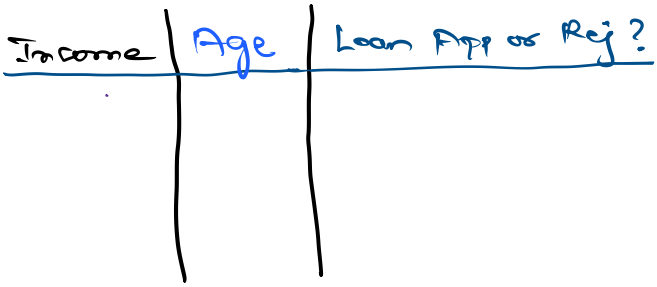
(27%)

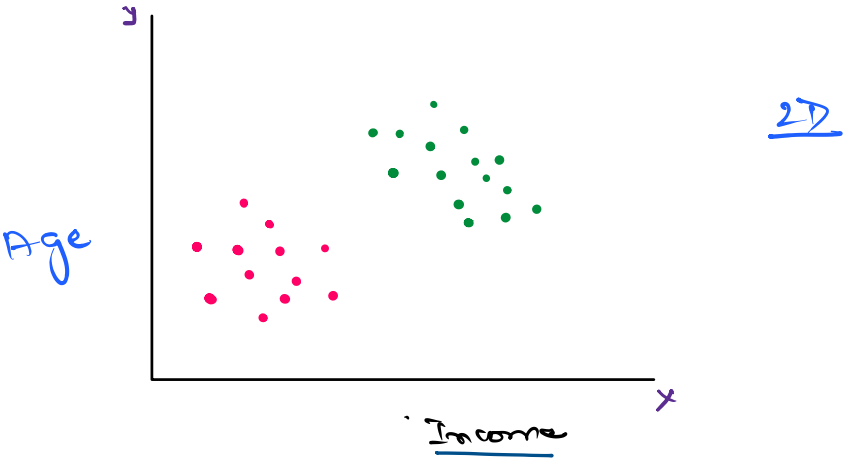


Mathematical Intuition of PCA →

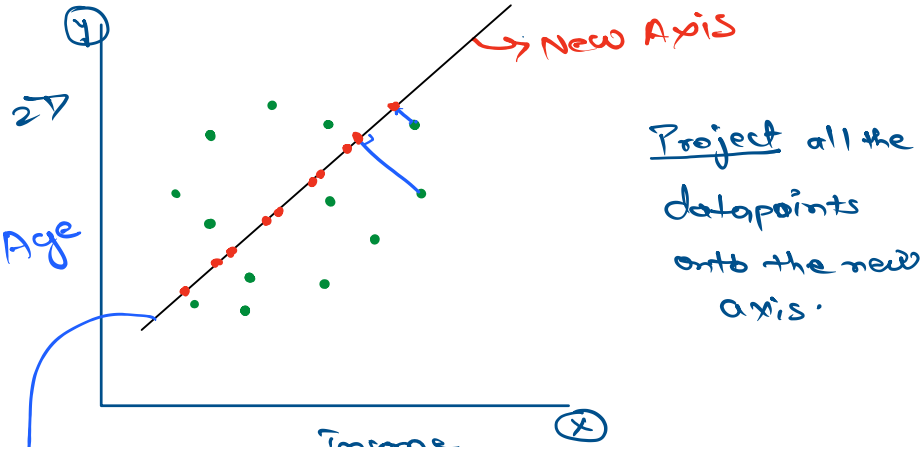
Inp ↓ Out ↓

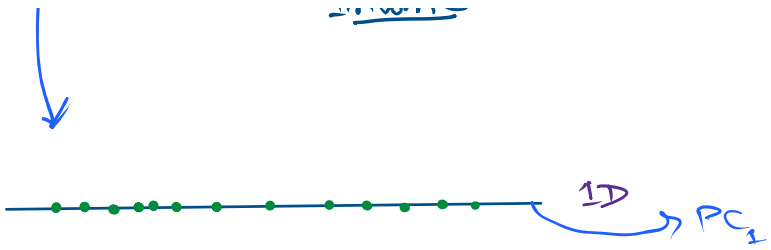






How to reduce the dimension using PCA →

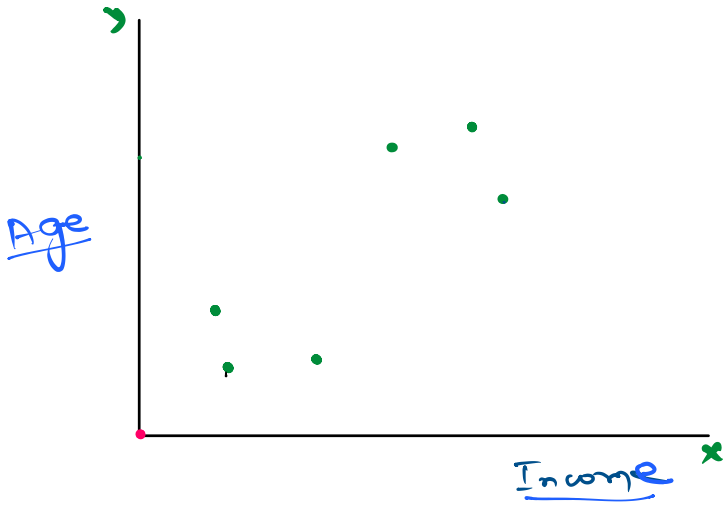




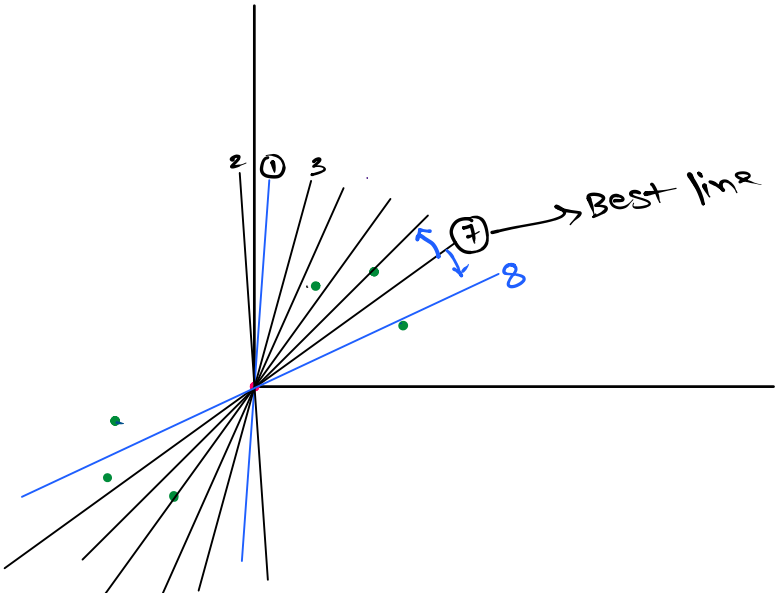
