

모델평가

학습한 모델의 과적합 방지


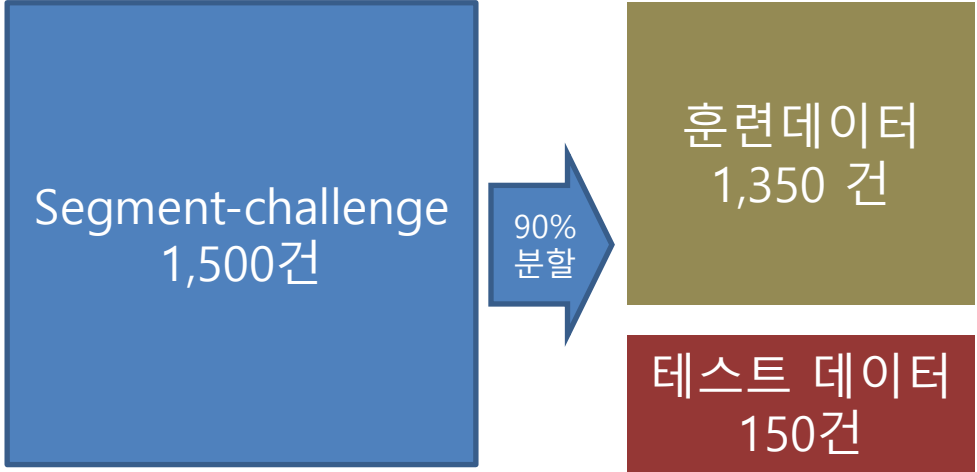
Why 모델평가 방법이 다양한가?

모델링을 위한 원본데이터 확보가 어려운 바,
그나마 확보된 원본데이터로 돌려막기하는 것

Why 어려워? 필터링 통한 축소, 원래 건수 적음 등

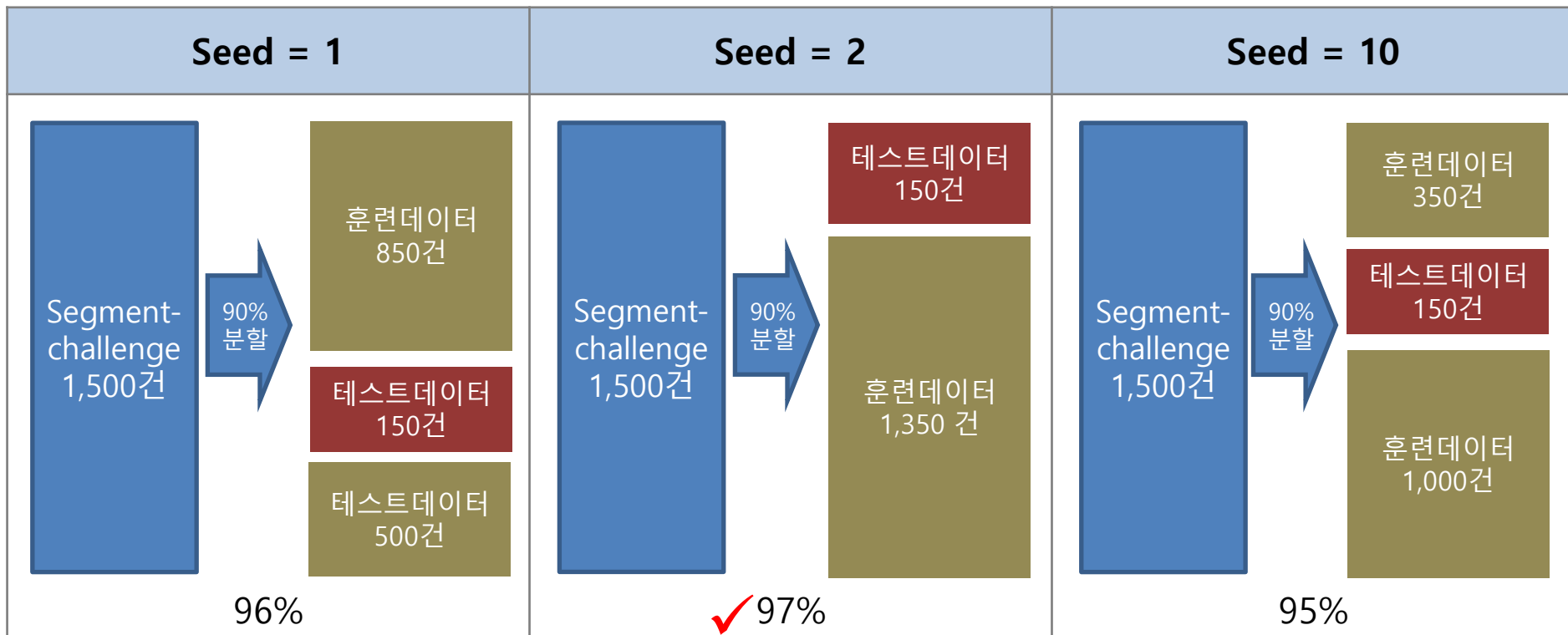
1. 분할검증(HOLDOUT)

