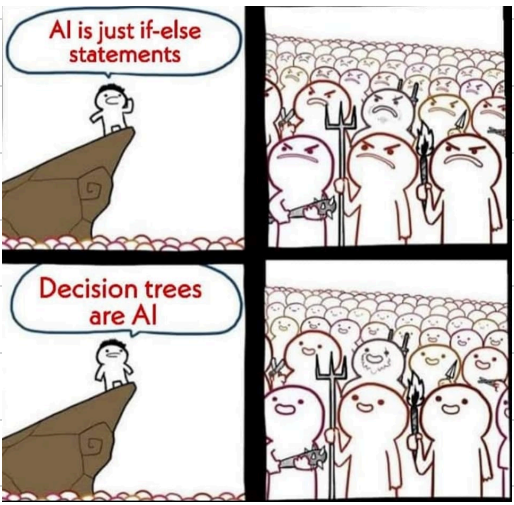
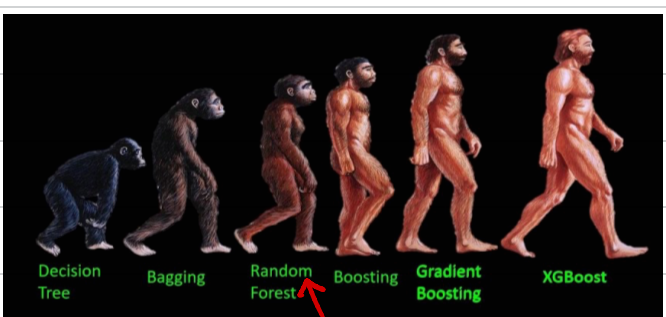
Session-5

Bagging and Random Forest (RF)



Agenda

1. Ensemble
2. Bagging
3. RF - Random Forest
4. OOB - Score
5. Hyper-Parameter Tuning - GRID Search & Random Search



Where we will reach today

DT- Regression

$$G.I: \quad 1 - \sum_{i=1}^{K} p(y_i)^2$$

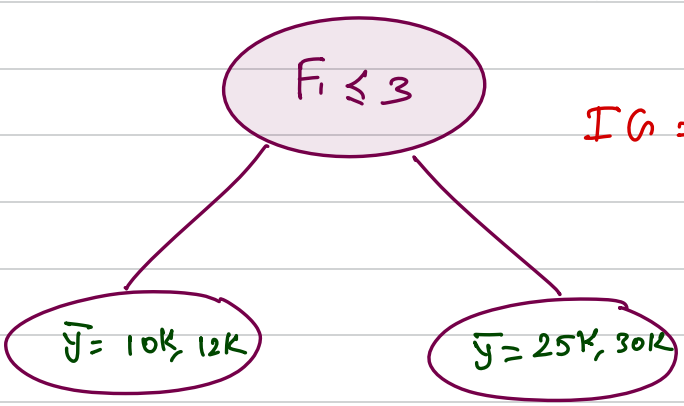
$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

F_1	y
3	10K
2	12K
20	25K
23	30K

$\bar{y} = 19.25$

$$MSE = \frac{1}{4}\left((10-19.25)^2 + (12-19.25)^2 + (25-19.25)^2 + (30-19.25)^2\right)$$

=1/4 * ((10-19.25)^2 + (12 - 19.25)^2 + (25-19.25)^2 + (30-19.25)^2) = 71.68



IG = 71.68 - 3.37 = 68.38

=1/2 * ((10 - 11)^2 + (12 - 11)^2) = 1

=1/2 * ((25 - 27.5)^2 + (30 - 27.5)^2) =6.25

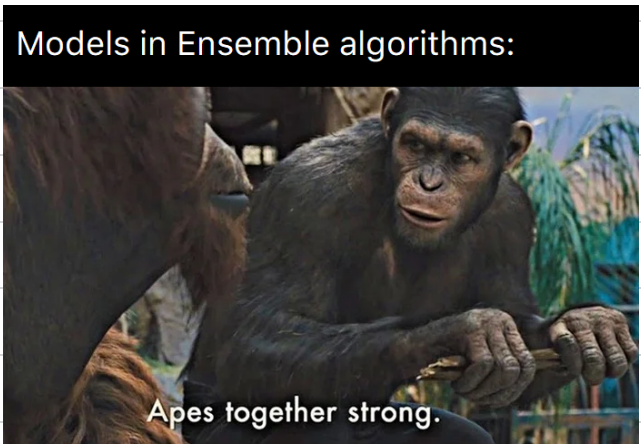
$\frac{2}{4} \times 1 + \frac{2}{4} \times 6.25 = 3.37$

