2016 11.10

Evaluating Model Performance

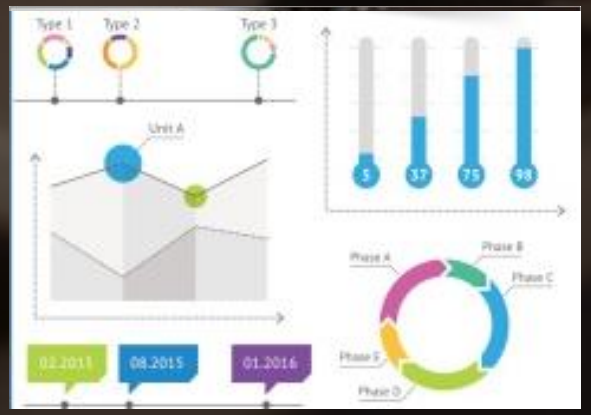
Chapter 10 presentation

[C조] 구민수 구유림 윤소라 손범호

Index

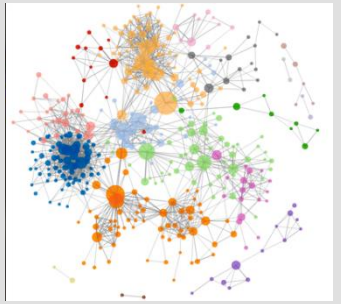
CHAPTER 2
Cross Validation

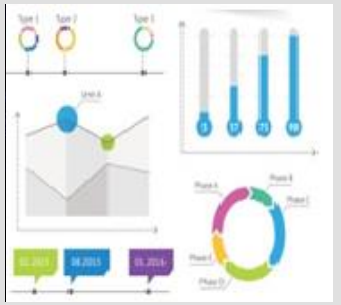
CHAPTER 1
Evaluating Model Performance



CHAPTER 01

Evaluating Model Performance

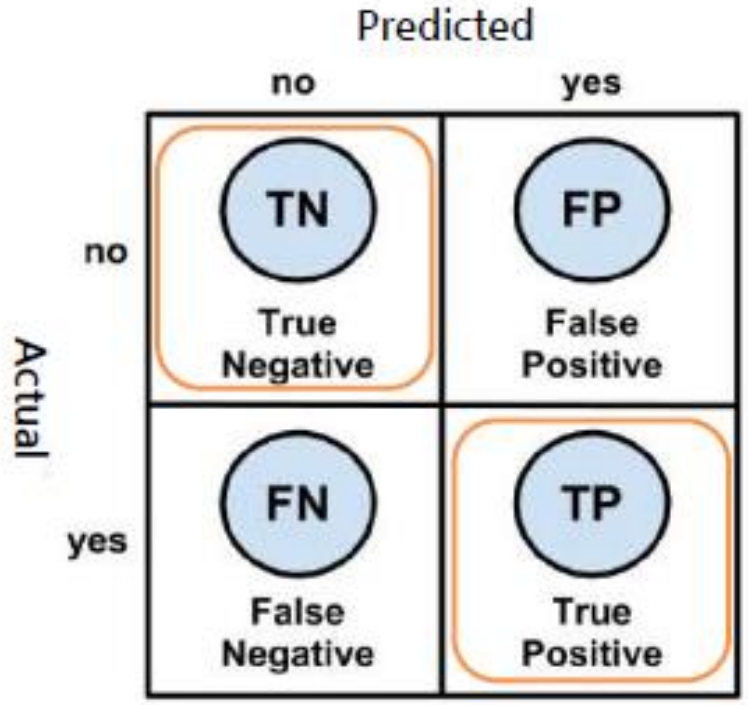




NO.01 Evaluating Model Performance

Confusion Matrix

= contingency table, an error matrix
알고리즘의 성능을 평가할 때 평가하는 지표 중 하나

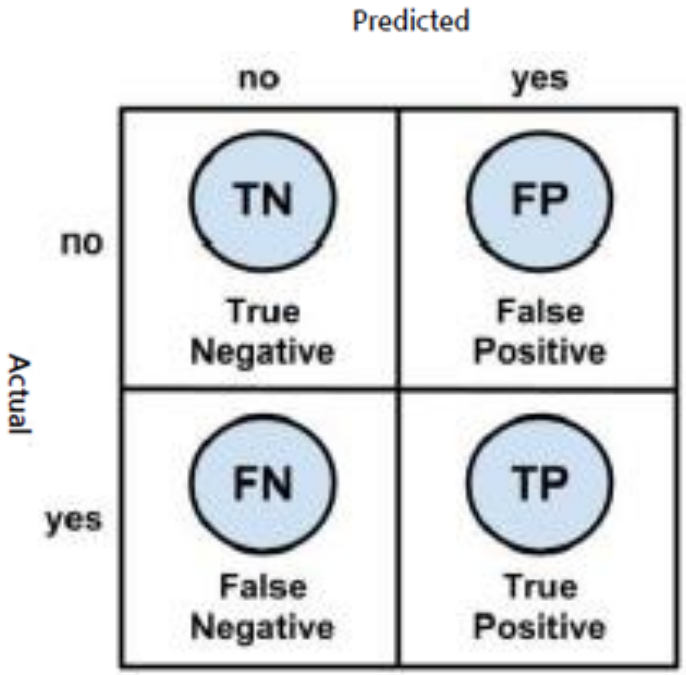


- True/False는 값을 맞췄는지를 나타냄
- Ex) (1인데 1), (0인데 0)
- Pos/Neg는 예측한 값이 1 or 0을 나타냄

- TP의 경우 1이라 예측했는데 실제로 1
- TN 0-0
- FP 1-0
- FN 0-1

NO.01 Evaluating Model Performance

Confusion Matrix



Accuracy: Actual과 Predicted가 같은 것

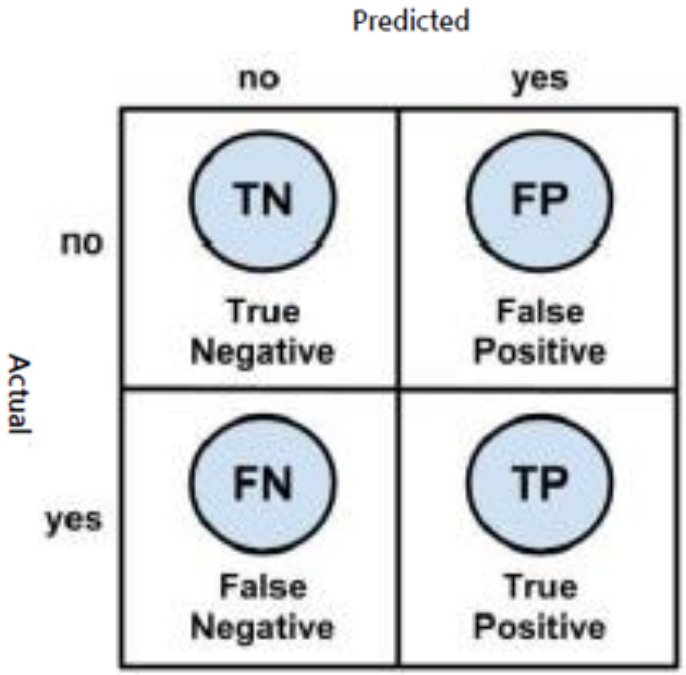
