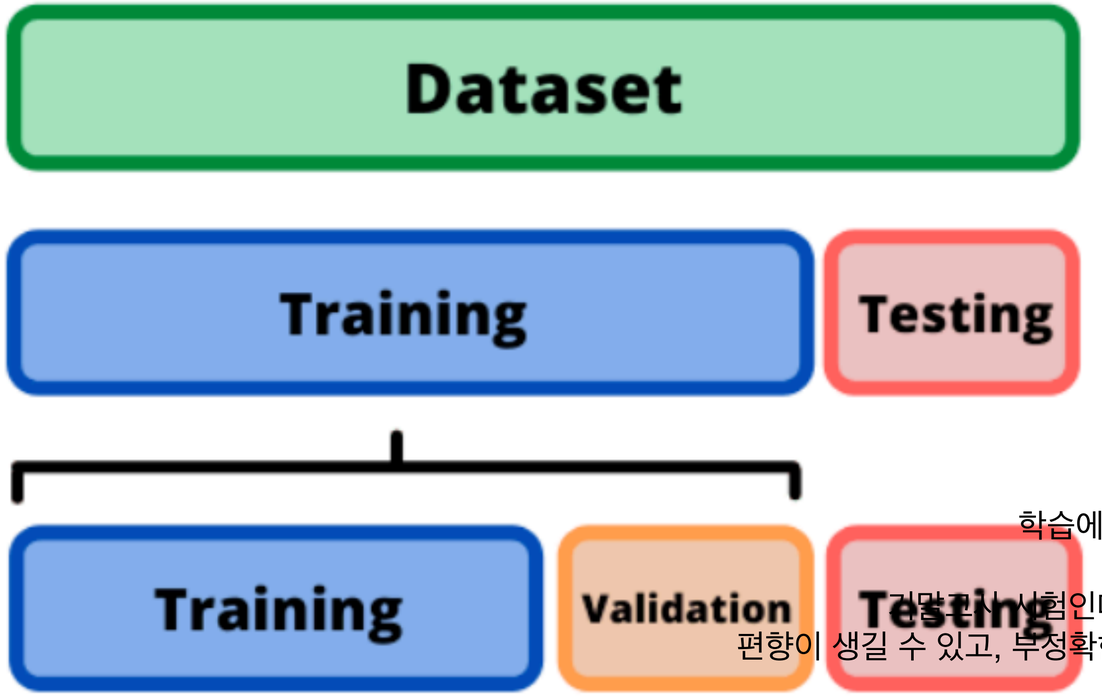
Train/Test/Validation

20240203_7기_파문응수업자료

Train/Test/Validation

과적합과 일반화를 하기 위해서 이런 작업들을 진행한다.



validation은 무엇인가?
test 검증하기 전에 train data를 가지고 하나의 test 를 다시 해보는 것
검증하는 것

왜 데이터를 나눌까?

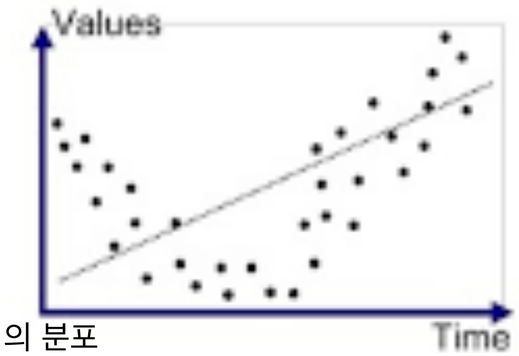
train 데이터로만 학습 -> 검증을 test
학습을 할 때는 test 데이터를 사용하면 안 된다.

학습에 test 까지 넣게되면 -> 모델은 test까지 패턴을 학습하고 기억한다.
test 데이터를 넣으면 좋은 성적이 나온다.
편향이 생길 수 있고, 부정확한 결과가 나온다. (test는 모르는 상태에 봐야하는데 문제를 미리 알게 된 거니 이 사람이 시험인데, 선생님이 몇 문제 나온다고 알려줬다. 최소한 이 몇 문제는 맞출 확률 높다.
진짜 이 문제를 풀수 있는 능력이 있는지 의문 ?)

test로 검증이라는 것은? 일반화를 하기 위한 작업
23년 1년치 데이터를 가지고 24년 예측 하는 경우도 생길 것
23년의 데이터를 가지고 Train/test 나누고 -> 그 후에 24년들의 데이터의 예측을 진행할 것
vip 고객 유무

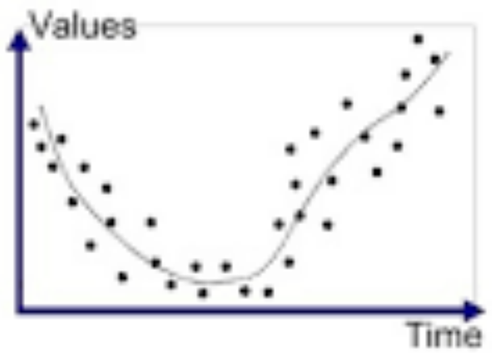
Overfitting, Underfitting

Overfitted , Underfitted 되는 이유?