Ensemble

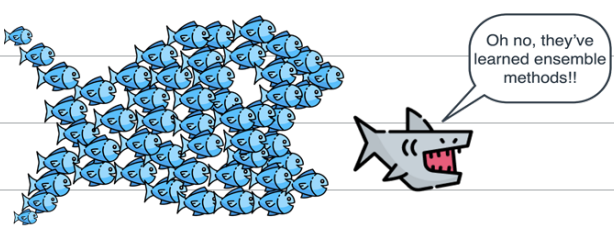
1. Log-Reg → 1
2. Perceptron → 0
3. Decision Trees → 1

Voting.

1



→ Types Of Ensemble Learning.



1. Bagging - R.F
2. Boosting (GBDT)

xgboost, lightgbm

3. Stacking
4. Cascading

} Never in comp's, but used in competitions

Which one of these is a type of ensemble learning technique?

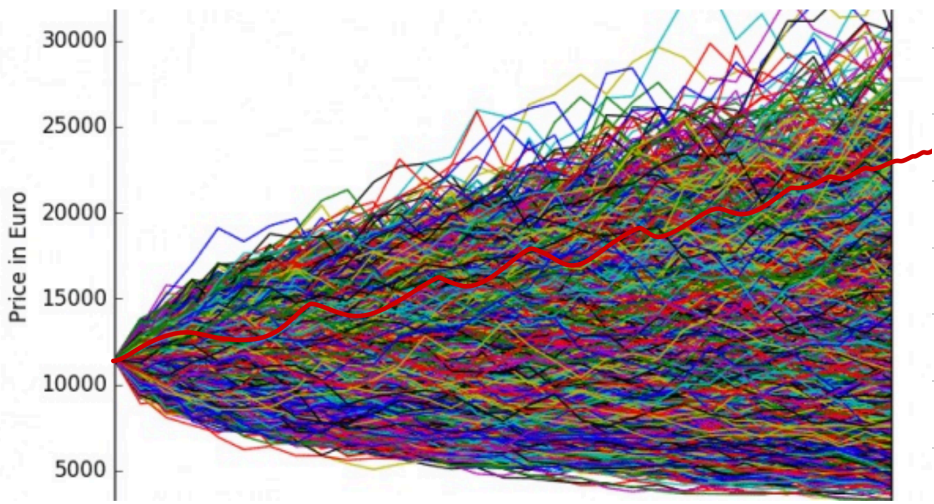
0 users have participated

A	Selecting	0%
B	Collecting	0%
✓ C	Bagging	0%
D	Randomizing	0%

End Quiz Now

based on all quizzes from the session

2	Soumyajit Misra	1/1 95.67
1	Karthik	1/1 96.38
3	Deepak	1/1 95.67
4	Shoreya gupta	1/1 95.13
5	Perisetla Pavan Kalyan	1/1 94.53
6	Gamidi Sri Valli Supraja	1/1 94.06
7	Vishwajit Verma	1/1 94.00
8	OM PRAKASH S	1/1 93.73
9	SHASHANK JHA	1/1 93.59
10	Sumanth Andhavarapu	1/1 92.87

