Random Forest – 호텔 예약취소 예측

1 프로젝트 개요

Random Forest – 호텔 예약취소 예측

8 Classification Report & AUC Score

Classification Report

	precision	recall	f1-score	support
0	0.86	0.92	0.89	29919
1	0.84	0.75	0.79	17572
accuracy			0.86	47491
macro avg	0.85	0.83	0.84	47491
weighted avg	0.85	0.86	0.85	47491

Precision (정밀도)

1이라고 예측한 것 중, 얼마 만큼을 제대로 맞추었는가?

Precision (정밀도)

1이라고 예측한 것 중, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

Precision (정밀도)

1이라고 예측한 것 중, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

$$\frac{TP}{FP + TP}$$

Precision (정밀도)

1이라고 예측한 것 중, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

$$= \frac{TP}{FP + TP} = \frac{13149}{2432 + 13149}$$

Precision (정밀도)

1이라고 예측한 것 중, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 (TN)	2432 (FP)
1	4423 (FN)	13149 (TP)

$$= \frac{TP}{FP + TP}$$

$$= \frac{13149}{2432 + 13149}$$

$$= 0.8439$$

Precision (정밀도)

1이라고 예측한 것 중, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

$$= \frac{TP}{FP + TP} = \frac{13149}{2432 + 13149} = 0.8439$$

TYPE 1 에러와 관련

Recall (재현율)

실제 1인것 중에, 얼마 만큼을 제대로 맞추었는가?

Recall (재현율)

실제 1인것 중에, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

Recall (재현율)

실제 1인것 중에, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

$$\frac{TP}{FN + TP}$$

Recall (재현율)

실제 1인것 중에, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

$$\frac{TP}{FN + TP} = \frac{13149}{4423 + 13149}$$

Recall (재현율)

실제 1인것 중에, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

$$\frac{TP}{FN + TP} = \frac{13149}{4423 + 13149} = 0.7483$$

Recall (재현율)

실제 1인것 중에, 얼마 만큼을 제대로 맞추었는가?

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

$$\frac{TP}{FN + TP} = \frac{13149}{4423 + 13149} = 0.7483$$

TYPE 2 에러와 관련

F-1 Score

Precision과 Recall의 조화평균

F-1 Score

Precision과 Recall의 조화평균

$$2 \times \frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

F-1 Score

Precision과 Recall의 조화평균

$$2 \times \frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

Precision	Recall	산술평균	조화평균
0.4	0.6		
0.3	0.7		
0.5	0.5		

F-1 Score

Precision과 Recall의 조화평균

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision	Recall	산술평균	조화평균
0.4	0.6	0.5	
0.3	0.7	0.5	
0.5	0.5	0.5	

F-1 Score

Precision과 Recall의 조화평균

$$2 \times \frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

Precision	Recall	산술평균	조화평균
0.4	0.6	0.5	0.48
0.3	0.7	0.5	0.42
0.5	0.5	0.5	0.5

ROC Curve

실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

FPR (False Positive Rate)

TPR (True Positive Rate)

ROC Curve

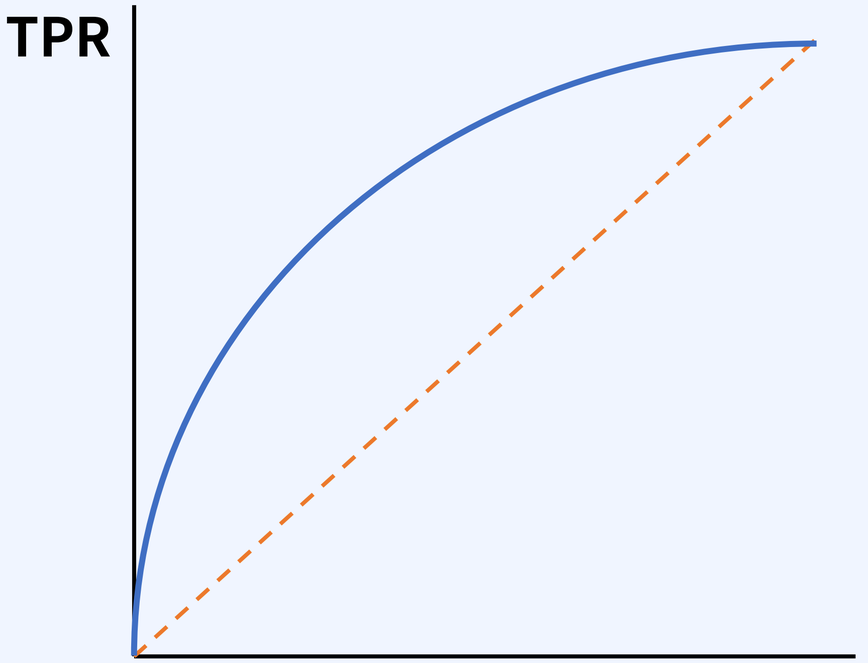
실제값 \ 예측값	0	1
0	27487 TN	2432 FP
1	4423 FN	13149 TP

