Algorithms

Regression : LR

Classification : Logistic Regression
KNN
Naive Bayes
Decision Trees

C
R

Classification Metrics

Regression : R2 score

?

→ way to measure the quality

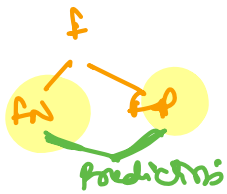
S1 — M1 (LR)

S2 — M2 (DT)

Confusion Matrix

→

(Actual) True label	Predicted label: 0	Predicted label: 1
0	True Negatives (TN)	False Positive (FP)
1	False Negative (FN)	True Positive (TP)



$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}$$

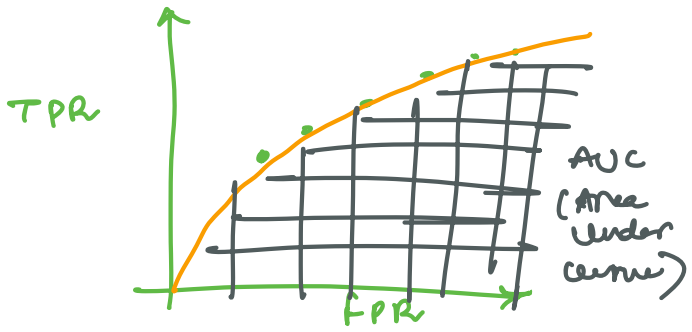
$$\text{Precision} = \frac{TP}{TP + FP}$$ } out of all predicted +ve, how many were actually correct

$$\text{Recall} = \frac{TP}{FN + TP}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

↓
HM of Precision & Recall

ROC Curve (Reciever Operating Characteristics)



Different values of threshold

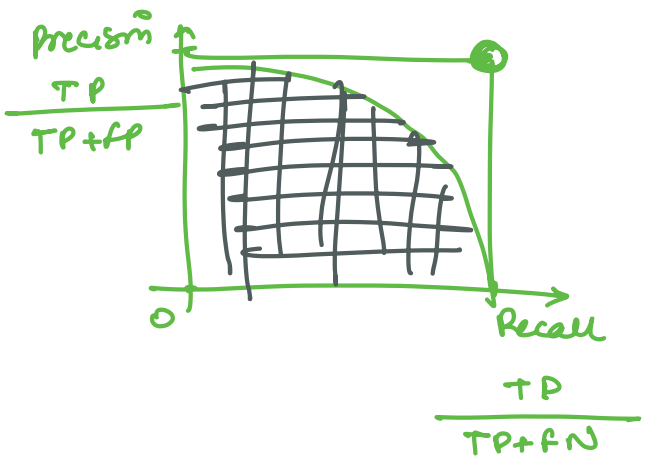
threshold → 0.5 : $<0.5 \rightarrow 0$, $\geq 0.5 \rightarrow 1$

0.8 : $<0.8 \rightarrow 0$, $\geq 0.8 \rightarrow 1$

$$\text{True Positive Rate} = \frac{TP}{TP+FN} * 100$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} * 100$$