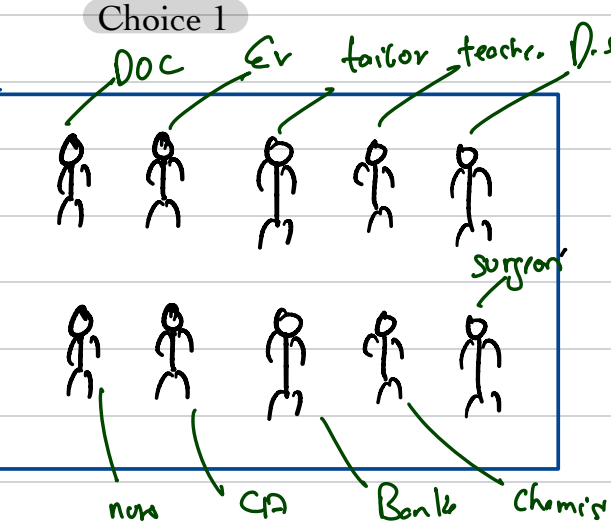


Population density

Bagging

Boot-strapped Aggregation.

Choice 1

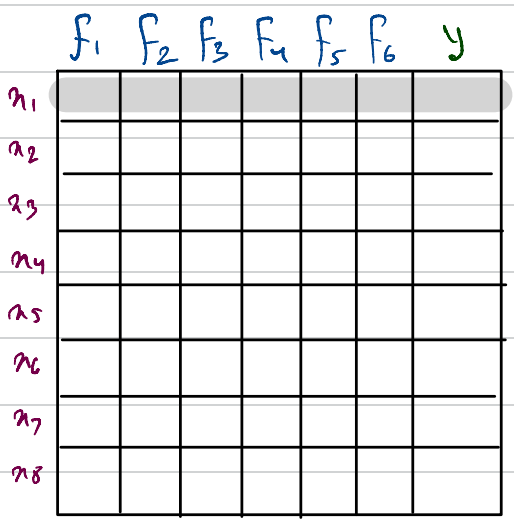


Choice 2

In bagging, each base-learner typically has high variance, also they're called "strong learners".

Dataset
n rows
m Featur

- DT1 — Sample 1k rows out of n rows, with replacement.
- DT2 — Sample 1k rows out of n rows, with replacement.
- DT3 — Sample 1k rows out of n rows, with replacement.
- DT4 — Sample 1k rows out of n rows, with replacement.
- DT5 — Sample 1k rows out of n rows, with replacement.
- DT6 — Sample 1k rows out of n rows, with replacement.
- DT7 — Sample 1k rows out of n rows, with replacement.



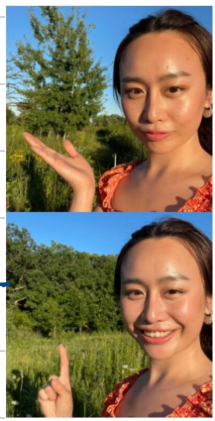
x4, x1, x3, x5, x7, x5

x1, x2

million row

10k rows

Random Forest



1 mill rows
1000 Featur

DT -> 10K rows
20 Featur

10K x 20

Dataset
n rows
m Featur

DT1 Sample 1k rows out of n rows, with replacement.
Sample x features, out of m features, without replacement.

DT2 Sample 1k rows out of n rows, with replacement.
Sample x features, out of m features, without replacement.

DT3 Sample 1k rows out of n rows, with replacement.
Sample x features, out of m features, without replacement.

DT4 Sample 1k rows out of n rows, with replacement.
Sample x features, out of m features, without replacement.

DT5 Sample 1k rows out of n rows, with replacement.
Sample x features, out of m features, without replacement.