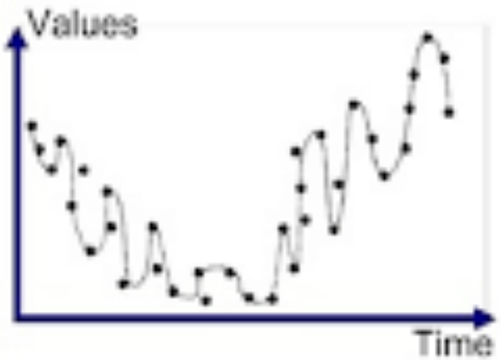


데이터는 비선형의 분포
모델은 선형적 예측
데이터 부족한 것도 있고, 모델이 너무 단순한 경우

Underfitted



Good Fit/Robust

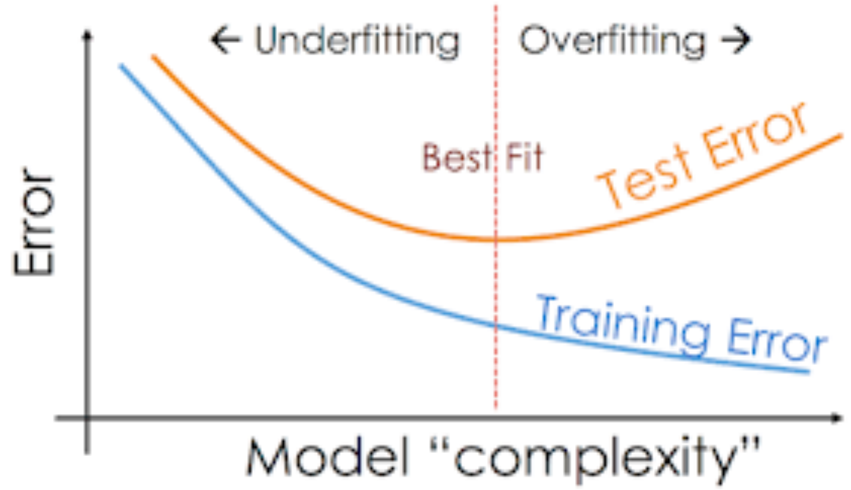


Overfitted

train data에 너무나 학습을 잘 하다보니
이 모든 패턴을 다 외운 것

만약 이 train data 안에 대부분 학생이 전국 상위 10%
데이터가 대부분을 이루고 있다.

