**<기타 분석 용어 정리>**

* 다중 공선성의 정의

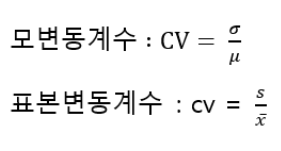
회귀 분석의 다중 공선성은 모형의 일부 예측 변수가 다른 예측 변수와 상관되어 있을 때 발생하는 조건이다. 중대한 다중 공선성은 회귀 계수의 분산을 증가시켜 불안정하고 해석하기 어렵게 만들기 때문에 문제가 된다.

* VIF(Variation Inflation Factor):분산 팽창 계수

다중공선성을 파악하기 위한 수치적 지표로 예측 변수들이 상관성이 있을 때 추정 회계 계수의 산포 크기를 측정하는 것이며, 산포가 커질 수 록 회귀 모형은 신뢰 할 수 없게 된다.

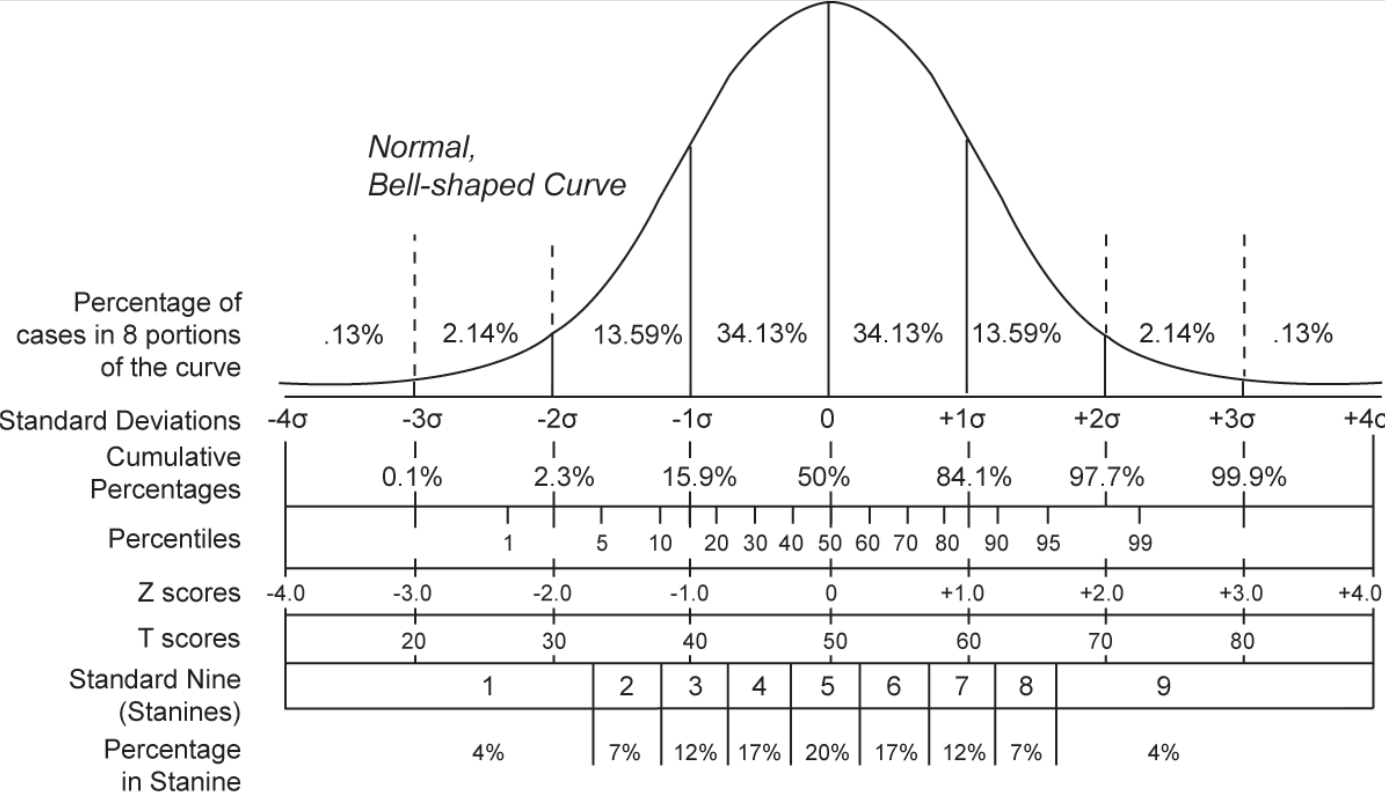
* 변동계수( coefficient of variation)

각 데이터 집단끼리 표준편차를 통해서 변동성을 나타내는 지표를 비교하려할 때 , 데이터 집단끼리 단위가 다르면 비교조차 할 수 없다. 이를 해결하기 위해서 변동계수를 사용한다. 변동계수는 표준편차를 평균으로 나누어 계산한다. 이를 통해 다른 데이터 관측치들로부터 서로 변동성에 대해 비교 할 수 있게된다.

