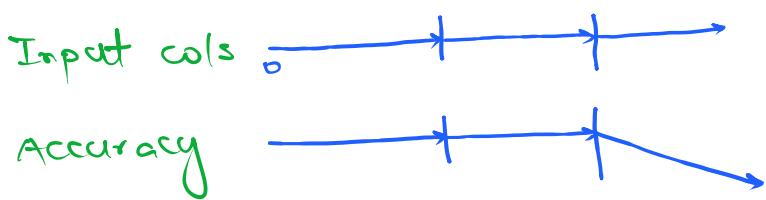


Locality Pincode
High correlation

Length width Area independent
High correlated

Curse of Dimensionality



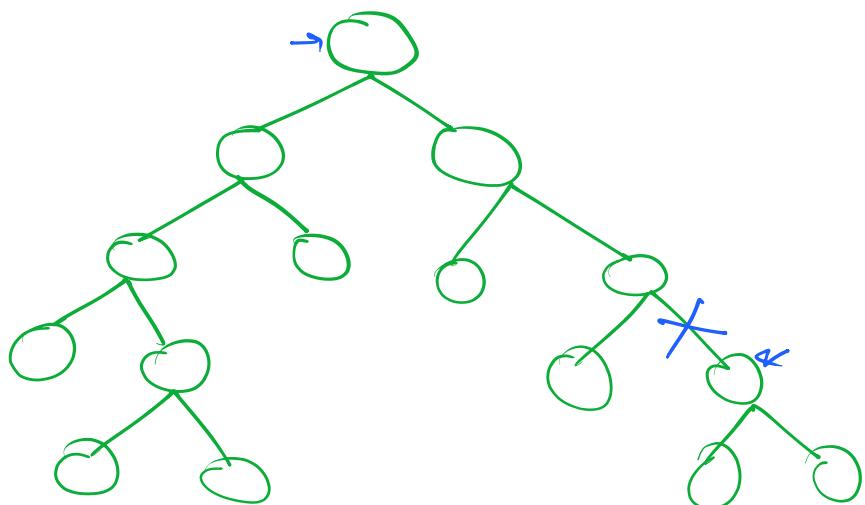
overfitting
↓
Unnecessary
cols
which are basically

Multicollinearity
↓
Redundant
features

Noise in the data.

Solutions to deal with the curse of Dimensionality →

- ① Hyperparameter Tuning.
- ② VIF (variance inflation factor) for reducing multicollinearity.
- ③ RFE (Recursive Feature Elimination)
- ④ Tree based algorithms like (DT, RF, XGBoost) have inbuilt feature selection methods.



Dimensionality Reduction
Techniques

Principal Component Analysis (PCA)

can be applied
on all types of
data.

Linear Discriminant Analysis (LDA)

can be applied
only on
classification
data.

PCA: Instead of deleting the cols to avoid the curse of dimensionality, let's combine given input cols to create new more powerful input cols which are much less in number compared to original input cols.

Data(100 input cols)

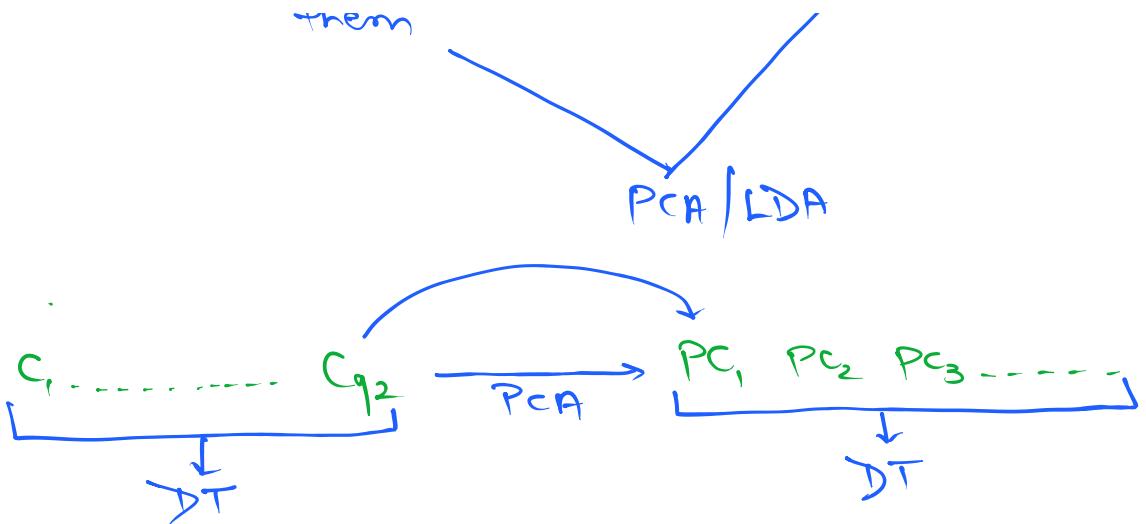
8 unnecessary cols
(delete it)

92 cols

Domain knowledge
+
DS Tools

46 cols have
very less info
contained in
...

46 cols
are more
imp input cols
/



$$C_1 + C_{35} + C_{48} = PC_1$$



Principal components are linear combination of existing input cols.

$$ax + by + cz$$

$$2 \cdot PC_1 + 3 \cdot 8 \cdot C_{35}$$

	C_1	C_{35}	
Rooms	2	1	PC_1
Floors	2	4	(3)
2			
3			
8			
6			

Loss of interpretability / explainability
in the data after PCA transform.

- (99.1%) - i.e. 99.1%