머신러닝 과정

데이터 가공
1. 데이터 클렌징
2. 데이터 확인 (시각화)
3. 데이터 전처리

데이터 분석

모델 학습 예측
4. 머신러닝 알고리즘에 적합한 데이터로 준비
5. 머신러닝 알고리즘 적용

인코딩 — scikit-learn, tensor, pytorch

평가 반복
6. 머신러닝 알고리즘 평가
7. 머신러닝 알고리즘 개선
8. 결과 나올 때까지 1~7 반복

- 회귀와 분류
  - 평가 지표가 다르다.
  - 결과 차이 분명
  - 회귀: RMSE, RMSLE
  - 분류: F1 score

- 회귀 평가 준비

MAE	Mean Absolute	$\frac{1}{N}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$	실제, 예측 값 차이 절댓값 변환 ⇒ 평균
MSE	Mean Squared	$\frac{1}{N}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	차이제곱 평균
RMSE	Root MS	$\sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$	$\sqrt{MSE}$ MSE는 너무 큼.
MSLE	Mean squared log	$\frac{1}{N}\sum_{i=1}^{N}(\log(y_i+1)-\log(\hat{y}_i+1))^2$	MSE에 log 적용 log(y+1)
RMSLE	Root MLS	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i+1)-\log(\hat{y}_i+1))^2}$	$\sqrt{MSLE}$
R^2	R Square	예측값 분산 / 실제값 분산	

- 분류 평가 지표
  - 정확도 : $\frac{\text{예측결과 동일 data}}{\text{전체 예측 data}}$ 직관적, label 불균형 한 경우 성능 판독 가능성

오차 행렬 (혼동 행렬) confusion matrix

실제 클래스 y	예측 클래스 ŷ: Negative 0	예측 클래스 ŷ: Positive 1
Negative 0	TN (F→F)	FP (F→T)
Positive 1	FN (T→F)	TP (T→T)

⇒ 예측 클래스에 대한 평가
① 참이냐 거짓이냐?
② 예측을 어떻게 했냐?

* 오차 행렬에서의 정확도 $= \frac{TN+TP}{TN+FP+FN+TP}$

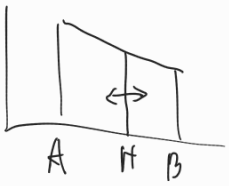
- 정밀도와 재현율 : positive 데이터에 초점
- 정밀도 $\frac{TP}{FP+TP}$ 양성 예측도 — 모델이 true로 분류한 데이터 중 실제 true 비율.
  ex) 예측한 날씨가 실제로 얼마나 맞았나?
- 재현율 $\frac{TP}{FN+TP}$ 민감도 — 실제 true인 데이터 중 모델이 true로 분류한 데이터 비율.
  ex) 실제 날씨가 맑은 날에 대해 모델이 맑다고 예측한 비율.

* 정확도 관계.

- F1 스코어 : 정밀도, 재현율 결합한 지표.

$$F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 \times \frac{precision \times recall}{precision + recall}$$

(조화평균)
산술평균에 비해 bias가 ↓



⇒ label이 불균형한 경우에도 적용 가능.

- ROC 곡선과 AUC

이진 분류 모델 성능 측정

$$-FFR = \frac{FP}{FP+TN}$$