$$\text{FPR (False Positive Rate)} = \frac{FP}{TN + FP}$$

$$\text{TPR (True Positive Rate)} = \frac{TP}{FN + TP}$$

ROC Curve

$$\frac{TP}{FN + TP}$$



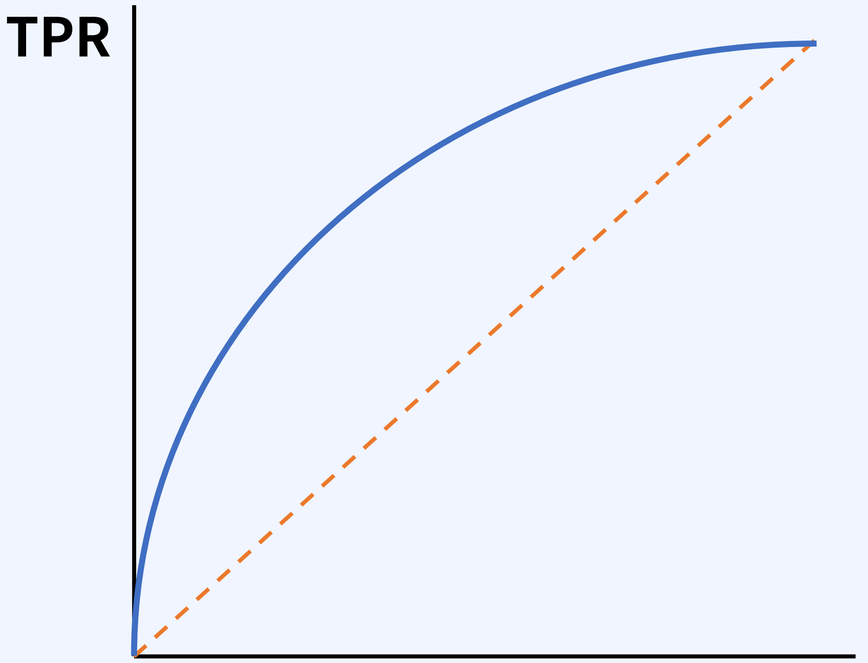
$$\frac{FP}{TN + FP}$$

실제값 \ 예측값	0	1
0	27487 (TN)	2432 (FP)
1	4423 (FN)	13149 (TP)