일반화로 실제 나머지 90% 나머지를 예측할 때
기준이 너무 높게 잡힌 상태



Generalization

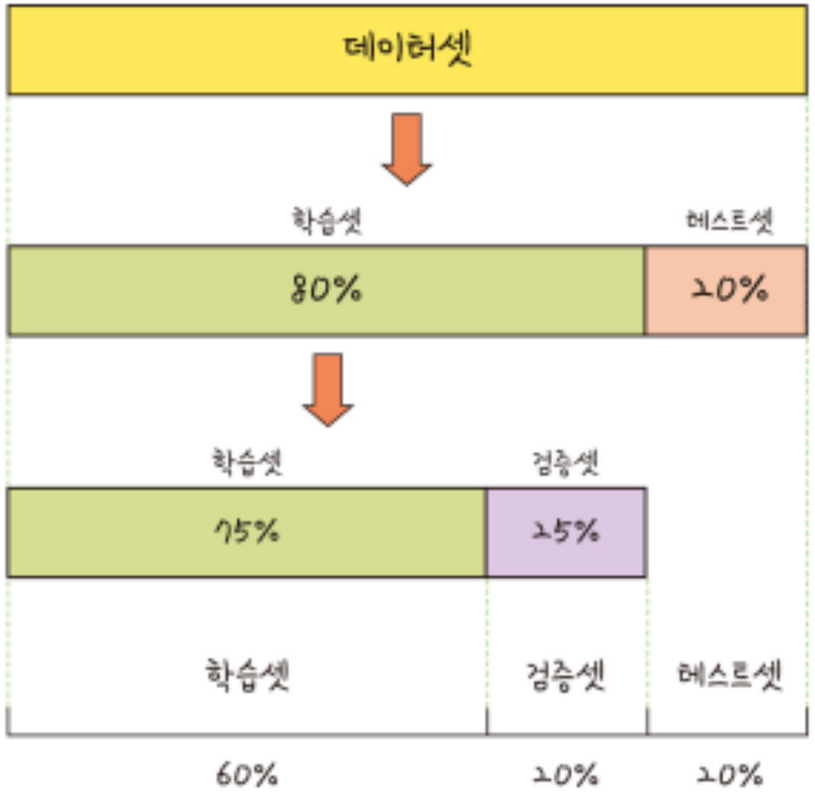