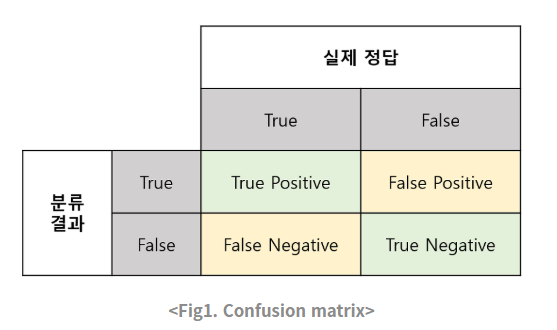


* 표준 점수(Z score)

통계학적으로 정규분포를 만들고 개개의 경우가 표준편차상에 어떤 위치를 차지하는지를 보여주는 수치



* P-value
* 분류 성능 평가 지표: Precision(정밀도), Recall(재현율), Accuracy(정확도), f1-score  
  (참조: <https://sumniya.tistory.com/26>)



True Positive(TP): 실제 True인 정답을 True로 예측(정답)

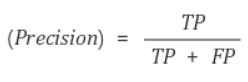
False Positive(FP): 실제 False인 정답을 True로 예측(오답)

False Negative(FN): 실제 True인 정답을 False로 예측(오답)

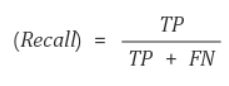
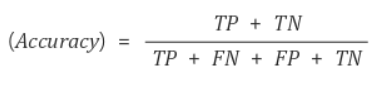
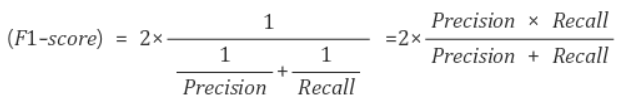
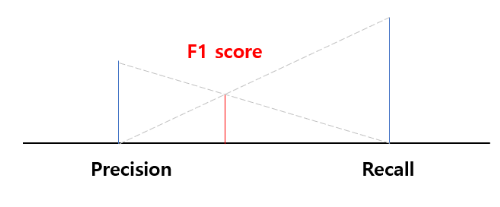
True Negative(TN): 실제 False인 정답을 False로 예측(정답)

1. Precision(정밀도)

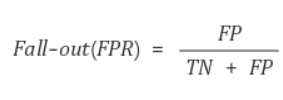
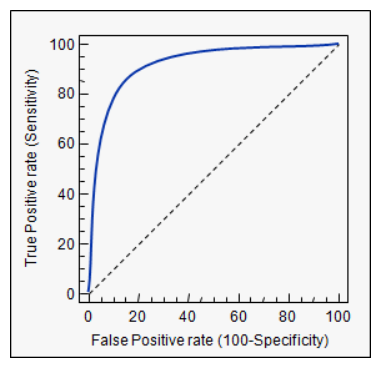
모델이 True라고 분류한 것 중에서 실제 True인 것의 비율



PPV(Positive Predictive Value)라고도 불리며, 날씨 예측 모델이 ‘맑다’로 예측했는데, 실제 날씨가 맑았는지 살펴보는 지표

1. Recall(재현율)  
   실제 True인 것 중에서 모델이 True라고 예측한 것의 비율  
     
   sensitivity 또는 hit rate라는 용어로 사용되기도 한다. 실제 날씨가 맑은 날 중에서 모델이 맑다고 예측한 비율을 나타낸 지표  
   (Precision 은 모델의 입장에서, Recall 은 실제 정답(data)의 입장에서 정답을 맞춘 경우를 뜻함 🡪 두 값을 함께 고려해야 제대로 된 모델의 평가가 된다.)
2. Accuracy(정확도)  
     
   가장 직관적으로 모델의 성능을 나타낼 수 있는 평가 지표이다. 하지만 domain의 bias를 고려해야한다. 예를 들어, 예측하고자 하는 한달 동안 특정 기후에 부합하여 비 오는 날이 흔치 않다고 생각할 경우, 해당 data의 domain이 불균형하게 되므로 예측하는 성능은 높지만, 비가 오는 것을 예측하는 성능은 매우 낮을 수 밖에 없다. 따라서 이를 보완할 지표가 필요하다.
3. F1 score  
   Precision과 Recall의 조화 평균이다.  
     
   F1 score는 데이터 label이 불균형 구조일 때, 모델의 성능을 정확하게 평가할 수 있으며, 성능을 하나의 숫자로 표현할 수 있다.   
   

* 그 외 다른 지표들

1. Fall-out  
   FPR(False Positive Rate)으로도 불리며, 실제 False인 데이터 중에서 모델이 True라고 예측한 비율이다.   
   
2. ROC(Receiver Operating Characteristic) curve  
   여러 임계값들을 기준으로 Recall-Fallout의 변화를 시각화 한 것이다.   
     
   curve가 왼쪽 위 모서리에 가까울수록 모델의 성능이 좋다고 평가된다. 즉 Recall이 크고 Fall-out이 작은 모형이 좋은 모형이다.
3. AUROC  
   ROC curve는 그래프이기 때문에 명확한 수치로 비교하기 어렵다. 따라서 그래프 아래의 면적 값을 이용하게 된다. 최대값은 1이며 좋은 모델일수록 1에 가까운 값이 나온다.