$$\frac{TP+TN}{TN+FN+FP+TP}$$

Error rate : Actual과 Predicted가 다른 것

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

NO. 01 Evaluating Model Performance

Confusion Matrix



Sensitivity or Recall: 실제 긍정일 때 맞춘 비율

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

Specificity : 실제 부정일 때 맞춘 비율

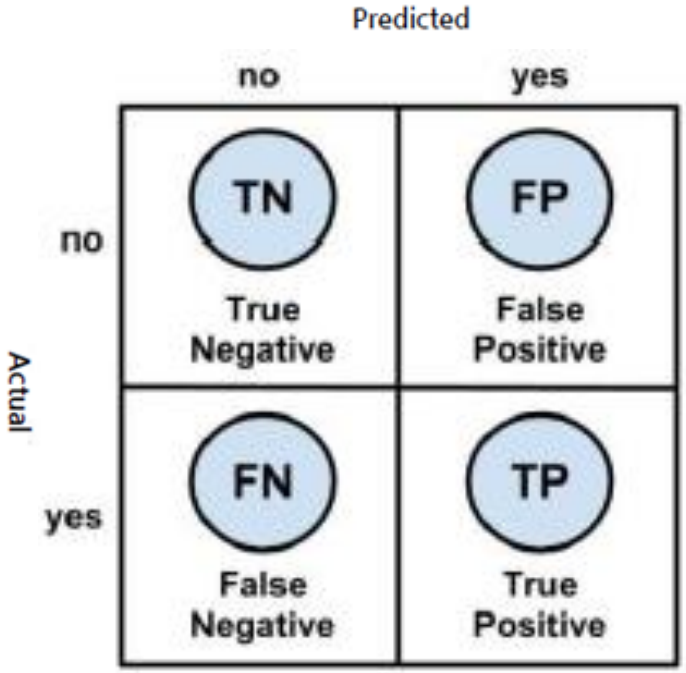
$$\frac{TN}{TN+FP}$$

Precision: 긍정 예측시 실제 긍정

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

NO.01 Evaluating Model Performance

Confusion Matrix



F-measure

같은 단위로 정밀도와 재현율을 구했을 때 정밀도와 재현율을 합쳐 모델의 성능 측정

$$\frac{2*precision*recall}{Recall+precision}$$

$$\frac{2*TP}{2*TP+FP+FN}$$

NO.01 Evaluating Model Performance

Contingency table– 응용

Example – 마케팅 캠페인 고객 응답 여부

성별(Gender)	거주지역(Location)	응답여부(Respond)
M	A	Y
M	B	Y
M	A	Y
M	C	Y
F	B	N
F	A	N
F	B	N
M	C	N
M	A	N
M	A	Y