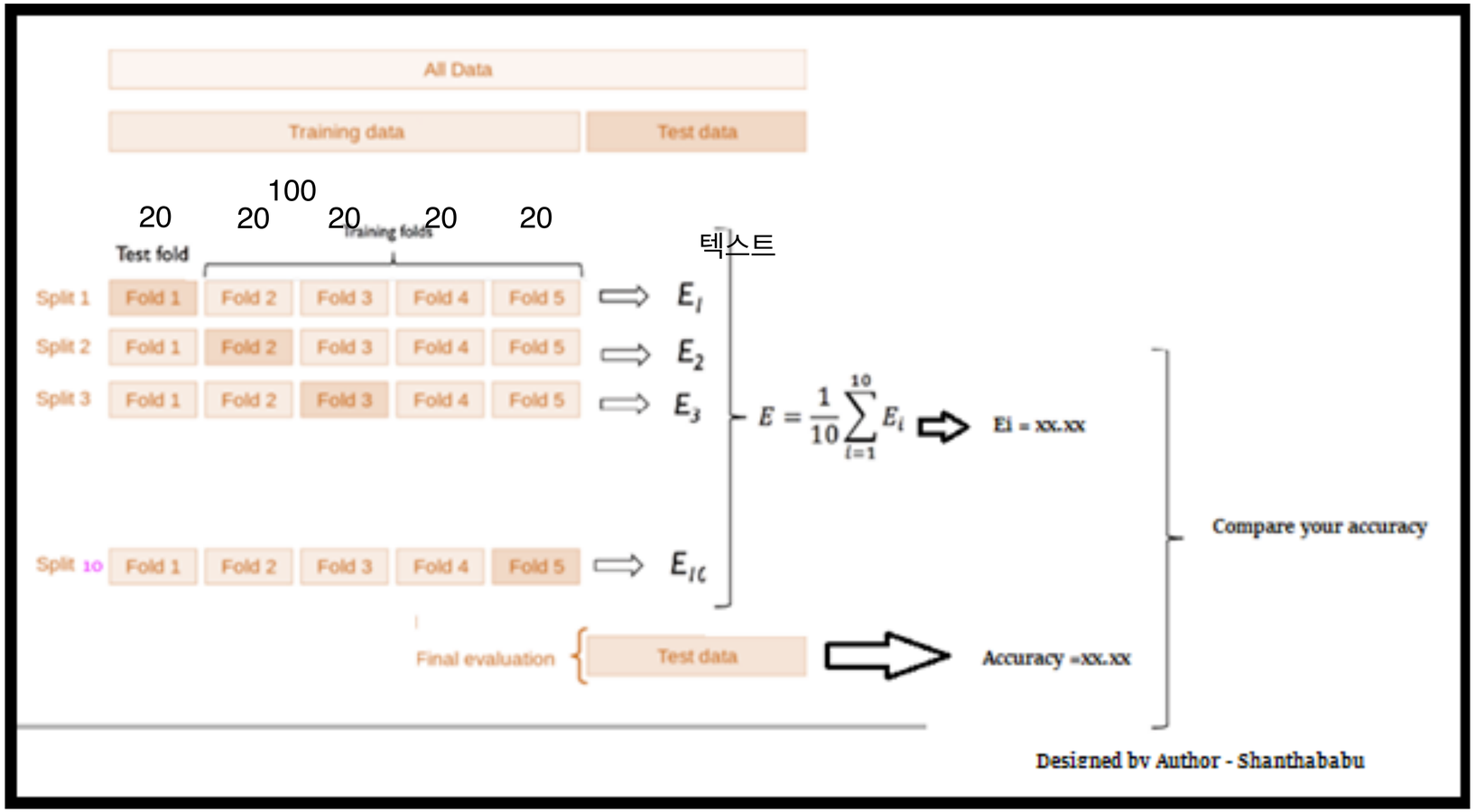
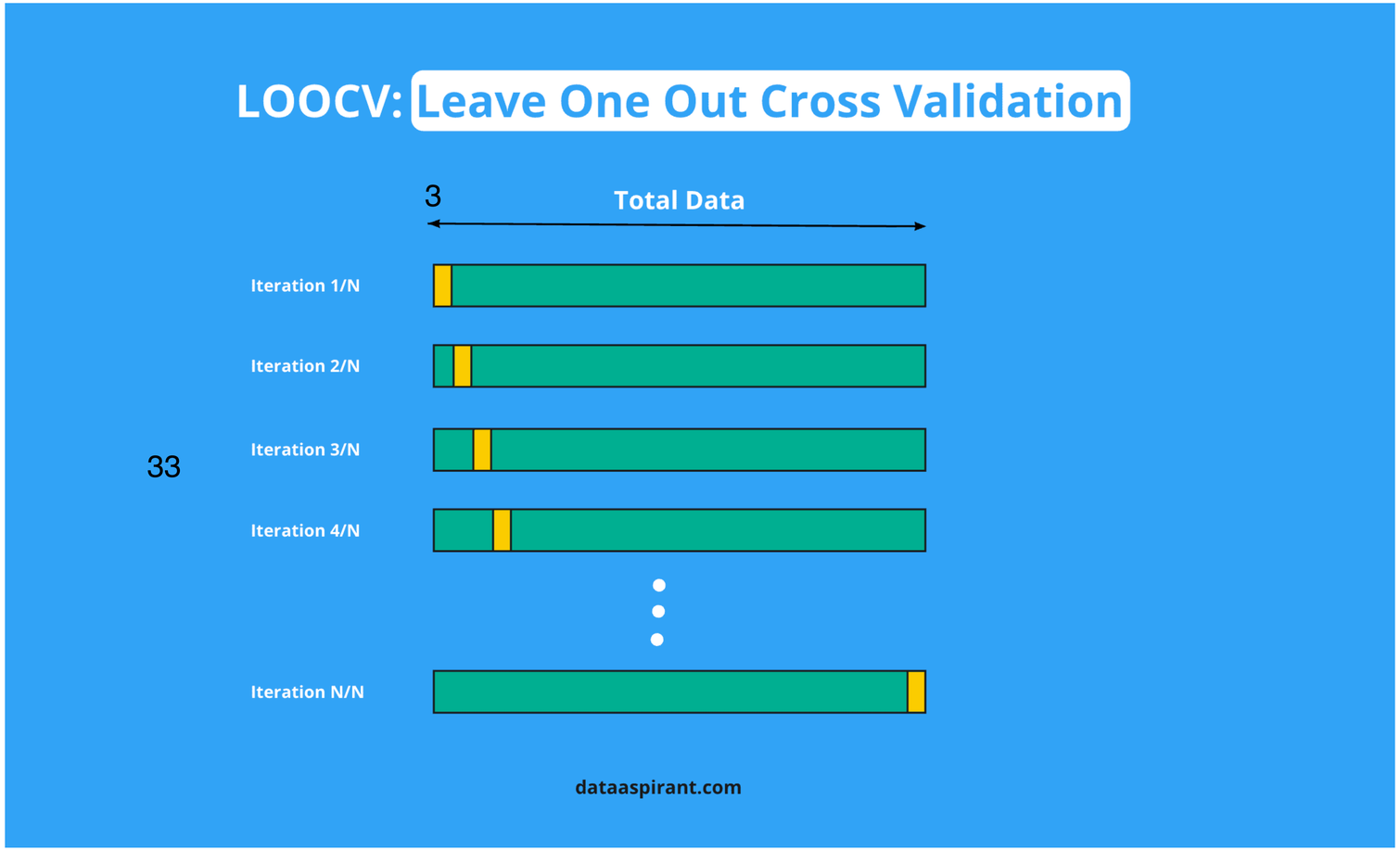