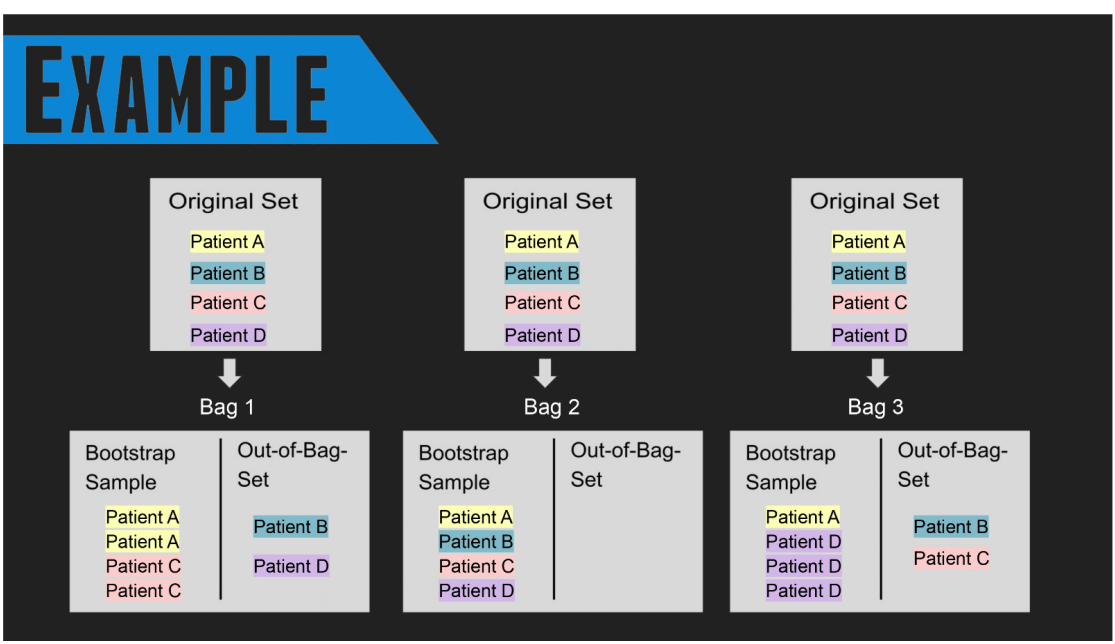
DT6 Sample 1k rows out of n rows, with replacement.
Sample x features, out of m features, without replacement.

DT7 Sample 1k rows out of n rows, with replacement.
Sample x features, out of m features, without replacement.

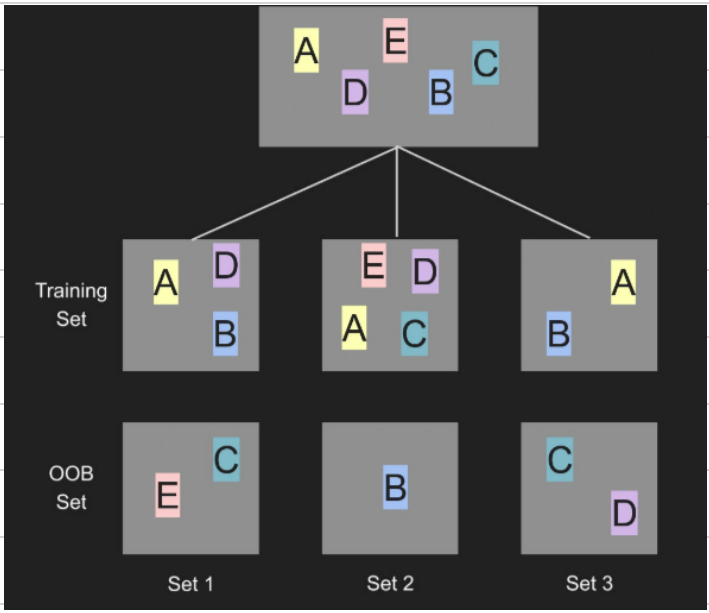
10K rows

8K — training
1K — testing
1K — validation

Rich man's evaluation



OOB - Score



	f_1	f_2	f_3	f_4	f_5	f_6	y
x_1							
x_2							
x_3							
x_4							
x_5							
x_6							
x_7							
x_8							

DT1 → x_1, x_2, x_3

$(x_4 - x_8)$ test

DT2 → x_1, x_6, x_7

test $(x_2 - x_5?), x_8$

Bias & Variance Tradeoff

What is the reason for introducing Row/Column Sampling in Random Forests?