Threshold = 0.5

ROC Curve

$$\frac{TP}{FN + TP}$$



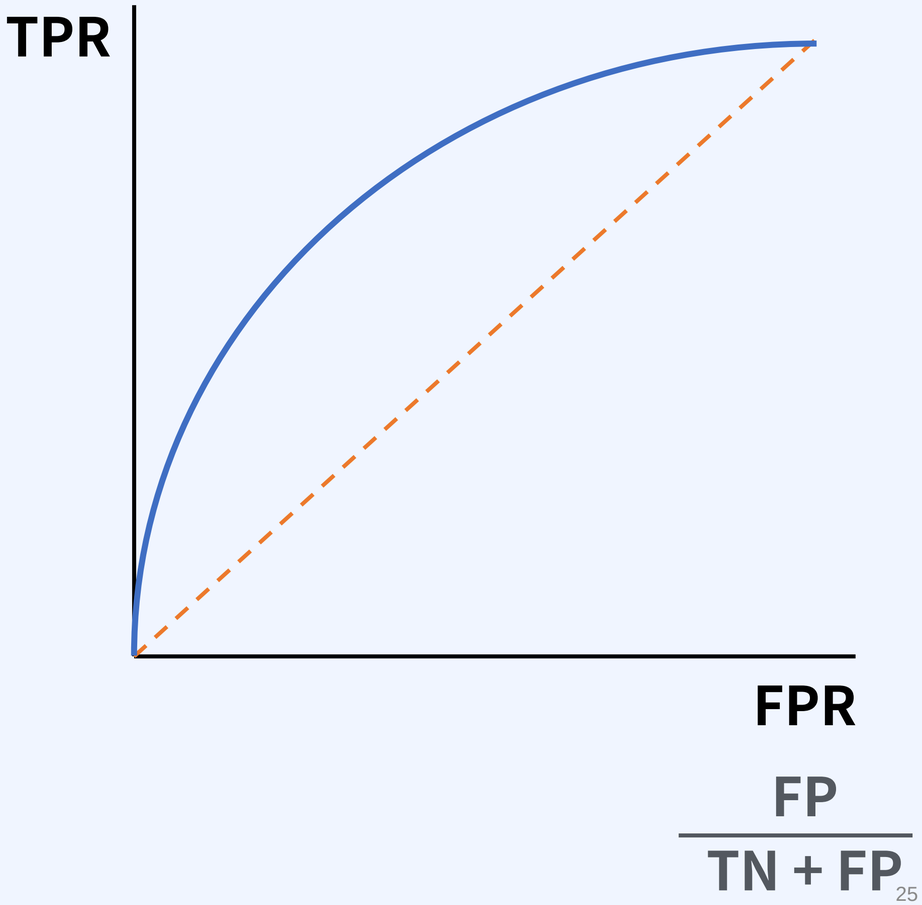
$$\frac{FP}{TN + FP}$$

실제값 \ 예측값	0	1
0	0 (TN)	29919 (FP)
1	0 (FN)	17572 (TP)

Threshold = 0

ROC Curve

$$\frac{TP}{FN + TP}$$



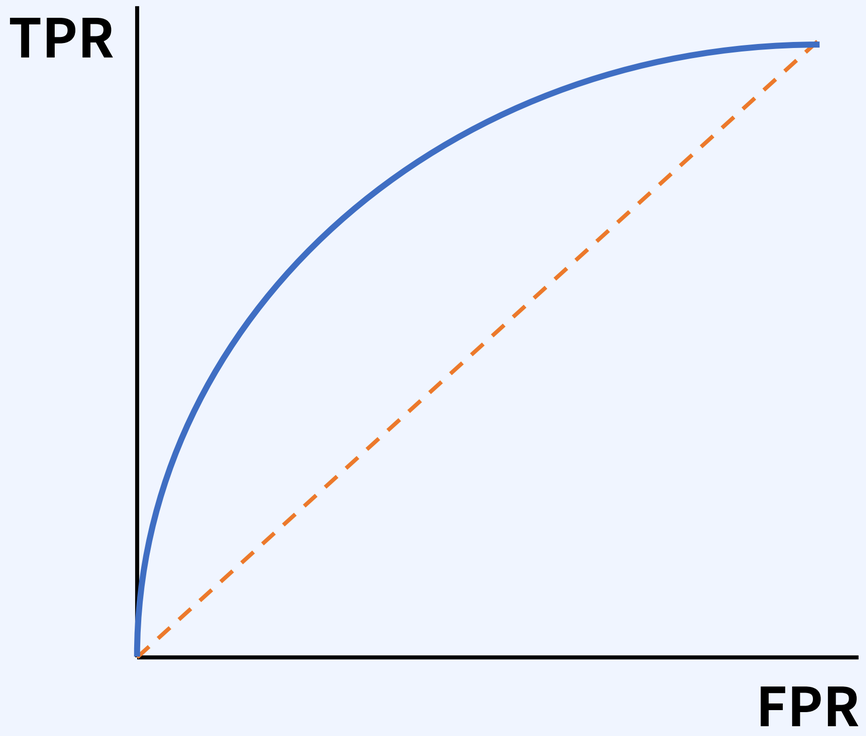
$$\frac{FP}{TN + FP}$$

예측값

실제값	0	1
0	29919 TN	0 FP
1	17572 FN	0 TP

Threshold = 1

AUC



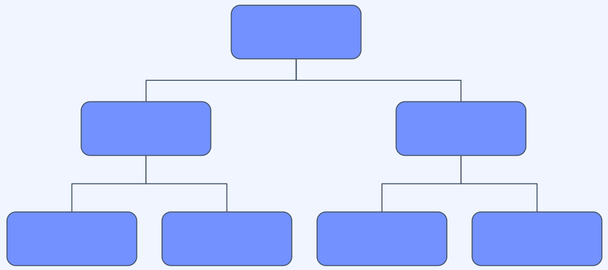
AUC = Area Under the Curve
(0.5 ~ 1.0)