NO.01 Evaluating Model Performance

Contingency table– 응용(Gini 척도= 클래스 비율의 제곱의 합)

모집단에서 두 개의 집단을 추출하고,
그 집단에서 동일 클래스에 있는 개체가 반복해서 나올 확률

- 1. Gender

$$G = \frac{3}{10}\{(\frac{0}{3})^2 + (\frac{3}{3})^2\} + \frac{7}{10}\{(\frac{5}{7})^2 + (\frac{2}{7})^2\} = 0.714286$$

- 2. Location

$$G = \frac{5}{10}\{(\frac{3}{5})^2 + (\frac{2}{5})^2\} + \frac{3}{10}\{(\frac{1}{3})^2 + (\frac{2}{3})^2\} + \frac{2}{10}\{(\frac{1}{2})^2 + (\frac{1}{2})^2\} = 0.526667$$

Gini 값이 큰 Gender를 분기 기준으로 선택!

		Respond		Total
		Y	N	
Gender	F	0	3	3
	M	5	2	7
Total		5	5	10

		Respond		Total
		Y	N	
Location	A	3	2	5
	B	1	2	3
	C	1	1	2
Total		5	5	10

NO.01 Evaluating Model Performance

Contingency table–응용(Chi-Square)

통계학적 유의성에 대한 검정, 관측된 표본들 간의 차이가 우연에 의한 확률
그빈도에 대한 기대값과 관측값의 표준화된 차이의 제곱들의 합

- 1. Gender

$$\chi^2 = \frac{(0-1.5)^2}{1.5} + \frac{(3-1.5)^2}{1.5} + \frac{(5-3.5)^2}{3.5} + \frac{(2-3.5)^2}{3.5} = 4.29$$

		Respond		Total
		Y	N	
Gender	F	0 (1.5)	3 (1.5)	3
	M	5 (3.5)	2 (3.5)	7
Total		5	5	10

- $\chi^2 = \Sigma_{i,j}(f_{ij}-e_{ij})^2/e_{ij}$
- f_{ij} : (i,j)셀의 관측 빈도
- e_{ij}: (i,j)셀의 예측 빈도
  - $e_{ij} = (f_{i.} * f_{.j})/ f_{..}$
- 자유도 (r-1)(c-1)의 카이제곱 분포를 따름

NO. 01

Evaluating Model Performance

Contingency table–Entropy 감소(정보 이익)

=1 * (P(A)log2P(B) + P(A)log2P(B))

- 1. Gender

		Respond		Total
		Y	N	
Gender	F	0	3	3
	M	5	2	7
Total		5	5	10

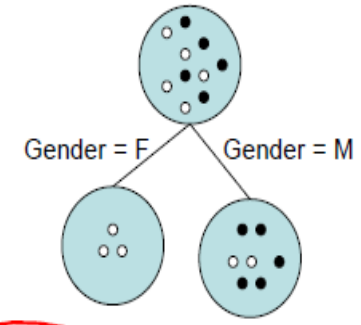
$$E_{before} = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1$$

$$E_{left} = -0 - \frac{3}{3}\log_2\frac{3}{3} = 0$$

$$E_{right} = -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7} = 0.863121$$

$$E_{after} = \frac{3}{10}\times 0 + \frac{7}{10}\times 0.863121 = 0.604185$$

$$IG_{Gender} = E_{before} - E_{after} = 1 - 0.604185 = 0.395815$$



- 2. Location

		Respond		Total
		Y	N	
Location	A	3	2	5
	B	1	2	3
	C	1	1	2
Total		5	5	10

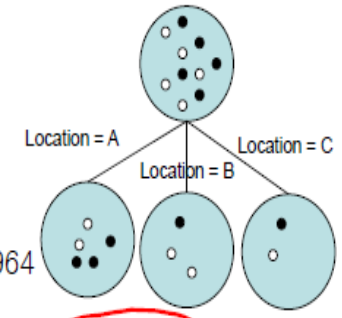
$$E_{before} = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1$$

$$E_{left} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.970951$$

$$E_{middle} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918296$$

$$E_{right} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$E_{after} = \frac{5}{10}\times 0.970951 + \frac{3}{10}\times 0.918296 + \frac{2}{10}\times 1 = 0.960964$$



$$IG_{location} = E_{before} - E_{after} = 1 - 0.960964 = 0.039036$$

IG값이 큰 Gender가 분기 기준으로 선택!

