머신러닝 모델링의 흔한 실수들

60+ 사례 중 데이터 관련 사례

2021.02.24.

한국에너지기술연구원 플랫폼연구실

이제현

머신러닝 workflow





데이터 생성/수집

데이터 저장/관리

전처리 preprocessing

- 통계분석
- 시각화
- Feature Engineering
- Feature Extraction

기계 학습 Machine learning

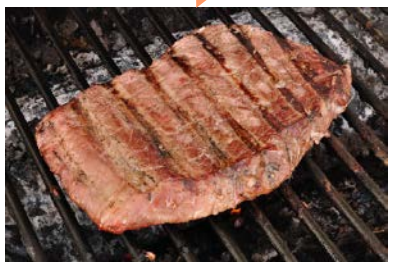
- 모델 선정
- Hyperparameter tuning
- 데이터 학습

오차분석 Error analysis

- 검증데이터
- 테스트데이터
- 잔차분석

모델 배포

머신러닝 workflow



"음식이 이 따위인데 사망자는 안 나왔나요?"

"기름이 하도 많아서 미국이 침공하겠다!"

"이 고기는 너무 안 익어서
실력 있는 수의사가 살릴 수도 있겠다!"

데이터 생성/수집

데이터 저장/관리

전처리 preprocessing

- 통계분석
- 시각화
- Feature Engineering
- Feature Extraction

기계 학습 Machine learning

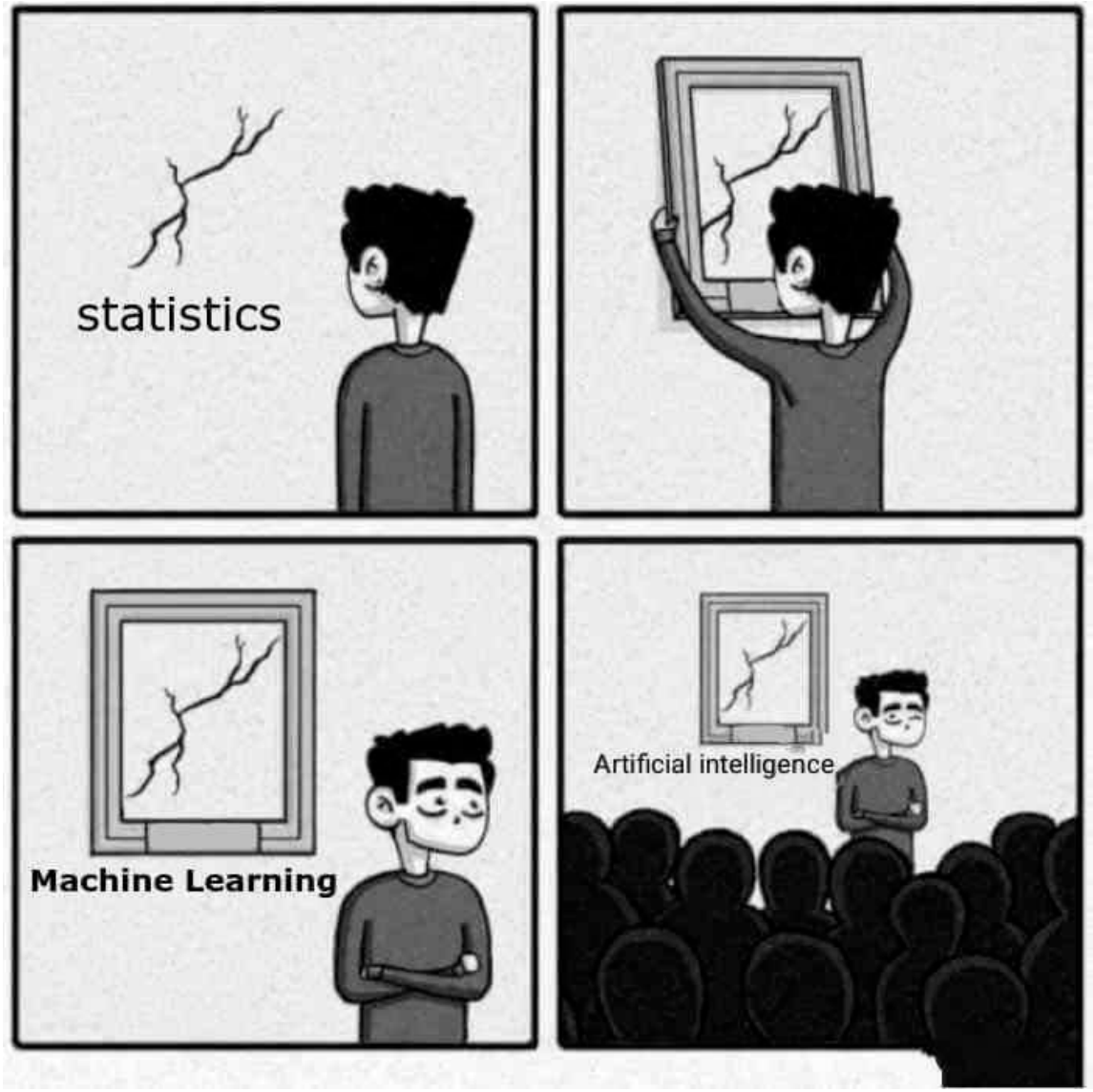
- 모델 선정
- Hyperparameter tuning
- 데이터 학습