0 users have participated

A	For decreasing the training time of model.	0%
✓ B	For tackling the problem of overfitting in base learners.	0%
C	For tackling the problem of underfitting in base learners.	0%
D	None of the above.	0%

End Quiz Now

	K (2)	VV (1)	SM (3)
	Karthik	Vishwajit Verma	Soumyajit Misra
	2/2 184.98	2/2 186.66	2/2 179.44
4	SHASHANK JHA	2/2	175.95
5	Kiran Hebasur	2/2	175.71
6	Sumanth Andhavarapu	2/2	172.64
7	Praveen	2/2	171.99
8	Sri Harsha Nanduri	2/2	171.26
9	Deependu Ghosh	2/2	168.35
10	Arpita Saha	2/2	166.23

If a dataset contains "n" rows, and "m" of these rows are sampled to train the base learners in Random Forest, what will be the cross-validation data for each of the models?

1 user has participated

A	Complete dataset with "n" rows	0%
B	A part of "m" sampled rows	100%
✓ C	Remaining "n-m" rows after sampling	0%
D	None of the above.	0%

End Quiz Now

Leaderboard

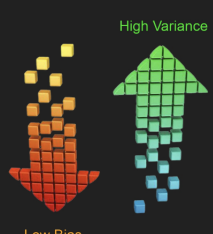
Based on all quizzes from the session

	VV (2)	K (1)	SM (3)
	Vishwajit Verma	Karthik	Soumyajit Misra
	3/3 277.29	3/3 279.91	3/3 257.89
4	SHASHANK JHA	3/3	256.88
5	Praveen	3/3	255.68
6	Sri Harsha Nanduri	3/3	254.33
7	Deependu Ghosh	3/3	252.87
8	Kiran Hebasur	3/3	250.86
9	OM PRAKASH S	3/3	250.39
10	Arpita Saha	3/3	244.45

Bias - Variance Tradeoff

The base learners in RF are Deep Decision Trees.

- So they slightly overfit on the sub sample of data
- Which means :

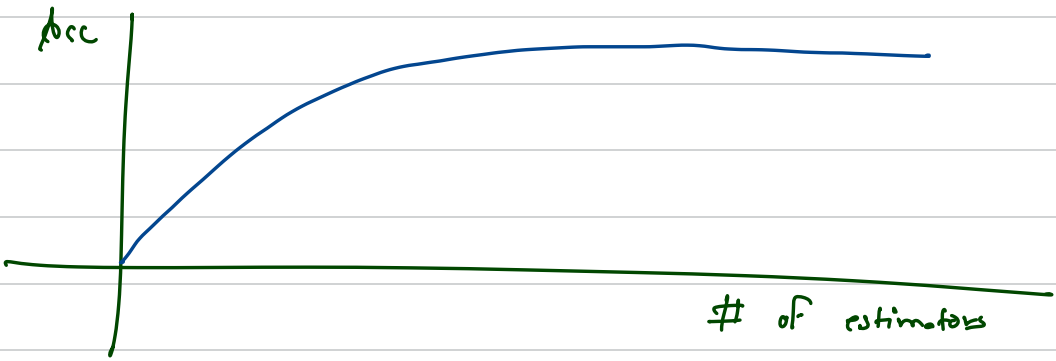


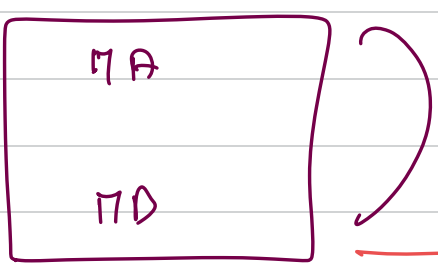
Impt. Hyperparameters- RF

of trees ↓ → High variance

1 DT
↓
overfit on sample data

# of trees ↑↑





100 — 80 – MA
20 – MB

1000 — 800 — MA
200 – MB

ccp-alpha

Ridge

Log-Reg. Loss function + $\lambda \times \|w\|^2$

D.T → DT + ccp-alph × # of Leaf nodes.

Loss function.

ccp-alpha tries to reduce the # of Leaf node

↑ ccp-alph → ↑ bias

↓ ccp-alpha → ↑ variance

What is the use of the hyperparameter 'ccp_alpha'?

