Bias & Variance - Model Performance Tuning

Bias
Prejudice.

Bias represents how much the model is inclined to interpret the data in a certain way before seeing the data itself.

In simple terms, bias is the difference between the predicted and actual value
           (expected)

\# A model with a higher bias would not match the data set closely.

\# A low biac model will closely match the training data set

Variance

— is the ==variability of predictions== i.e., how spread out model predictions are.
↓

K-fold cv: K-fold cross validation

variance refers to changes in the model predictions when using different portions of the training data set.

© apc

Nokia 1600   **Samsung Ultra**

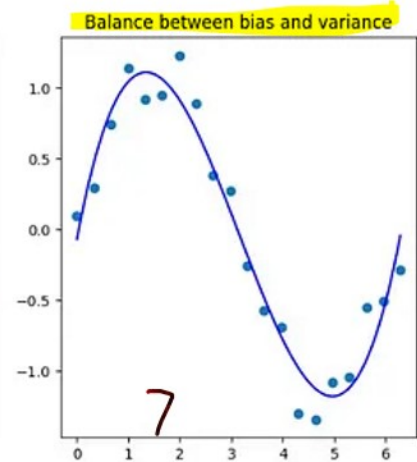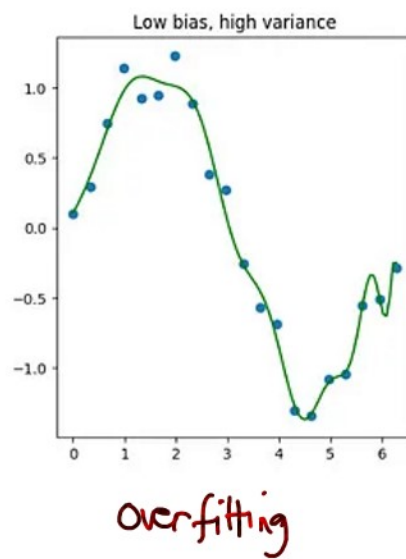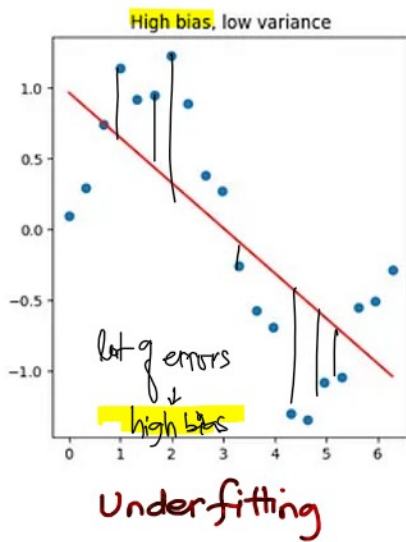simple   complicated

very less 10 features   [lots of features > 100]

Low Variance   High Variance

Underfitting

High Bias

Truth

Low Bias

Bill Howe, UW
src: domingo 2012

Overfitting

High bias, low variance

Low bias, high variance

Balance between bias and variance

lot of errors
↓
high bias

7

Underfitting   Overfitting
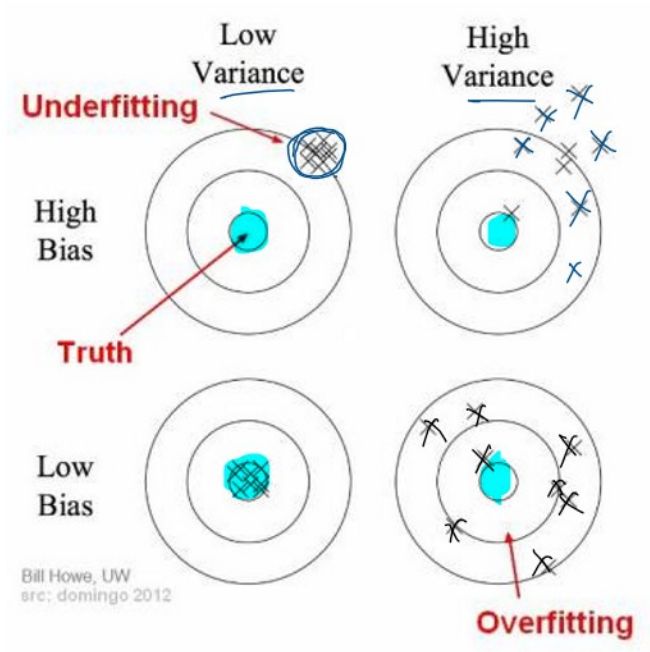
K-fold cross validation
Approach #①