오차분석 Error analysis

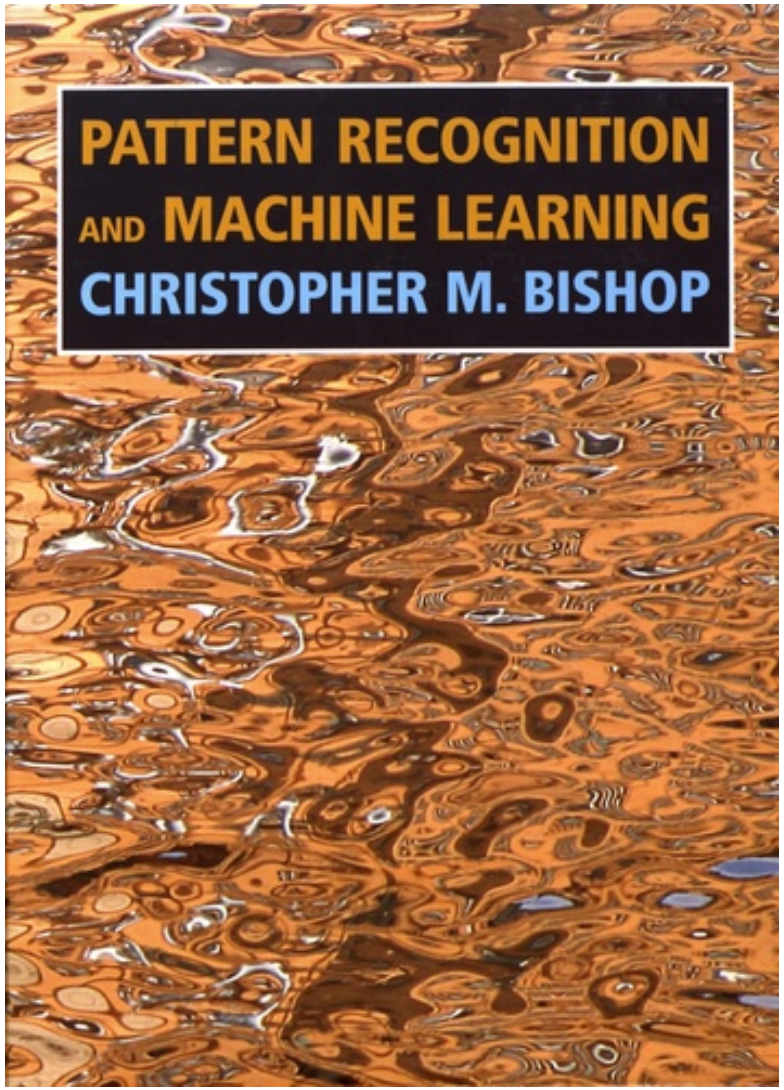
- 검증데이터
- 테스트데이터
- 잔차분석

모델 배포

인공지능, 머신러닝, 그리고 통계학



머신러닝의 목적



- **Pattern Recognition** : by C. Bishop
  automatic discovery of regularities in data through the use of computer algorithms

- **Machine Learning** : www.sas.com
  a method of data analysis that automates analytical model building.

- **주어진 데이터의 패턴을 파악하여 새로운 데이터에 적용**
  - **지도학습:** X-Y 인자간 관계를 파악, 새로운 X로 Y 예측
  - **비지도학습:** X 인자를 분석, 새로운 X를 군집분류 또는 데이터 변환
  - **강화학습:** 다양한 시도의 결과로부터 더 나은 시도를 도출