Random Forest – 호텔 예약취소 예측

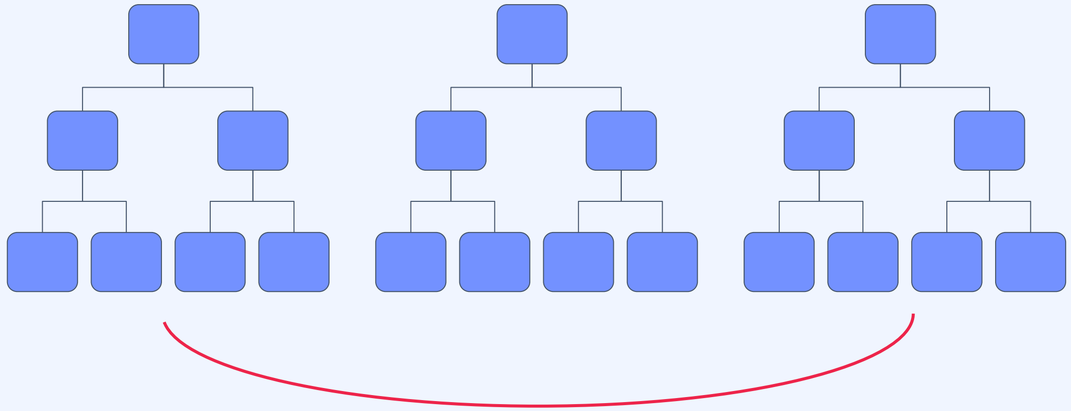
12 Random Forest 알고리즘의 이해

Ensemble 기법

Decision Tree



Random Forest



Random Forest의 특징

1)복원추출을 통한 여러 개의 Subset 사용 → Bagging

2)일부의 피처(독립변수)만 사용

3)각 트리는 서로 독립적이다

Bagging

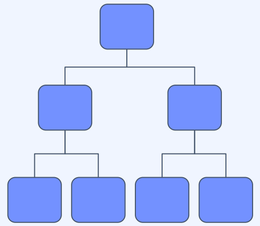
	X1	X2
1	A	13
2	B	2
3	C	5
4	D	7
5	E	10

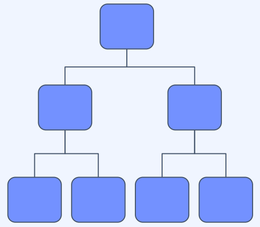
Bootstrap