NO. 01 Evaluating Model Performance

Contingency table–Information Gain Ratio

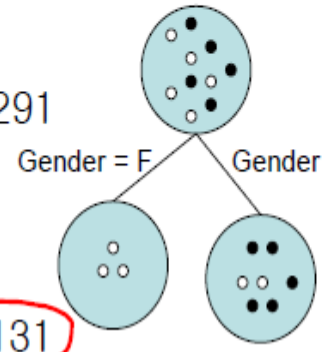
예) 무작위의 n개 분할의경우, -1* ((1/n)*log2(1/n)+…+(1/n)*log2(1/n)) = -log2(1/n)

- 1. Gender

		Respond		Total
		Y	N	
Gender	F	0	3	3
	M	5	2	7
Total		5	5	10

$IG_{Gender} = 0.395815$

$$IV_{gender} = -\frac{3}{10}\log_2\frac{3}{10} - \frac{7}{10}\log_2\frac{7}{10} = 0.881291$$



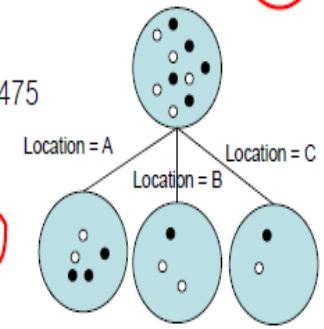
$$IGR_{gender} = \frac{IG_{gender}}{IV_{gender}} = \frac{0.395815}{0.881291} = 0.449131$$

- 2. Location

		Respond		Total
		Y	N	
Location	A	3	2	5
	B	1	2	3
	C	1	1	2
Total		5	5	10

$IG_{location} = 0.039036$

$$IV_{location} = -\frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10} = 1.485475$$



$$IGR_{location} = \frac{IG_{location}}{IV_{location}} = \frac{0.039036}{1.485475} = 0.026278$$

IGR값이 큰 Gender가 분기 기준으로 선택!

NO. 01 Evaluating Model Performance

Kappa statistic

우연히 정확한 예측할 확률 – 값=0~1
Poor = 0.2미만
Fair = 0.2~0.4
Moderate = 0.4~0.6
Good = 0.6~0.8
Very good = 0.8~1

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Pr(a) =
Pr(Actual A = Predicted A)
+
Pr(Actual B = Predicted B)

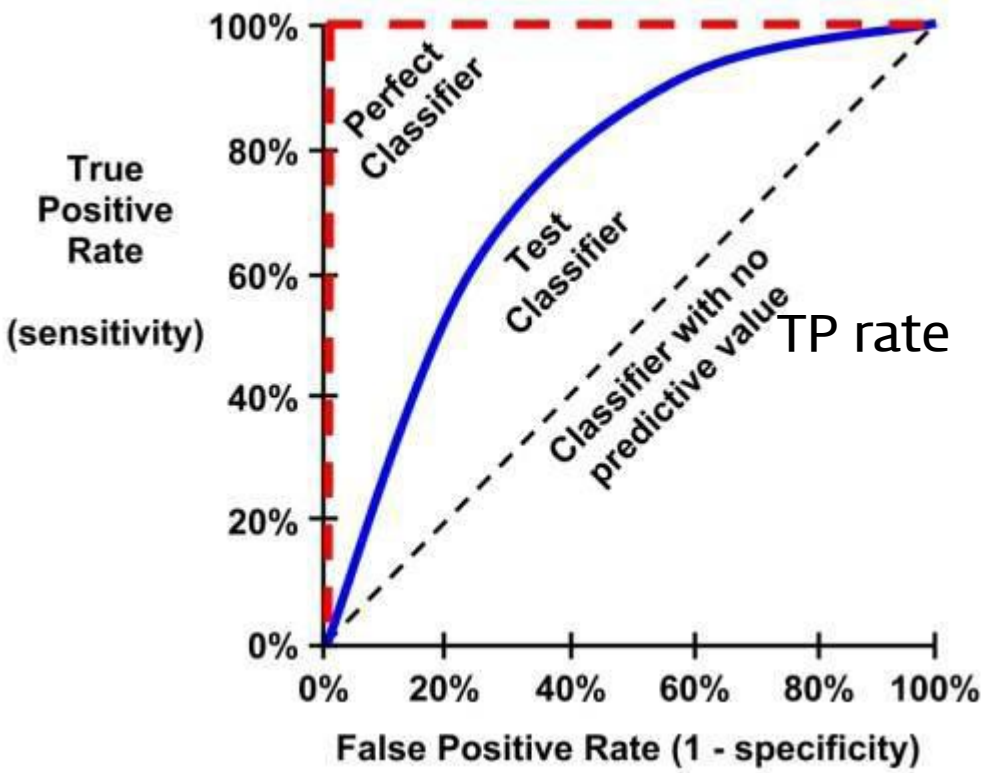
Pr(e) =
Pr(Actual A)*Pr(Predicted A)
+
Pr(Actual B)*Pr(Predicted B)