Projection of datapoints onto the new axis is nothing but linear combination of existing cols / dimension.

How to find proper new axis?

Mean centering of the data

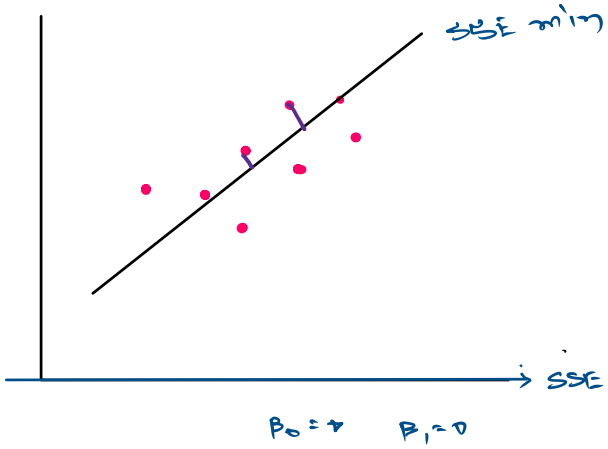


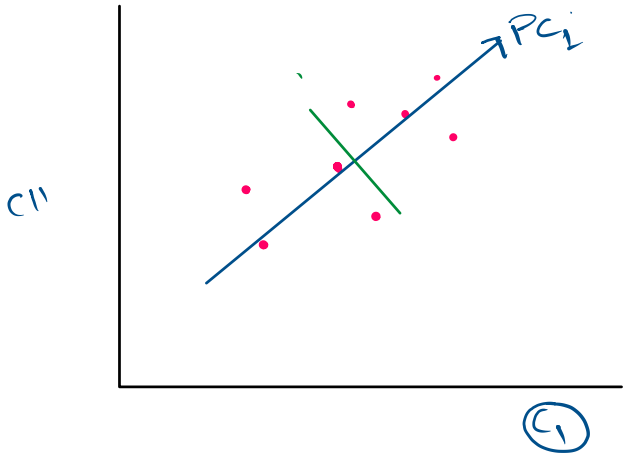
We have to move the data in such a way that its center is at 'origin'.