- 원본 데이터를 **일정비율로 분할**하여 훈련/테스트 데이터 생성후 검증 (단, **1회 실시**)

훈련/테스트 데이터를 별도 생성	원본 데이터를 일정비율로 훈련/테스트 데이터 분할
 <p>Segment-challenge 1,500건</p> <p>Segment-test 800건</p> <p>✓ 96%</p>	 <p>Segment-challenge 1,500건</p> <p>90% 분할</p> <p>훈련데이터 1,350 건</p> <p>테스트 데이터 150건</p> <p>95%</p>

2. 무작위 검증 (Random Seed)

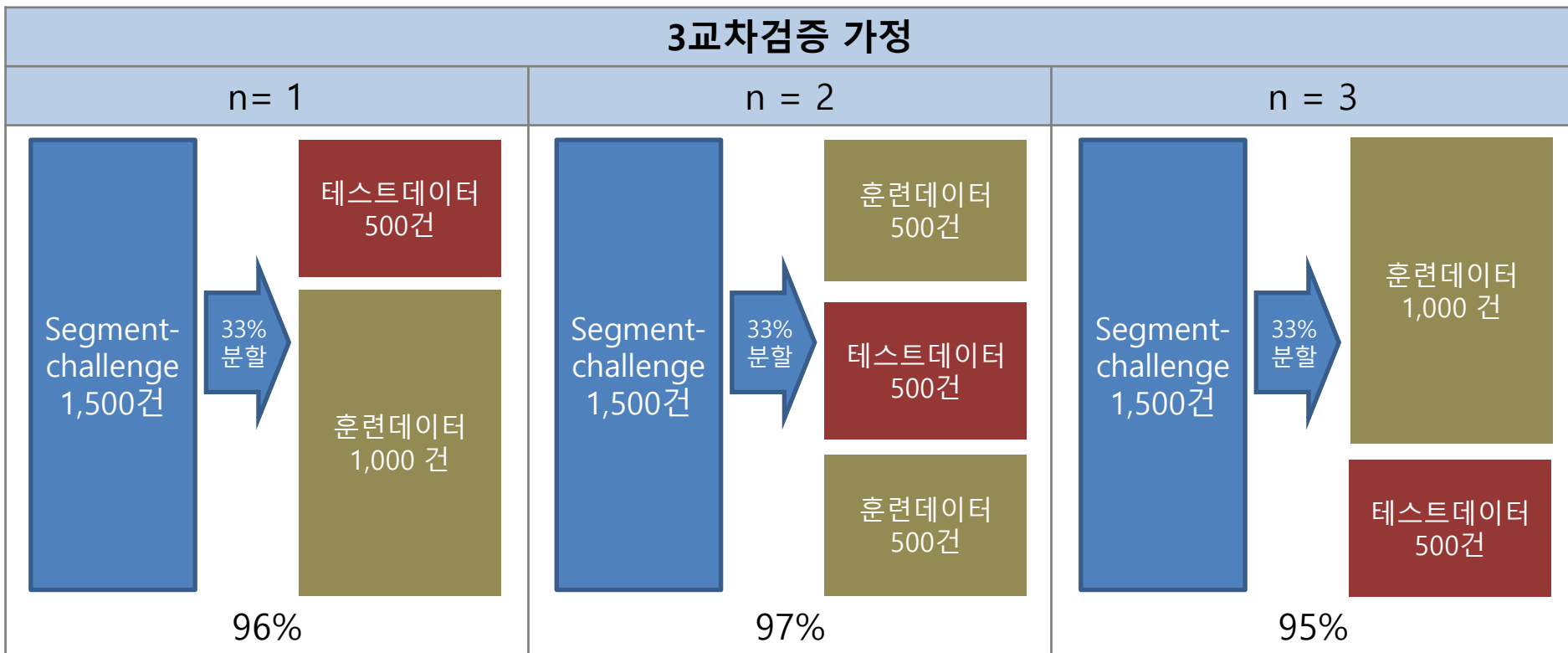
- 분할검증과 같이 원본데이터를 **무작위로** 일정비율로 분할후 검증하여 변동성 추정



3회 평균 : 96%, 분산 1%

3. 교차검증 (CrossValidation)

- 원본 데이터를 **n개 등분 분할**하여
n회 반복적으로 성과 측정후 **평균 산출**



3회 평균 : 96%

통상 10교차검증 사용

4. 기준과 상대적 비교

- 기준분류기 ZeroR 과 다른 분류기의
정분류율 비교하여 높으면 채택

ZeroR 정분류율 = 33%

J48 정분류율 = 98%

검증 방법은 동일하기만 하면 됨
└ 분할검증, 무작위 검증, 교차검증

5. 훈련데이터 = 테스트데이터

- 원본데이터를 분할/추출 없이 있는 그대로 훈련데이터와 테스트데이터로 간주