=> 그러나 0부터 1까지 값으로 판단하는 것은 지극히 주관적

NO. 01 Evaluating Model Performance

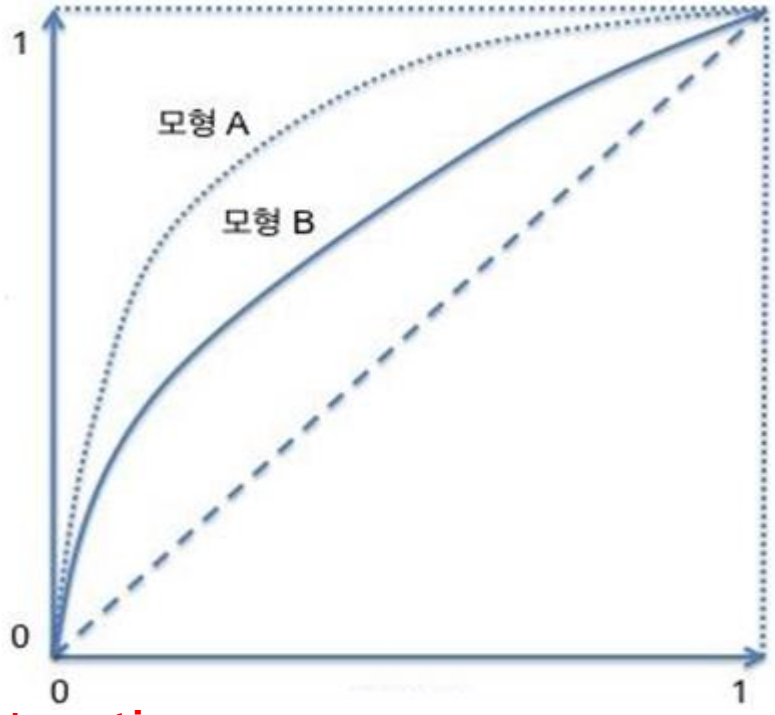
ROC 그래프

ROC(Receiver Operating Characteristic)
X축은 FP 비율(1-Specificity)
Y축은 TP 비율(Sensitivity)



TP rate

All Positive



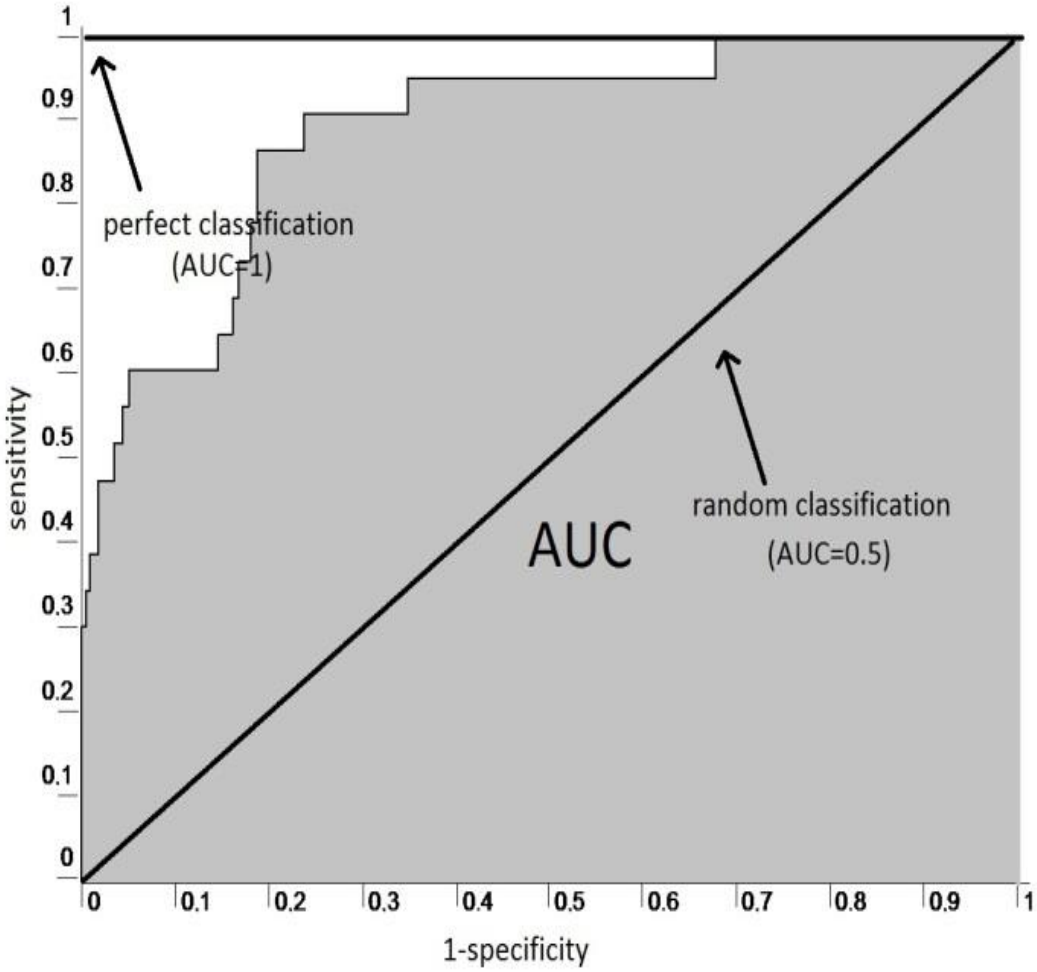
All Negative

NO.01 Evaluating Model Performance

AUC

Area Under the ROC curve
1에 가까울 수록 좋은 모형

- 0.9 - 1.0 = A (outstanding)
- 0.8 - 0.9 = B (excellent/good)
- 0.7 - 0.8 = C (acceptable/fair)
- 0.6 - 0.7 = D (poor)
- 0.5 - 0.6 = F (no discrimination)