오류 1 모델 선정

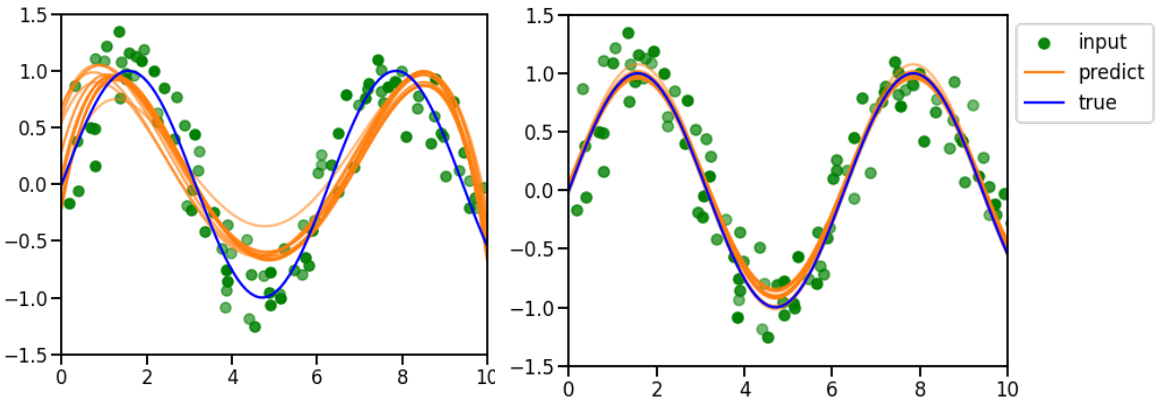
데이터를 적절하게 변환한 경우 ≈ 옳은 모델을 선택한 경우

$X = sin(x)$ Feature Engineering

ML model : $y = a\ sin(x)+b$

$a = 0.9536, b = 0.0398$

True data : $a = 1, b = 0$



모델 선정 방법 : 데이터에 대한 사전 지식

분야별 전문 지식 Domain Knowledge

① 물리적 의미
- 지배 방정식

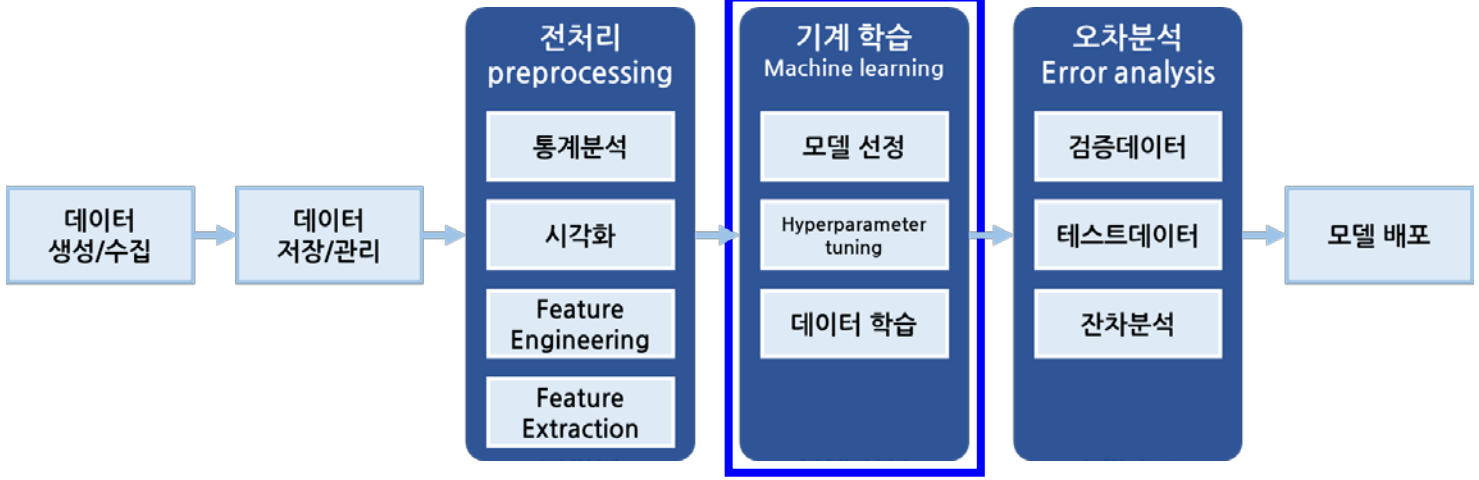
② 수치적 특성
- 주기성, 상/하한선 등

