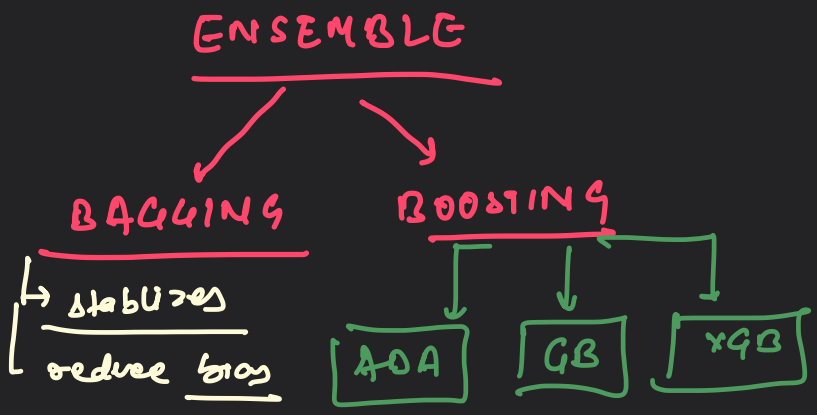
Bootstrap Aggregation

- Few

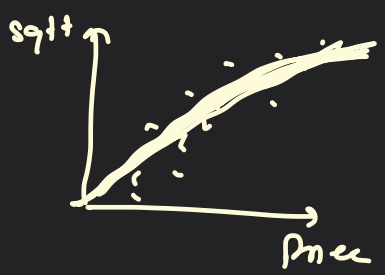
Let's wait till 7:07 AM

Bootstrap Aggregation (bagging)



Bias & Variance

LR (Pruned)



bias (underfit)

Variance (

objective

Low Variance : overly sensitive to the little changes in the data

Low bias : overly simple model & able to able to capture the real pattern properly

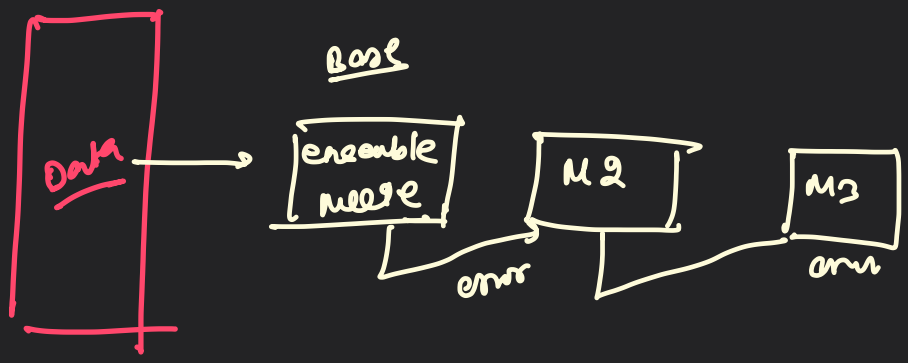
Simple Model (LR) → high bias → underfit
low variance

Complex model (NN) → low bias
high variance (overfitting)

Disadv

1. overfitted
2. Unstable (prev → new) unable to missing data
3. Difficult
4. Biased

Random Forest



① Bootstrap sample

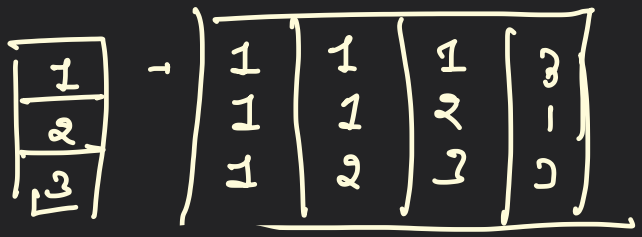
Randomly draw N sample out of given data

repeative sample

100T → RF

100 → bootstrap

→ sf



Dataset: 1000 records of custom (age, income, purchase)

① Create a bootstrap sample (100)

Total number of sample = 100

size of individual: Same as df (1000)

Random Samples

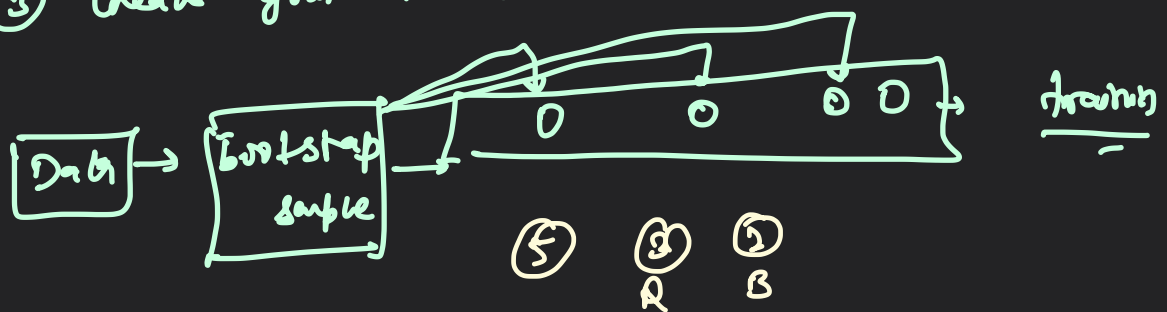
df → Random (repeated)

1000 → with replacement

100 bs have size 1000

100

③ Create your 100 dT



⑤ ② ①
R B

④ predict

Age	IB	pass/fail
18	x_1	P
21	x_2	F
24	x_3	P
17	x_4	F
26	x_5	P

→ bootstrap (10) sample

len(training)