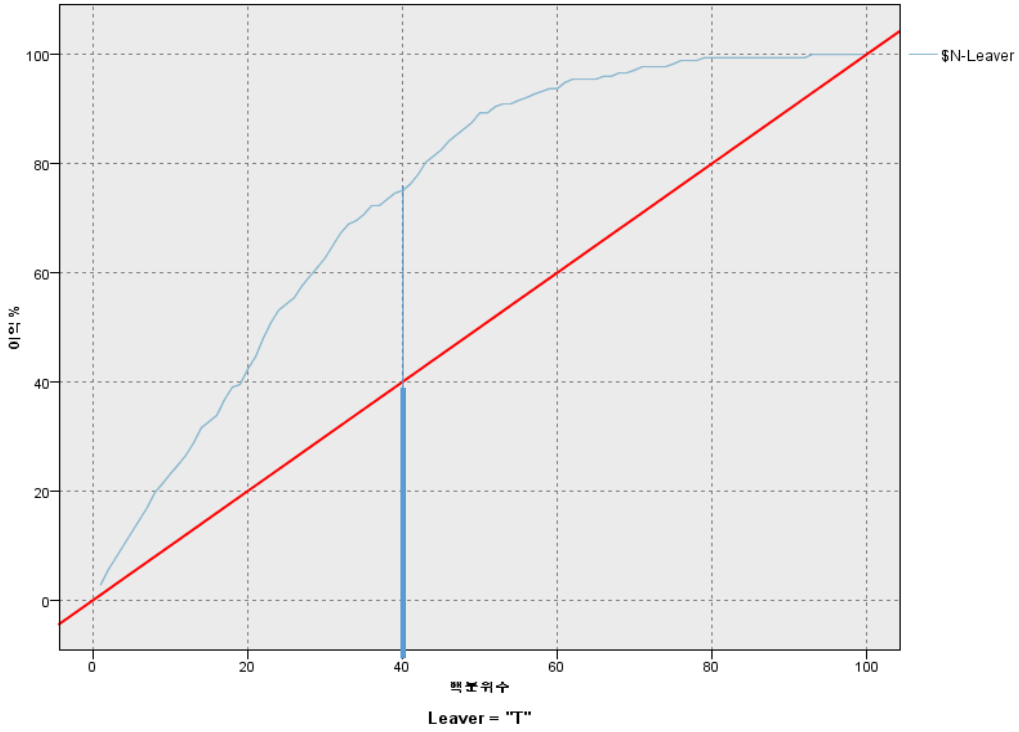
NO. 01 Evaluating Model Performance

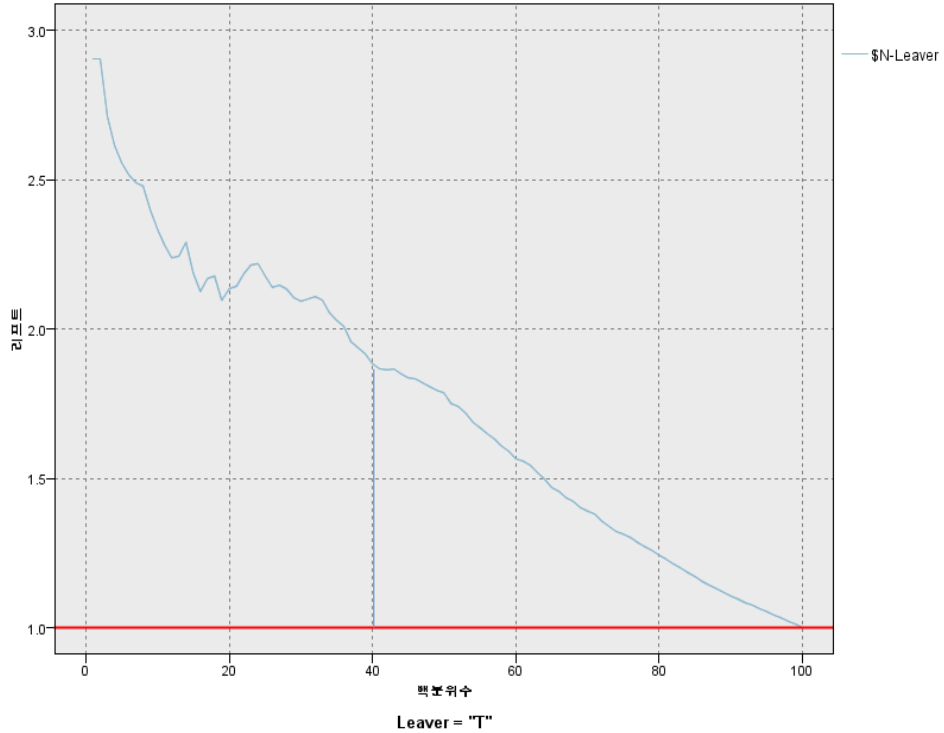
Gain Chart

모형을 사용하여 데이터를 선택하였을 때, 목표 집단의 레코드가 몇 퍼센트나 포함되는지에 대한 도표

Lift

무작위로 선정한 것에 비해서 향상된 정도





Ganji C조
Chapter 10-1

1. Evaluating Model Performance

NO.01 Evaluating Model Performance

But 예측 방법은?

일반적인 예측 모델의 성능 평가(회귀모형 등)

MAE(평균절대오류) $= \frac{1}{n}\sum|e_i|$

Average error(평균오류) $= \frac{1}{n}e_i$

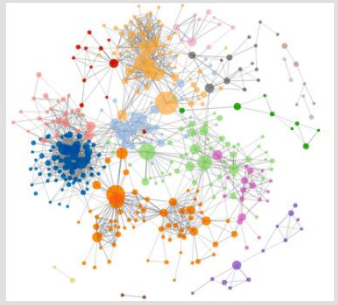
MAPE(평균절대 백분율오류) $= 100\% \times \frac{1}{n}\sum\left|\frac{e_i}{y_i}\right|$

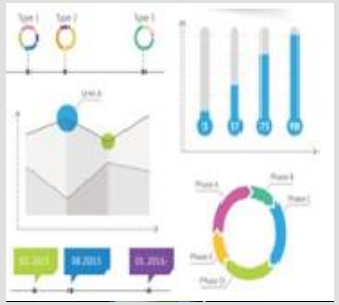
RMSE $= \sqrt{\frac{1}{n}\sum {e_i}^2}$

SSE(제곱오류 총합) $= \sum {e_i}^2$

CHAPTER 02