③ 데이터 특성
- 분포, 인자간 상관성 등

모델이 틀린 경우

ML model : $y = a_0x^0 + a_1x^1 + a_2x^2 + a_3x^3 + a_4x^4$

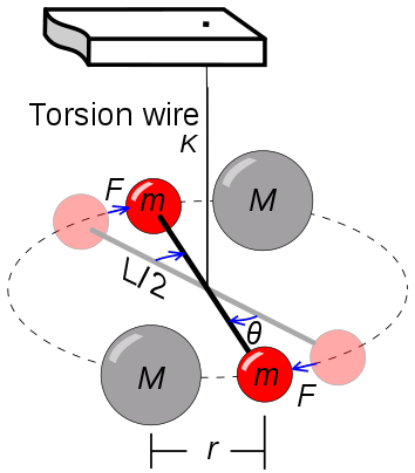
True data : $y = sin(x)$



오류 1 모델 선정

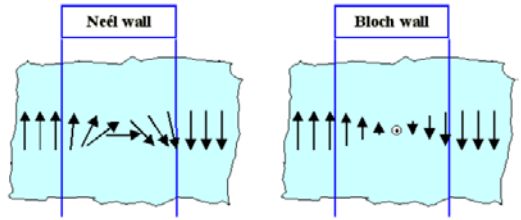
- **Feature Engineering + 모델 선정**
  - 다짜고짜 2차항, 3차항 만들고 log를 취하고 - 잘 맞아도 해석이 어려움
  - 지배방정식의 항term을 만든다는 생각으로 접근