0 users have participated

A	To set the column sampling ratio	0%
✓ B	To control underfitting or overfitting	0%
C	To optimize the learning rate of RF	0%
D	To set the depth of the base learners	0%

End Quiz Now

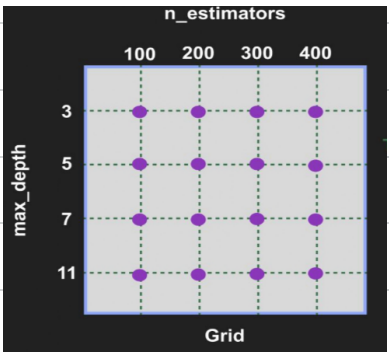
Leaderboard

Based on all quizzes from the session

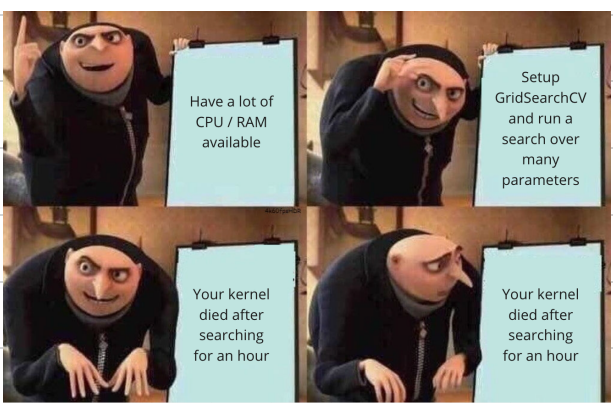
2	SHASHANK JHA	4/4 351.32
1	Soumyajit Misra	4/4 352.42
3	Deependu Ghosh	4/4 345.72
4	Sri Harsha Nanduri	4/4 336.33
5	Kiran Hebasur	4/4 330.37
6	Arpita Saha	4/4 328.72
7	Praveen	4/4 328.65
8	Narayana	4/4 317.29
9	Nayana	4/4 314.71
10	Karthik	3/4 279.91

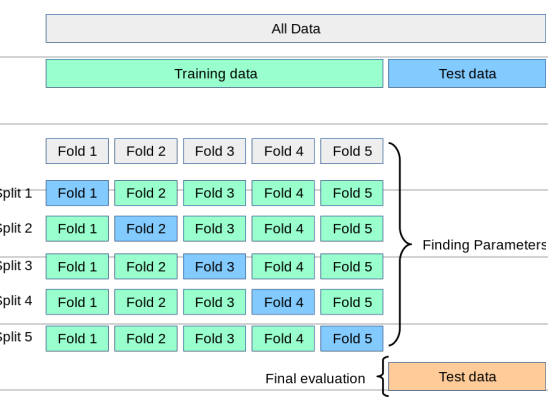
Fine-Tuning Hyper-parameters

```
# Define the hyperparameters grid for t
param_grid_dt = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [None, 5, 10, 15],
    'min_samples_split': [2, 10, 20],
    'min_samples_leaf': [1, 5, 10],
}
```



For crit in criterion:
    for depth in max_depth:
        for samp1 in min_samples_split:
            for samp2 in min_samples_leaf:
                train_model (crit, depth, samp1, samp2)
                store-acc





If you're performing GridSearchCV on a random forest model for the parameters,

- **'max_depth' having 3 values and**
- **'min_samples' having 4 values**
- **for 10 cross-validations**

How many times model.fit() is called?

0 users have participated

A	1440	0%
✓ B	120	0%
C	40	0%
D	30	0%

End Quiz Now

Based on all quizzes from the session

2	SHASHANK JHA	5/5 ⚡ 432.52
1	Soumyajit Misra	5/5 ⚡ 441.97
3	Kiran Hebasur	5/5 ⚡ 427.93

4	Deependu Ghosh	5/5	⚡ 425.00
5	Sri Harsha Nanduri	5/5	⚡ 416.91
6	Praveen	5/5	⚡ 407.32
7	Arpita Saha	5/5	⚡ 403.38
8	Narayana	5/5	⚡ 396.84
9	Vishwajit Verma	4/5	⚡ 362.22
10	Shoreya gupta	4/5	⚡ 355.70