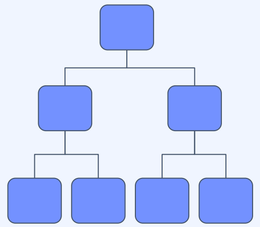
	X1	X2
1	A	13
4	D	7
5	E	10

	X1	X2
2	B	2
2	B	2
4	D	7

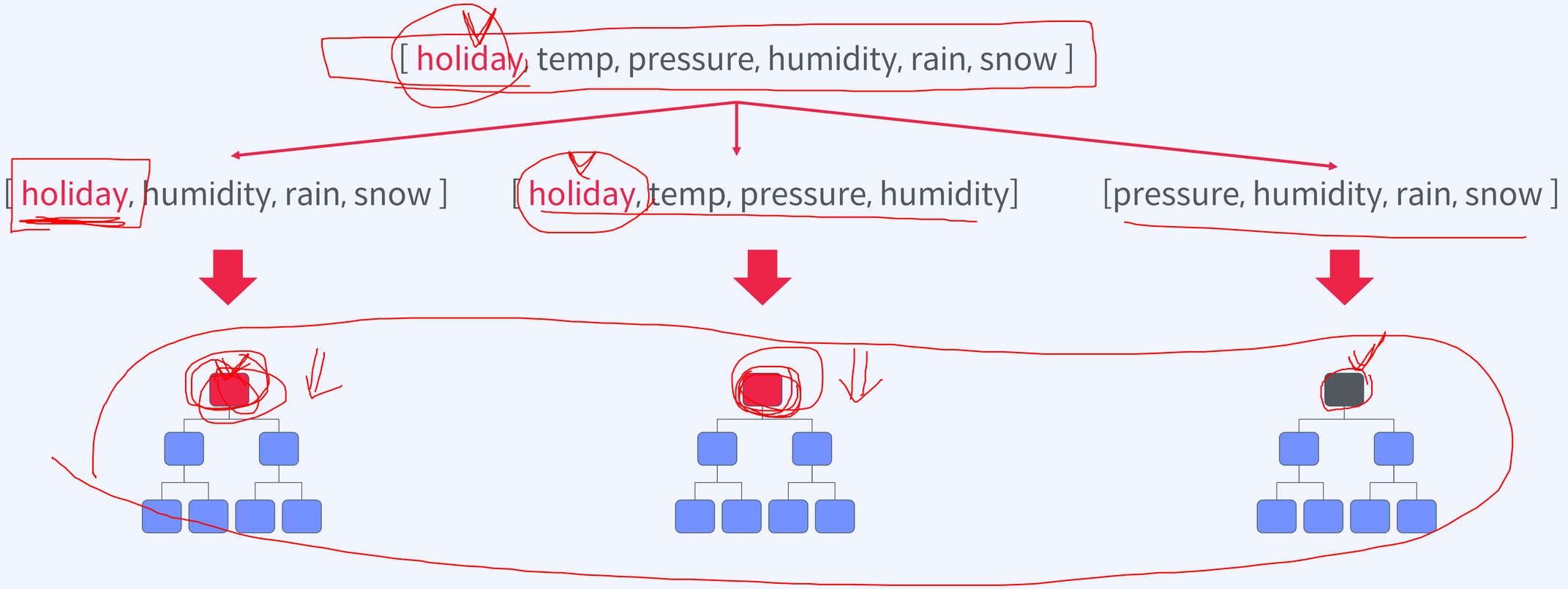
	X1	X2
1	A	13
3	C	5
3	C	5



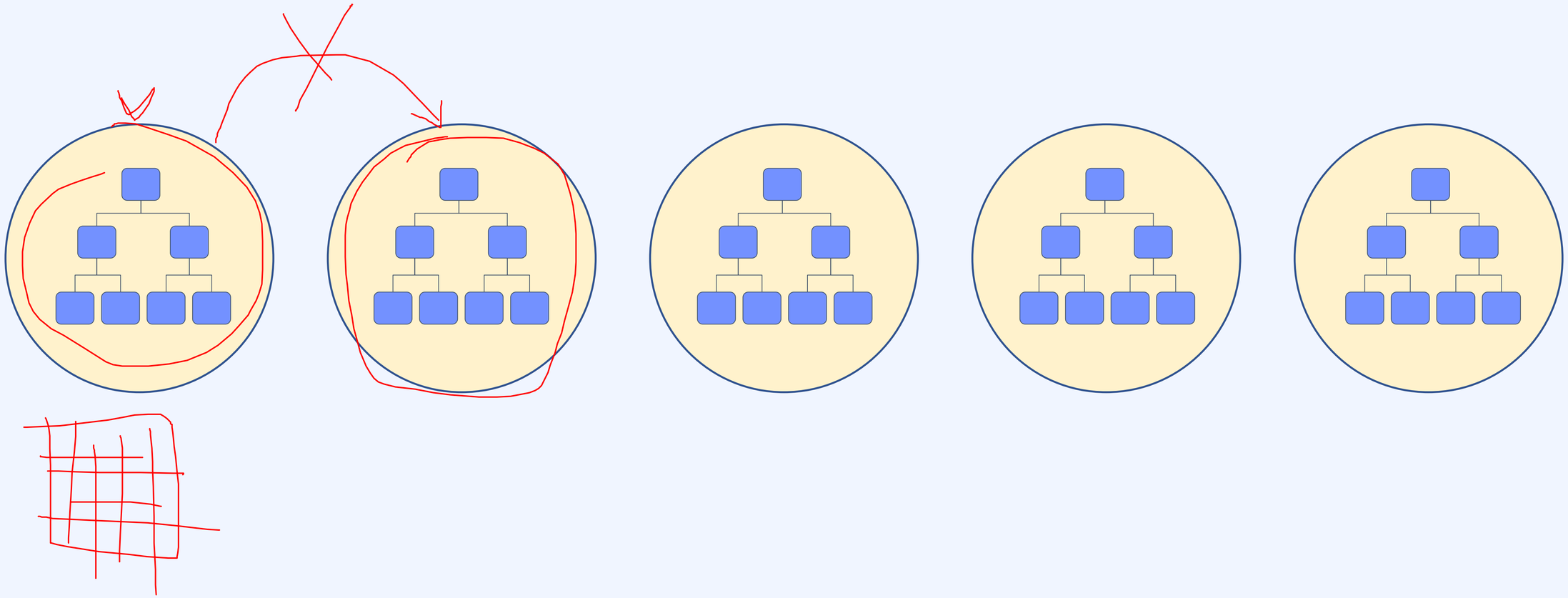




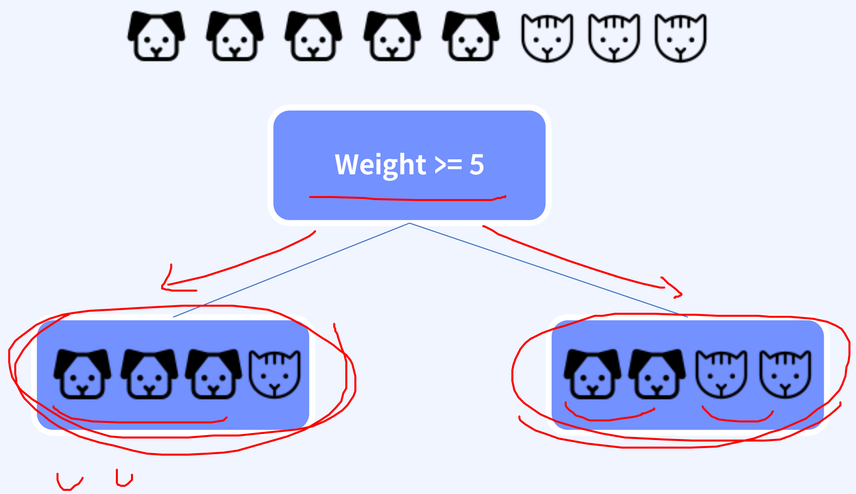
일부 피처를 사용 → 강력한 피처의 영향력 제한



각 트리가 독립적이라는 것의 의미

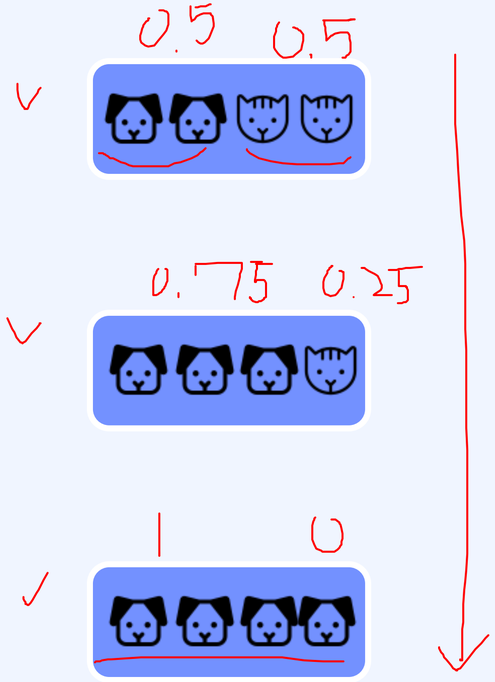


Tree - Classifier



GINI Index

$$1 - \sum_{i=1}^{n} (P_i)^2$$



$0.5^2 = 0.25 + 0.25$ $1-0.5 = 0.5$

$(0.75)^2 + (0.25)^2 = 1 - 0.625 = 0.375$

$1 + 0 = 1 - 1 = 0$

Cross Entropy

$$-\sum_{i=1}^{n} p_i * Log_2(p_i)$$

$0.5 \times Log_2(0.5)$

$-0.5 + -0.5 = -1 = +1$

0.75 0.25

$0.75 \times L_2(0.75) = -0.31$

$0.25 \times L_2(0.25) = -0.5$

$-0.415\ldots + -0.81\ldots = 0.8$

1 0

$1 \times L_2(1) = 0$

$0 \times L_2(0) = 0$

$= 0$