그림 14-1 | 학습셋, 테스트셋, 검증셋

k-Fold Cross Validation

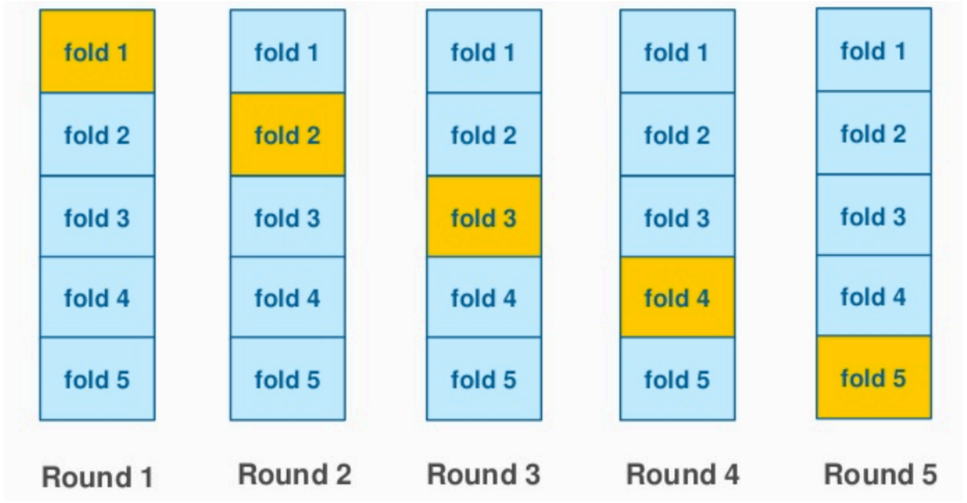


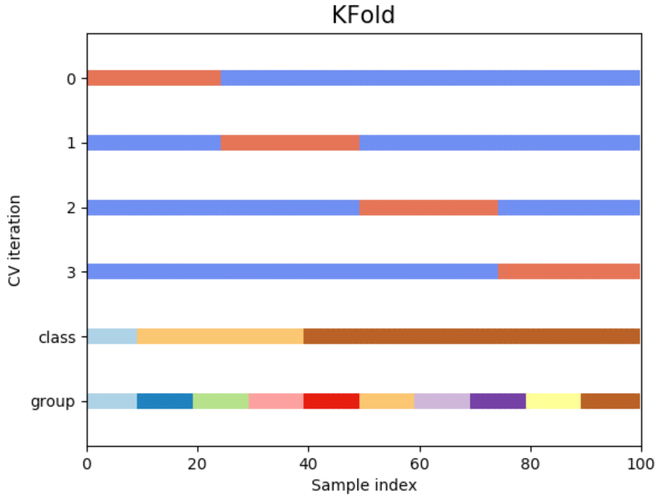
LOOCV(Leave-one-out Cross-validation)