Cross Validation;





1. Cross Validation

NO. 01 Cross Validation

분류 모델링의 기본 구조

Training data를 통한 학습 → Validation data를 통한 검증
Why? Validation data는 학습에 사용하지 않음으로써 모델이 overfitting되는 것을 방지한다

Training data의 학습을 통해 각 종류에 속할 확률을 구한 뒤, 그 확률을 기반으로 어떤 종류로 분류할 지 판단

(0, 1)두 종류로 구분한다고 할 때, '1'일 확률이 몇 이상이면 '1'로 분류할 것인가?
→ Cutoff value : 0.5? 꼭 0.5만이 정답은 아니다.

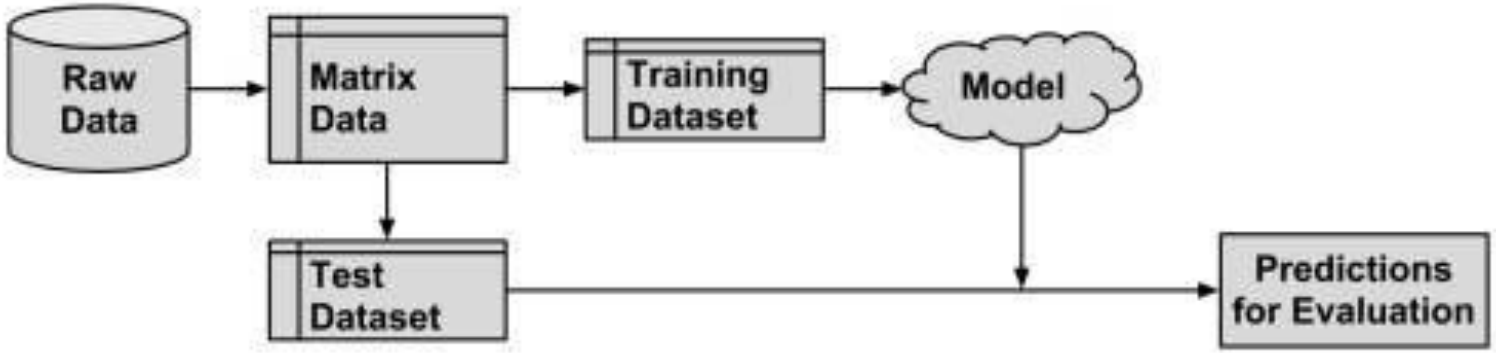
NO. 01 Cross Validation

1. k-fold cross validation

데이터의 (k-1)/k 만큼을 training data, 1/k 만큼을 validation data로 사용
k는 자유롭게 설정, 보통 10을 사용

NO.01 Hold-out 기법

일반적으로 7:3(훈련용:평가용) or 5:3:2(훈련용:검증용:평가용)으로 사용



범주의 비율이 낮거나 높을 수 있어서 층화 무작위 표본 추출 기법 을 사용
But 대표성을 보장하지 못할 가능성 有, 일부 수가 너무 많거나 적은 경우 편향될 수 있음