$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 \quad (SSD)$$

Any line which minimizes this sum will be our new axis.





PC_1 will be created in the direction of maximum variation.

mileage	# of doors	Price
17	4	
23	2	
18.6	4	
12	4	
13	2	
30	4	

Standardization →

Linear Regression

x_1 = # of rooms → 4

x_2 = # of floors → 2

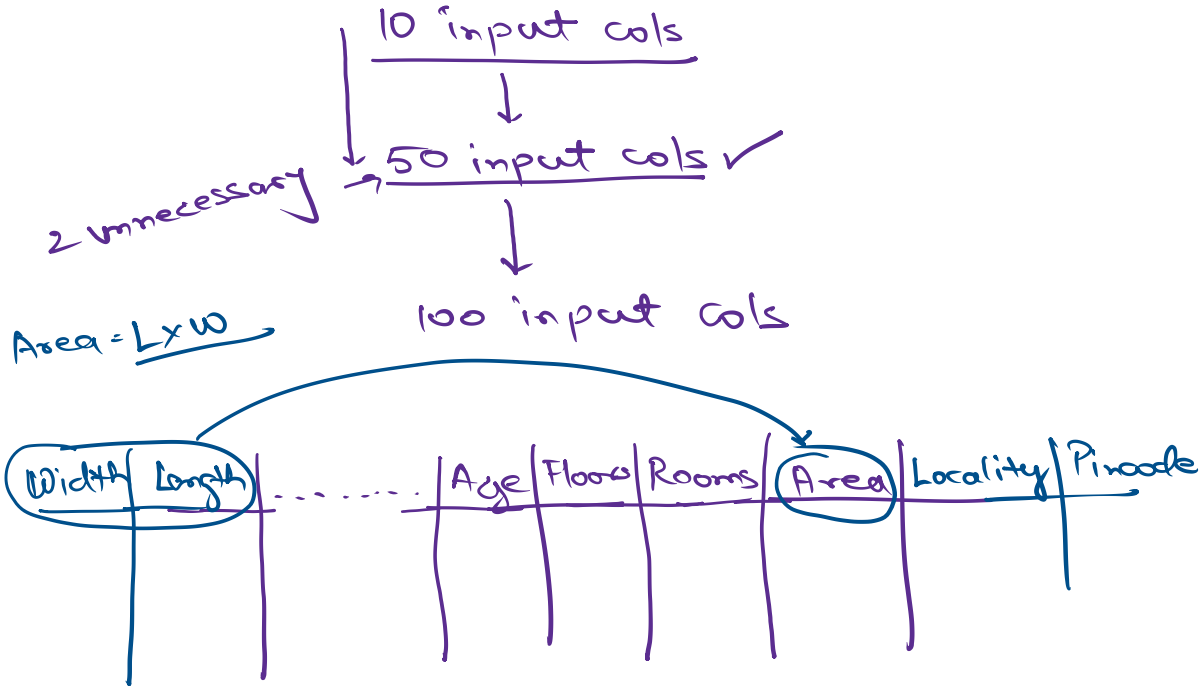
x_3 = Area of house. → 2000

$y\text{-pred} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$= 1.2 + 0.9 x_1 + 0.3 x_2 + 2.1 x_3$

$= 1.2 + 0.9 \times 4 + 0.3 \times 2 + 2.1 \times 0.2$

$= >4200$ crore.



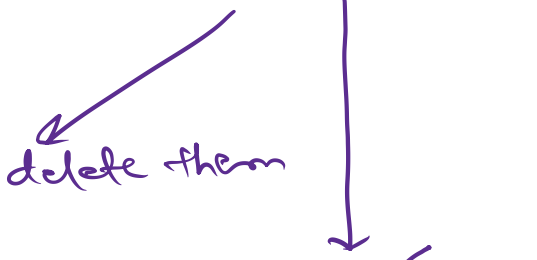
Curse of Dimensionality

Unnecessary & redundant cols

These will create noisy data

Less important

5 less important



Dimensionality Red.

$C_2 + C_7 + C_8 + C_{23} + C_{31}$ = New Col

→ PCA → unsupervised

C_1 C_2 C_3 C_4 ---- C_{100} ↓ output

↓

35