ex. 만유인력, 자기에너지

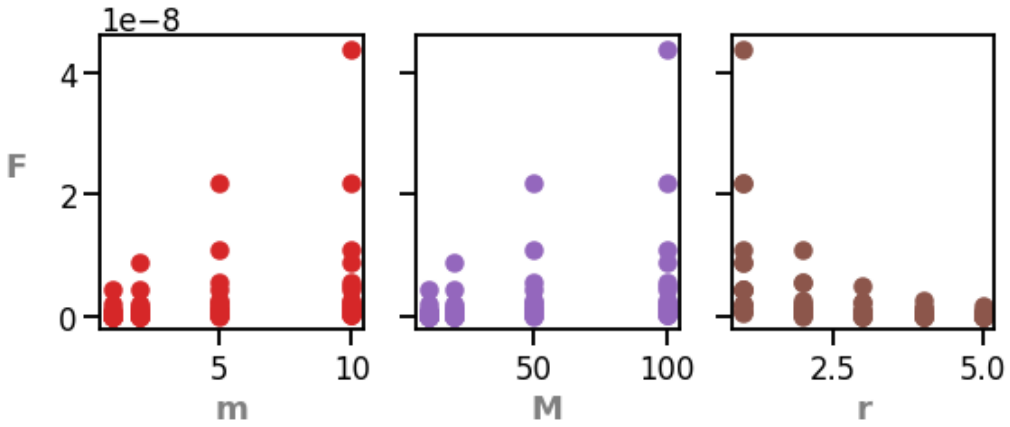


$$F = \frac{GmM}{r^2}$$

	m	M	r	F	mM/r2
0	1	10	1	6.929196e-10	10.000000
1	1	10	2	1.923696e-10	2.500000
2	1	10	3	9.967518e-11	1.111111
3	1	10	4	6.723212e-11	0.625000
4	1	10	5	5.221562e-11	0.400000
...	...	...	...	...	...
75	10	100	1	6.676552e-08	1000.000000
76	10	100	2	1.671052e-08	250.000000
77	10	100	3	7.441075e-09	111.111111
78	10	100	4	4.196770e-09	62.500000
79	10	100	5	2.695120e-09	40.000000



$$\gamma_{Ne'el} = 4\sqrt{A\frac{\mu_0 M_S^2}{2}}, \quad \gamma_{Bloch} = 4\sqrt{AK_1}.$$



오류 2 데이터 전처리

이상치 감지 방법 : 데이터에 대한 사전 지식 + 통계 분석

① **논리적 의미** - "네가 왜 거기서 나와?"

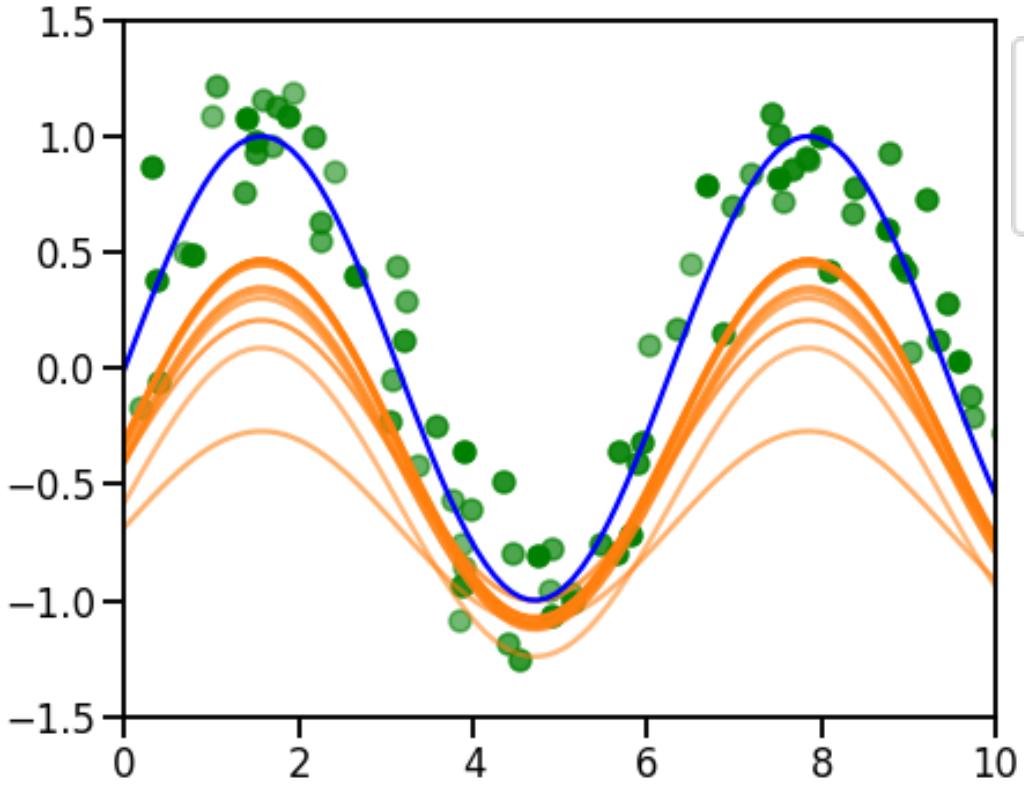
② **통계적 고립** - "남들 다 여기 있는데 쟨 왜 저기 있지?"