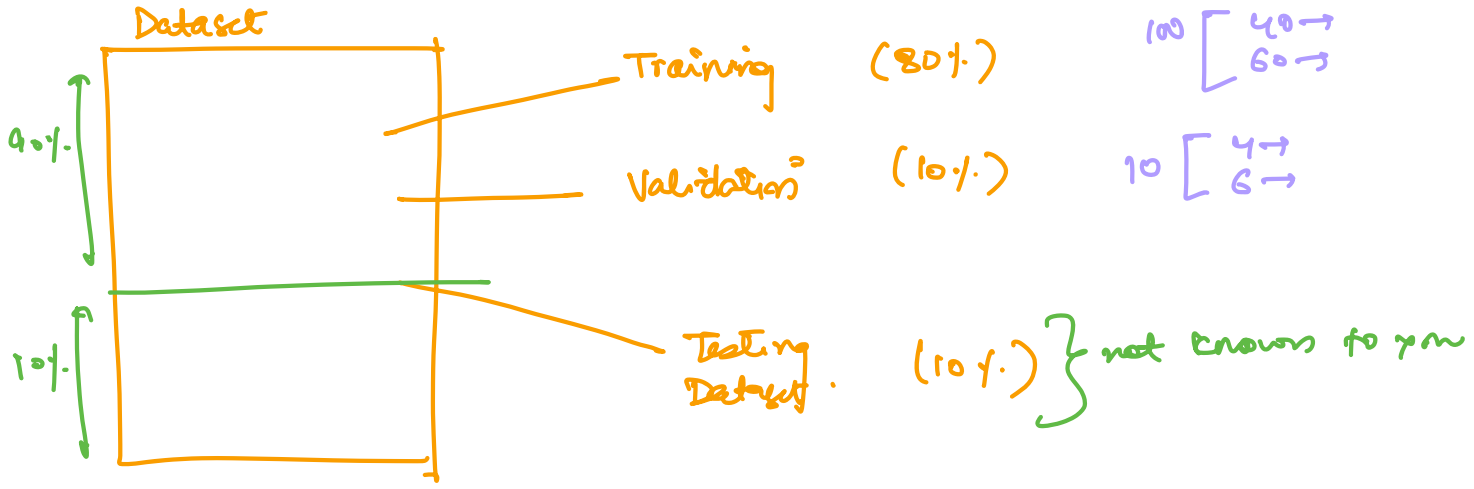
PR Curve



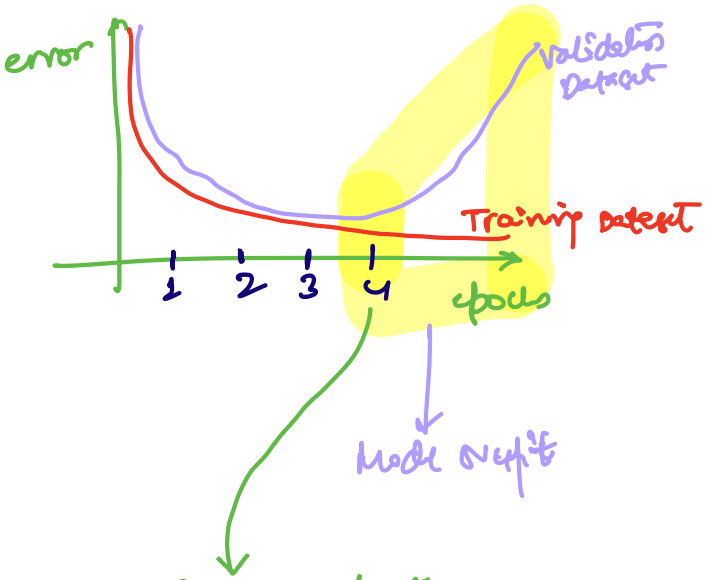
ideally:

high Precision $FP = 0$

high Recall $FN = 0$



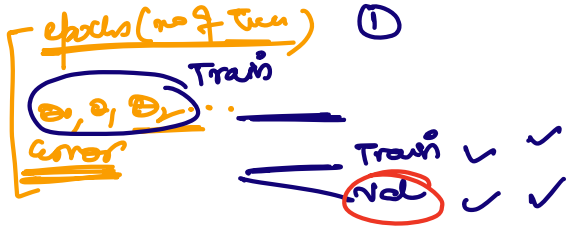
Logistic regression

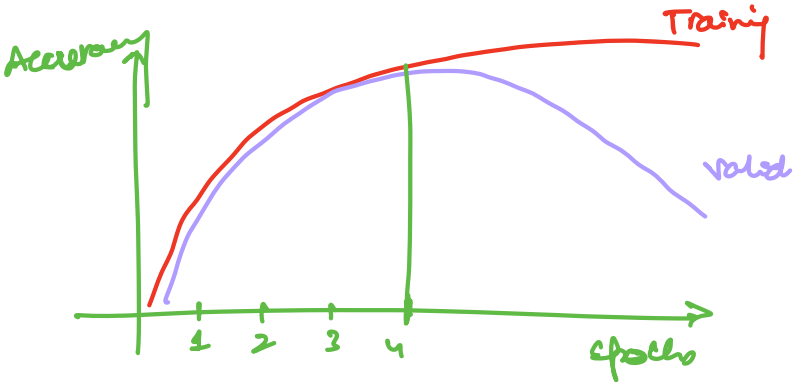


Train Data: Model Train

parameters learn:

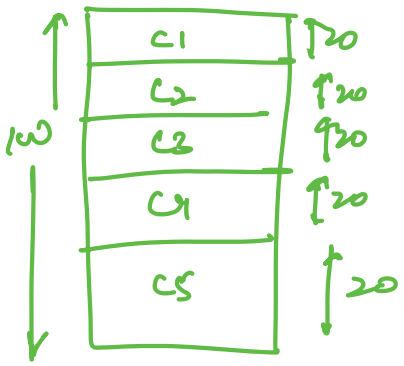
$\theta_0, \theta_1, \theta_2 \ldots$





K fold cross validation:

Train + Val



k = 5, 7 ...

step1: TD: C1 C2 C3 C4 VD: C5

2: TD: C1 C2 C3 C5 VD: C4

3: TD: VD: C3

4: TD: VD: C2

5: TD: VD: C1

Report Avg

KNN Naive Bayes DT Logistic Reg

Overfitting:

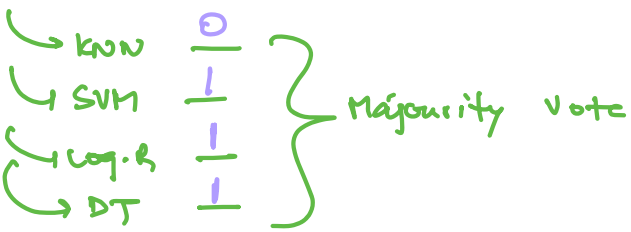
Model performs very well on training data but doesnot perform good on test data.

Solution:

→ k fold cross validation

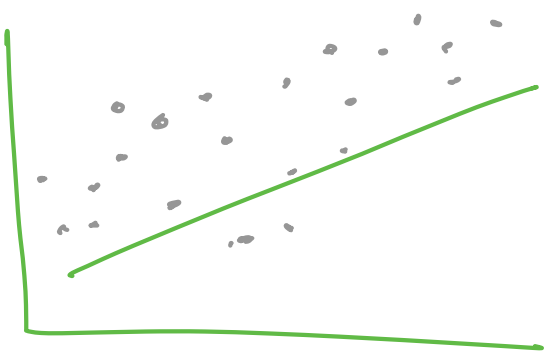
→ Sufficient data → Bootstrap

→ Ensembling techniques.

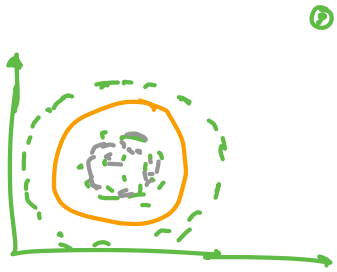


Underfitting:

Model is not going to learn patterns from training data.



Underfit model will give poor performance on train as well as test data.



Solutions:

→ No. of features increase

→ Models Complexity

→ Reduce noise

→ Increase the duration of training.

BIAS AND VARIANCE:

Bias: wrong assumptions about data — like assuming data is linear in reality is follows a complex fn.

↳ It is the inability of the model bcz of that there is diff in predicted value & actual value.

Low Bias: lower assumptions make a model which closely matches the training dataset. (Simple model)

High Bias: more assumptions model will not match training dataset closely. (Complex model)

Variance:

↳ measure of spread in data from its mean position.

↳ sensitive to a subset which follows same distribution as your training dataset.

Low variance: model is very less sensitive to changes.

High variance: model is very sensitive to changes and it can result in significant changes if trained on a different subset.

↓ overfitting.