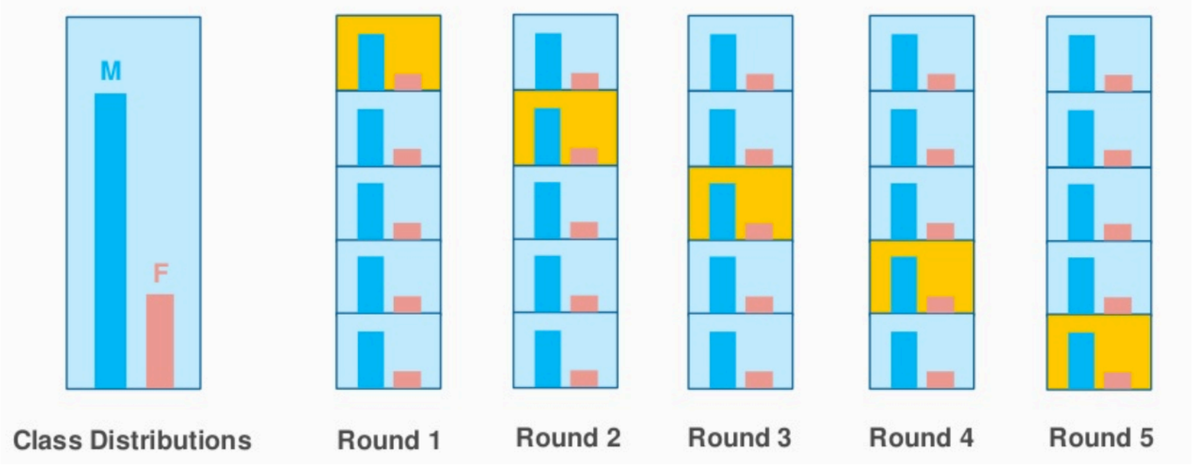


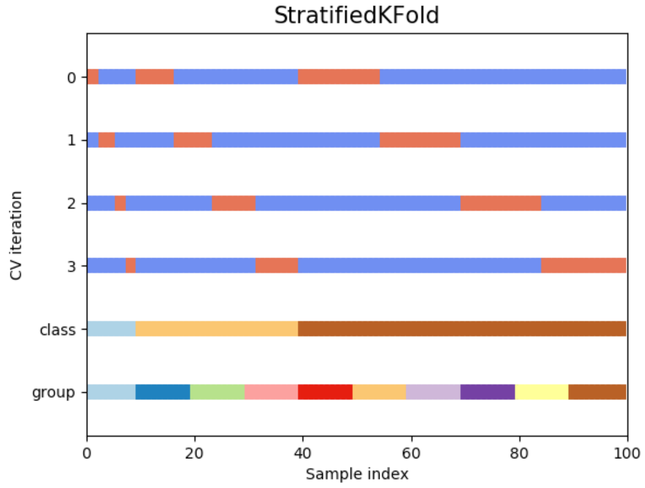
Stratified K-fold

기존의 fold 방식들의 문제점 > 클래스의 불균형을 잡지 못해서 발생하는 이슈

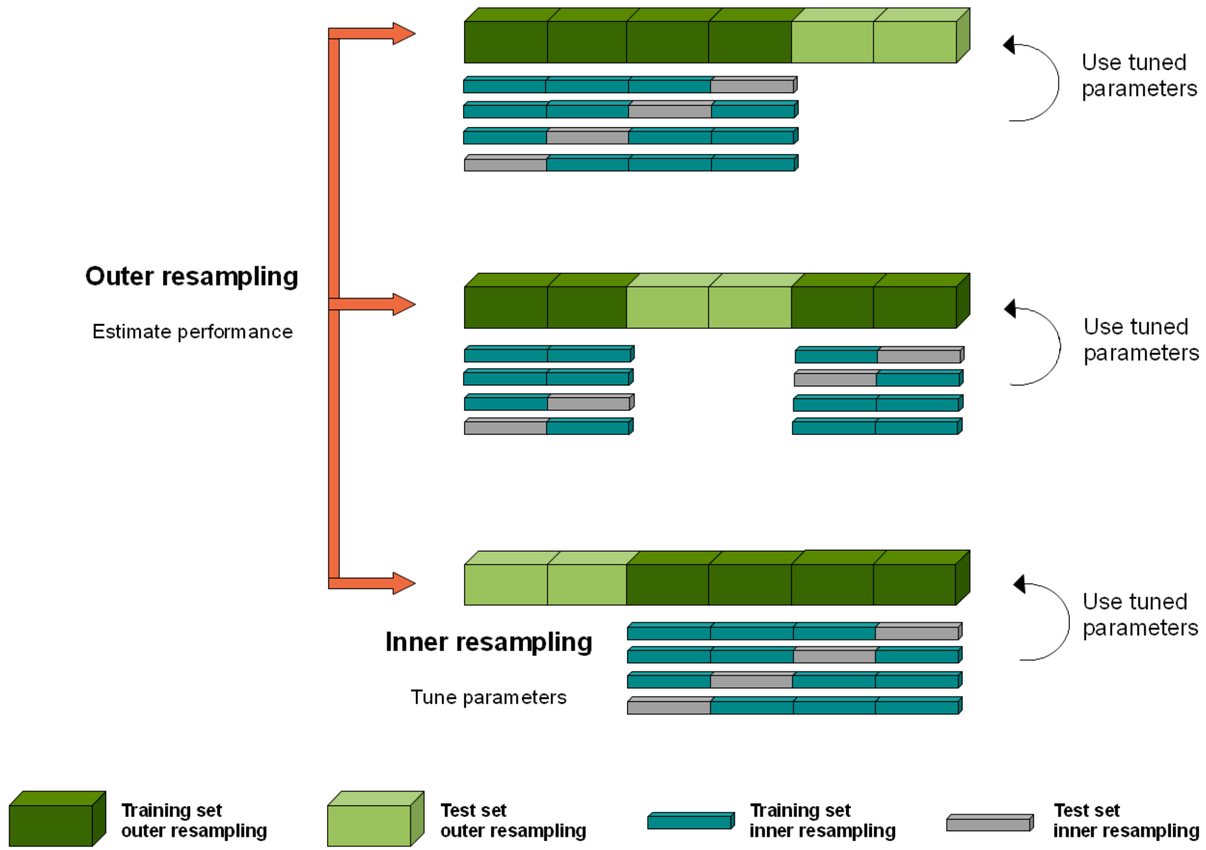








Nested Cross Validation



TimeSeriesSplit