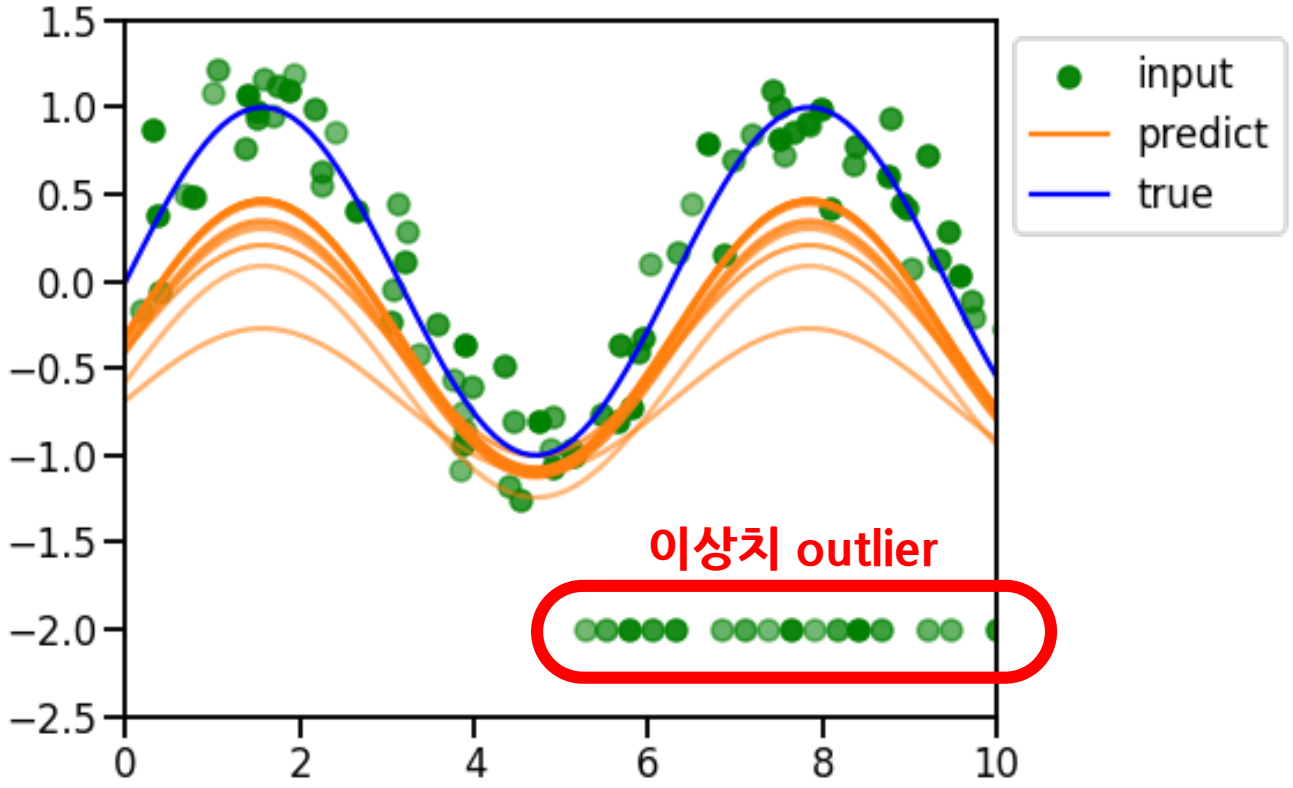
일반적 순서 : 통계적으로 고립된 데이터를 추린 후,
논리적으로 합당한지 체크. 뺄지 말지 결정.