-> 층화 Hold-out 기법 – 특정 변수를 일정한 비율로 포함하게 나눠줌

NO. 01 Cross Validation

2. Leave One Out Cross-Validation

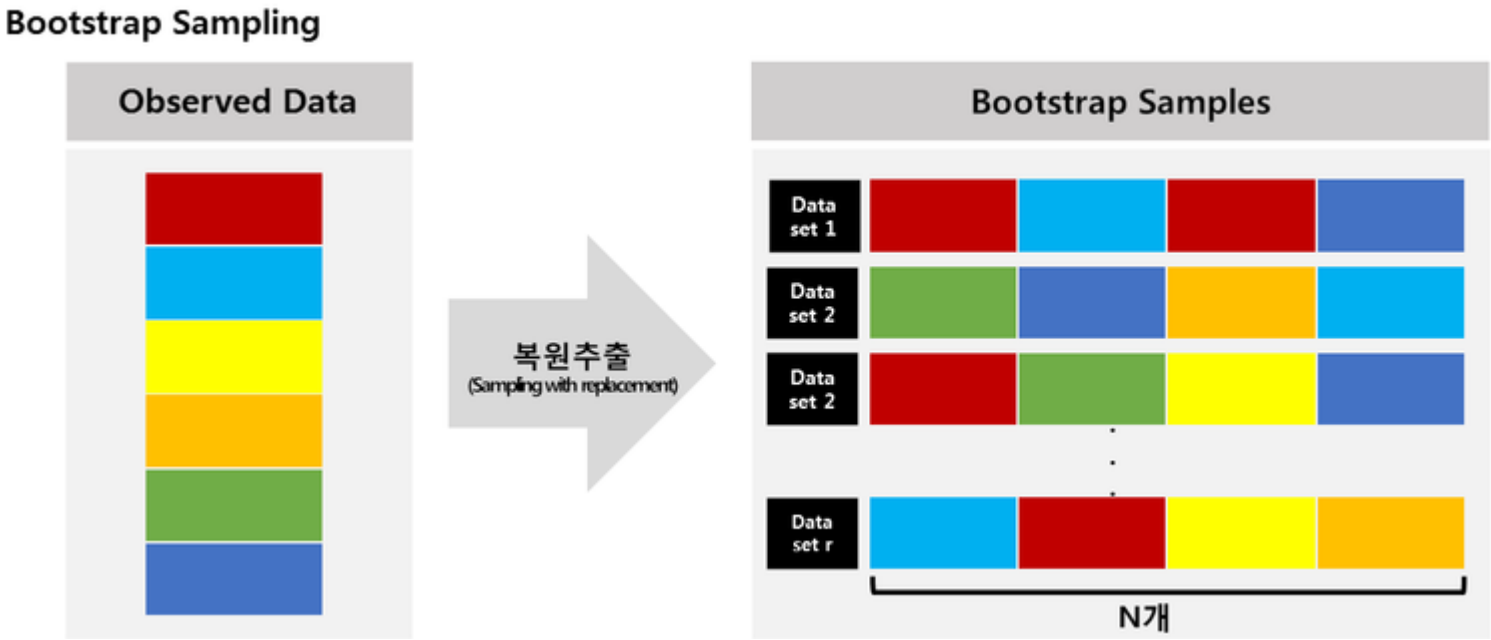
총 데이터가 N개 있다고 하면, 1개를 제외한 (N-1)개를 training data, 나머지 1개를 validation data로 하여 모델링을 진행한다.

이 같은 과정을 N번 반복하여 각 오차의 평균을 오차로 사용

Training에 사용할 데이터 자체의 수가 적을 때 사용한다.
모든 데이터에 대해 한번씩 다 test를 거치기 때문에 stable한 결과를 얻을 수 있지만 테스트가 결과적으로 N번 진행되어야 하므로 시간이 오래 걸리는 작업이다.

NO.01 Cross Validation

3. Boot Strap Sampling



훈련용 데이터를 복원 추출한다. 한번 추출된 데이터를 전체 데이터 집합에서 빼지 않고 다음 데이터를 추출 한다. 전체 데이터의 수가 N일 때 Boot Strap 방법으로 N개의 데이터를 추출하면 대략 전체 데이터의 63.2% 정도가 훈련용 데이터로 추출된다. 이는 각 데이터가 부트스트랩 표본으로 추출될 확률이 $1-(1-\frac{1}{N})^N$ 이고, 이 확률은 N이 충분히 클 때 점진적으로 $1-e^{-1}=0.632$ 에 수렴하기 때문이다. Boot Strap방법으로 추출되지 않은 표본은 수립하고, 이 모형을 시험용 데이터에 적용하여 정확도를 조사한다. 유사한 실험을 r번 반복하여 측정한 정확도들의 평균을 전체 모형의 정확도로 한다.

[C조] 구민수 구유림 윤소라 손범호