Principle of Component Analysis (PCA)

* Curse of dimensionality
  ⇓
  features.

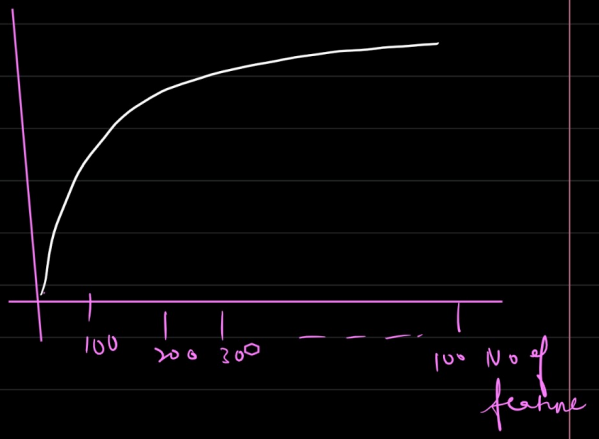
To predict price of house

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.

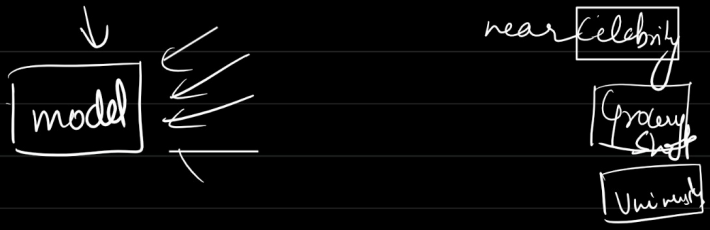
Broker

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)