10 random

RD10

RD1	
18	x_1
24	x_2
21	x_2
21	x_2
26	x_5

⑤ balls

①

Age	D/ND
x_1	D
x_2	ND
x_3	D
x_4	D
x_5	ND

OOB

OOB

OOB

Bagging

③

DT_1 DT_2 DT_3

DF_1

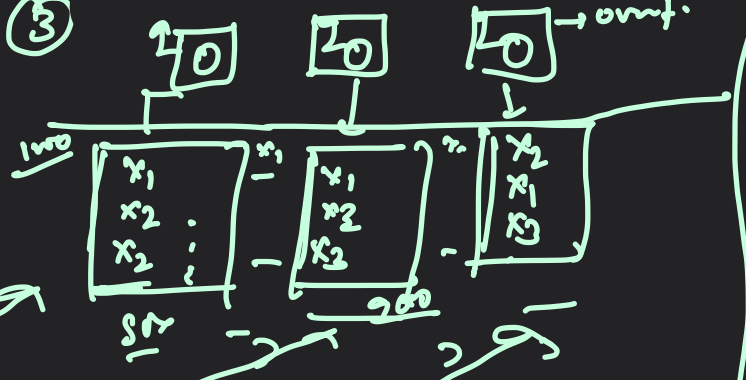
Age	D/ND
x_1	
x_1	
x_2	
x_3	
x_5	

df

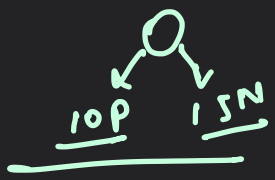
Suppose

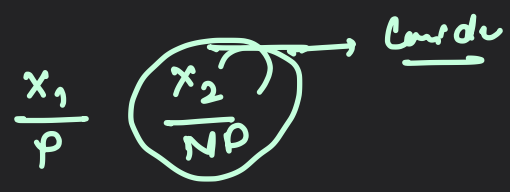
x_1	y_1
x_2	y_2
x_3	y_3

③



Final Predic (Voting)





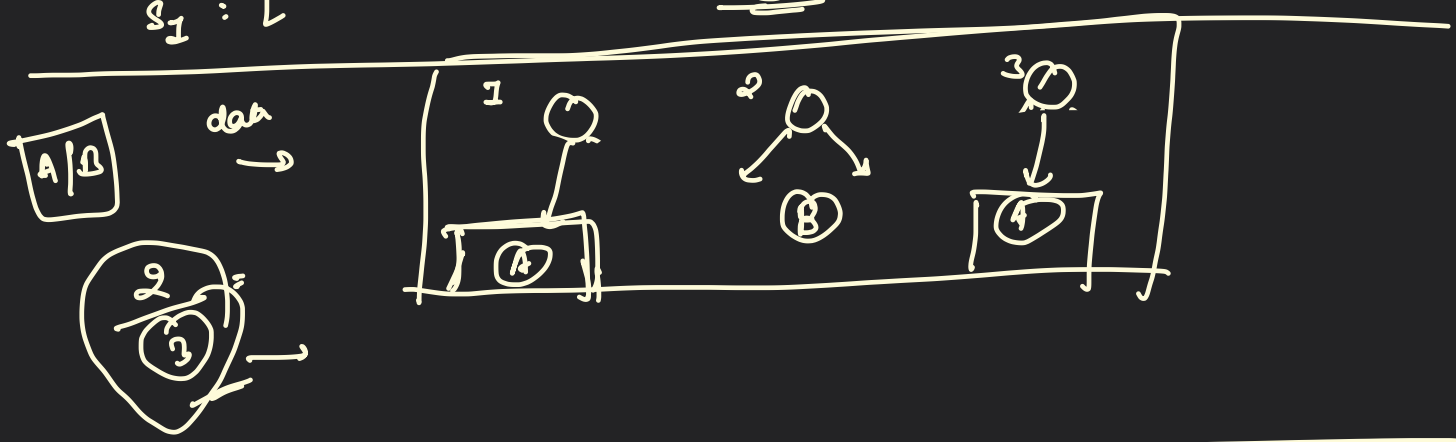
Dataset : [1, 2, 3, 4, 5] = random.5

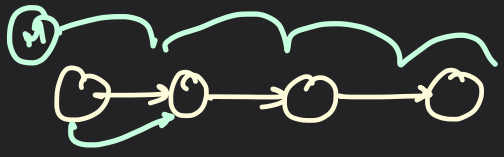
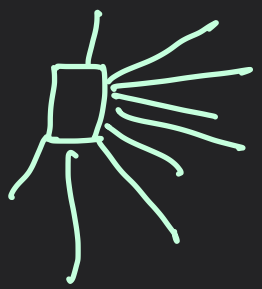
RFC (n_estimators)

① Determine the sample length len=5

② Bagging (n_estimators = 100) → 100

S_1 : [



Boosting	Bagging
 one-by-one model	 Parallel model
sequential ensemble	all-at-one model

Model

Model → Amount of data is very less

Model → Rel ✗ → Inaccurate

Underfitting

Model → trains data over time

Complex data ? Rel ✓

Training → Acc (Train)

Testing → (Test) ✗

Over fit

Train : 67% :

Test : 23% : 65% / 60%

Underfitting

training : 23% ✗

CF/

B/V

OF/UF