05.머신러닝 모델의 평가지표

- 지도학습 모델의 평가지표
  - 회귀(Regression) 모델의 평가 지표
    - RMSE (Root mean square error) :
      - 정밀도(precision)을 표현하는데 적합하며, 각각의 차이값은 잔차(residual)라고 함
      - 작을 수록 좋음

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

    - R-square (결정계수)
      - 선형회귀 모델일 경우 언제나 0에서 1사이 값을 지님
      - 0이면 모형이 굉장히 안좋은 상태
      - 1이면 완벽한 피팅상태, 오차가 0인 상태를 의미
      - 클수록 좋으며, 1에 가까울수록 좋음

$$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

    - MAE(mean absolute error)
      - 오차의 부호만 제거해서 이를 평균한 값
      - MAE가 10이면 오차가 평균적으로 10정도 발생한다는 것을 의미
      - 오차의 평균개념이므로 작을수록 좋음

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

    - MAPE(mean average percentage error)
      - 실제 값 대비 오차가 차지하는 비중이 평균적으로 얼마인지 확인

- 오차의 평균개념이므로 작을수록 좋음

$$100 \times \left(\frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right)$$

- 분류(Classification) 모델의 평가 지표
  - 정오분류표(confusion matrix) = 교차분류표

정오분류표		모형에 의한 예측	
		Negative	positive
실제 자료	Negative	A (TN, true negative)	B (FP, false positive)
	Positive	C (FN, false negative)	D (TP, true positive)

  - 정오분류표 설명
    - TN : 아니라고 추측한 값이 맞을 경우, 예측값 Y가 1, 실제값 Y가 1 일 경우 : 35건
    - FP : 맞다고 추측한 값이 틀릴경우, 예측값 Y가 1, 실제값 Y가 0 일 경우 : 12건
    - FN : 아니라고 추측한 값이 틀릴경우, 예측값 Y가 0, 실제값 Y가 1 일 경우 : 7건
    - TP : 맞다고 추측한 값이 맞을 경우, 예측값 Y가 0, 실제값 Y가 0 일 경우 : 40건

ID	X1	...	Xk	Y	P(Y=1) 예측값	Y예측값
1	0.5736		0.5	1	0.9960	1
2	0.9876		0.2	1	0.9875	1
3	0.4366		0.7	1	0.9845	1
4	0.8791		0.3	1	0.8893	1
5	0.8462		0.0	0	0.7628	1
6	0.2198		0.4	1	0.7070	1
7	0.2911		0.2	0	0.6808	1
89	0.1512		0.4	0	0.0480	0
90	0.9824		0.1	0	0.0383	0
91	0.6375		0.7	1	0.0249	0
92	0.4177		0.7	1	0.0218	0
93	0.0116		0.0	0	0.0161	0
94	0.5114		0.4	0	0.0036	0

분류기준값 : 0.5		예측범주	
		0	1
실제 범주	0	40	12
	1	7	35

  - 정확도, 정분류율(Accuracy)

- 전체 관찰치 중 정분류된 관찰치의 비중
- 실제 분류모델을 평가할때 정확도만으로는 불안정한 경우가 많음
- 실제 분류모델에선 관심범주와 실제범주가 1:1이 아니기 때문
- 관심범주의 비중이 작을 경우, 정확도는 높지만 예측이 좋다고 보기 어렵다.

$$\frac{A+D}{A+B+C+D}=\frac{TN+TP}{TN+FP+FN+TP}$$

		예측	
		Negative	positive
실제	Negative	A (TN)	B (FP)
	Positive	C (FN)	D (TP)

- 정밀도(Precision)
  - Positive로 예측한 것 중에서 실제 범주도 Positive인 데이터의 비율
  - FP가 작을수록 좋음

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

- 재현율(Recall)
  - 실제 범주가 Positive인 것 중에서 Positive로 예측된 데이터의 비율
  - FN가 작을수록 좋음

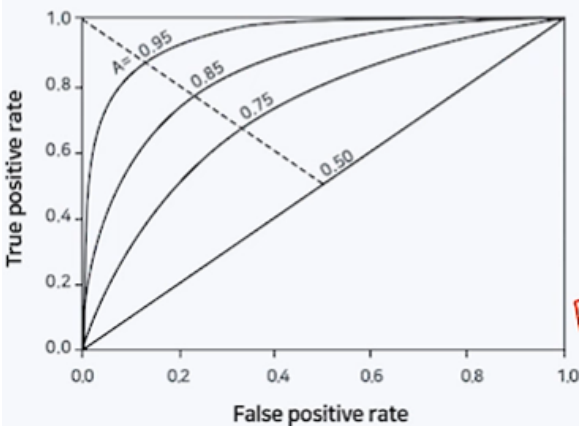
$$\frac{D}{C+D}=\frac{TP}{FN+TP}$$

- ROC(Receiver operating characteristic) 도표
  - 분류의 결정임계값(threshold)에 따라 달라지는 TPR(민감도, sensitivity)과 FPR(1-특이도, 1-specificity)의

조합을 도표로 나타냄

- Threshold(결정임계값) : 예측하는 범주 확률을 정하는 값
- TPR : True Positive Rate (=sensitivity(민감도))
  - 1인 케이스에 대해 1로 잘 예측한 비율
  - Positive로 예측했는데 True인 것
- FPR : False Positive Rate (=1-specificity(특이도))
  - 0인 케이스에 대해 1로 잘못 예측한 비율
  - Positive로 예측했는데 False인 것

- Threshold가 1이면 FPR=0, TPR=0
- Threshold를 1에서 0으로 낮춰감에 따라 FPR과 TPR은 동시에 증가
- FPR이 증가하는 정도보다 TPR이 빠르게 증가하면 이상적
  > 왼쪽 위 꼭지점에 가까울수록 좋음



- AUC(Area Under the Curve)
  - ROC 곡선 아래의 면적.
  - 가운데 대각선의 직선은 랜덤한 수준의 이진분류에 대응되며,
    TPR과 FPR이 같은 비율로 증가하는 상태
    이 때 전체 크기가 1이므로 AUC는 0.5이고
    이 경우가 그래프에서 나타내는 가장 안좋은 케이스
  - 0.5 보다는 커야하며, 1에 가까울수록 좋은 수치.
  - FPR이 작을 때 얼마나 큰 TPR을 얻는지에 따라 결정됨