1000 rows → random sampling and select 80% of data for training and **20%** of data for **validation**.
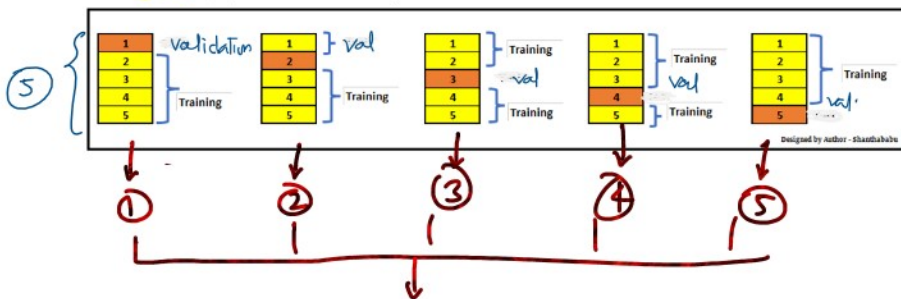
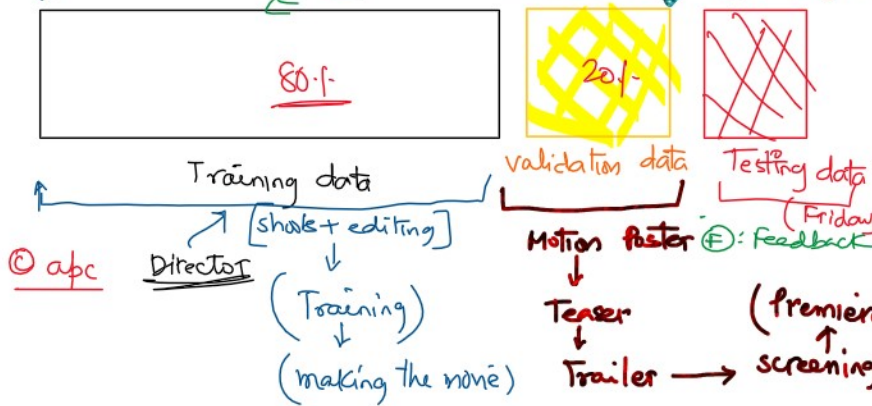K=5 fold cross-validation

Approach # 2
K(K=5) fold cross-validation

5 {
1 Validation | 1 } val | 1 } Training | 1 } Training | 1 } Training
2 | 2 | 2 | 2 | 2
3 | 3 val | 3 val | 3 | 3
4 Training | 4 Training | 4 | 4 val | 4 val
5 | 5 | 5 Training | 5 Training | 5

Designed by Author - Shanthababu

① ② ③ ④ ⑤

(avg. validation frequency) → Robust way of validation

(deploying the model to prod)

Model : Movie

data set

80%   20%   (testing data)

Training data
[shoot + editing]
© abc   Director
              ↓
           (Training)
              ↓
        (making the movie)

validation data
Motion Poster Ⓕ: feedback
    ↓
Teaser
    ↓
Trailer →

Testing data
(Friday release) on BO: Box Office

(Premiere)
   ↑
screening

Training # 3 years Jan'21 - Dec'23
Validation # Jan'24 - Jul'24.

M	SF.	Actual
Aug	200	300
Sep	150	180
Oct	180	200
Nov	—	110

P   A
(200 — 300)

$= \dfrac{(100)}{300} \times 100$

Error = 33.33%.

Chosen complexity level

← Underfitting   Overfitting →

Error/variance

Validation Error
validation

Both the errors are low

Variance

Training Error
→ Training

↓ desired outcome

Model "complexity"
(e.g., number of features)
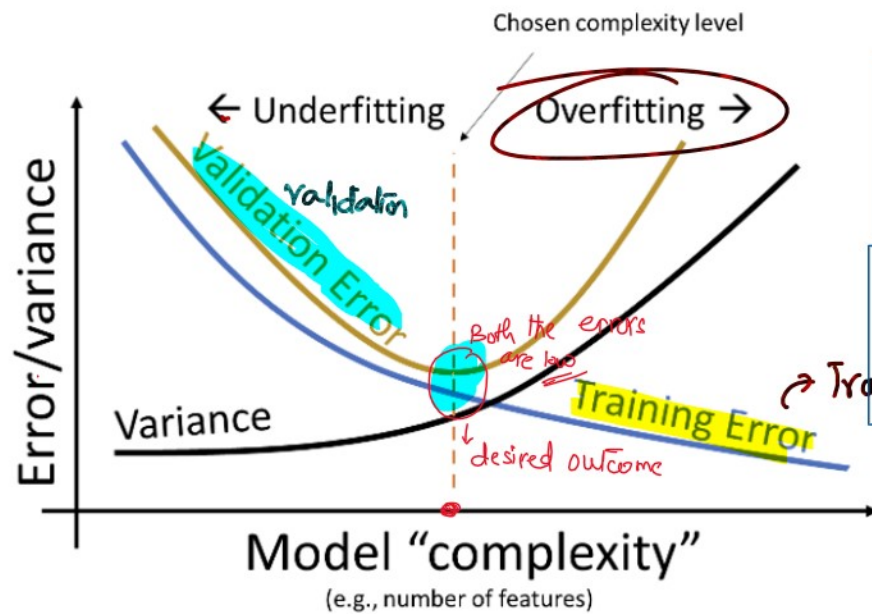
Training Error : Low
Validation Error: High] overfitting
(K-fold cv)

Training Error : High
Validation Error : High] Underfitting