Principle of Component Analysis (PCA)

* Curse of dimensionality
  ⇓
  features.

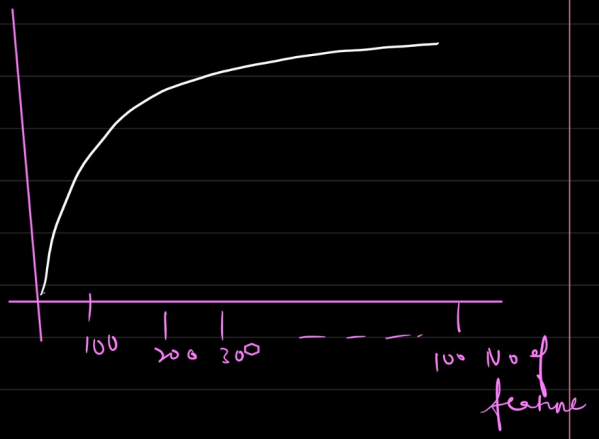
To predict price of house

f_1	f_2	...	f_{100}	y

20f	50f	100f	250f	500f	750f	1000f
M1	M2	M3	M4	M5	M6	M7

Acc Rsquare < Acc Rsquare ↑ < Acc Rsq ↑ < Acc Rsq ↑ ≈ Acc Rsq ≈ Acc Rsq ≈ Acc Rsq.

* With increase in no of features, after one point of time the Acc (rsquare) will not increase Rsq in that proportion.



⇓

Why?

→ Few of features will be multicollinear ($f_1 f_2 f_3 \approx f_4$)

→ few of features might be exactly same

f_1 — 1, 1, 1, 1, ... 1

f_2 — 1, 1, 1, ... 1

→ No variance / information in feature.

→ lots of duplicate entries

f_1
1
1.01
1.02
⋮

* With increase in no of feature performance of model degrades.

Execution time

Evaluation metrics

Analogy

* you went to buy a house.

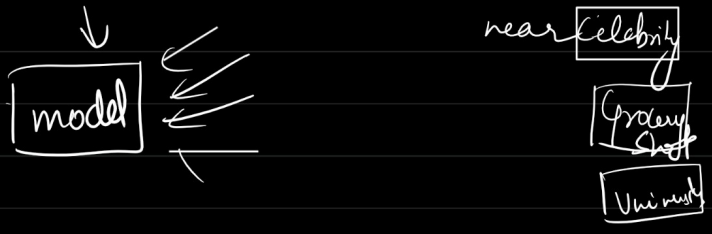
Broker

2 BHK → 60 Lakhs

3 BHK → 80 Lakhs

beach → ↑↑

Airport → ↑↑



* Curse of dimensionality → With increase in no of feature the performance of model degrades.

To remove COD :-

① Feature selection ② Feature Extraction

↓

PCA (Dimensionality reduction technique)

* Why we should remove COD ?

① To improve the performance of Model.
② Visualise the data → for insights
③ Prevents from overfitting.
④ Better interpretation.
⑤ To remove Curse of dimensionality

1000d → 3d

* Feature Selection

Area of house (X_1)	Near to airport (X_2)	Price of house y

① Correlation (Pearson)

$$Corr = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

$x \uparrow y \uparrow$
$x \downarrow y \downarrow$
$x \uparrow y \downarrow$
$x \downarrow y \uparrow$

② $$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) * (y - \bar{y})}{N-1}$$

$Corr(X_1, y) = 0.82$

$Cor(X_2, y) = 0.3$

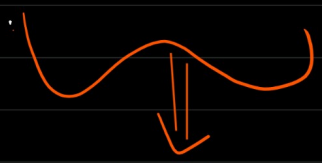
f_1 f_2 f_3 — — — — f_{1000} → y

0.82 -0.83

* Feature Extraction

(In feature Selection, you dropped the features. But let's if you want all the features and also curse of dimensionality, then comes the technique feature extraction)

Room Area	No. of rooms	Price of house



data transformation to extract New feature which represents both the the features.

House Area (using Domain expertise)	Price

but say → #of bathroom | distance from airport | distance from University.

Principal Component Analysis

→ Dimensionality reduction technique.