ex. "대학 중퇴자 재산이 왜 이렇게 많지? "

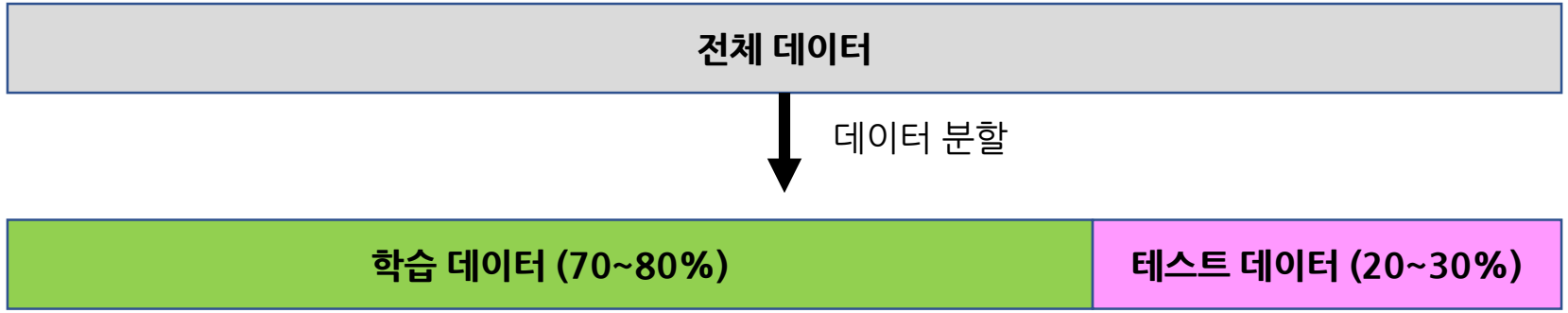
ML model : $y = a\ sin(x)+b$

$a = 0.7702,\ b = -0.3022$

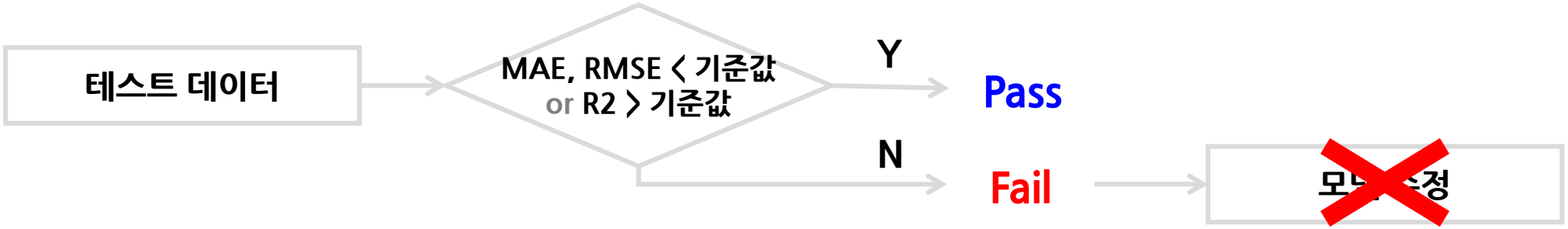




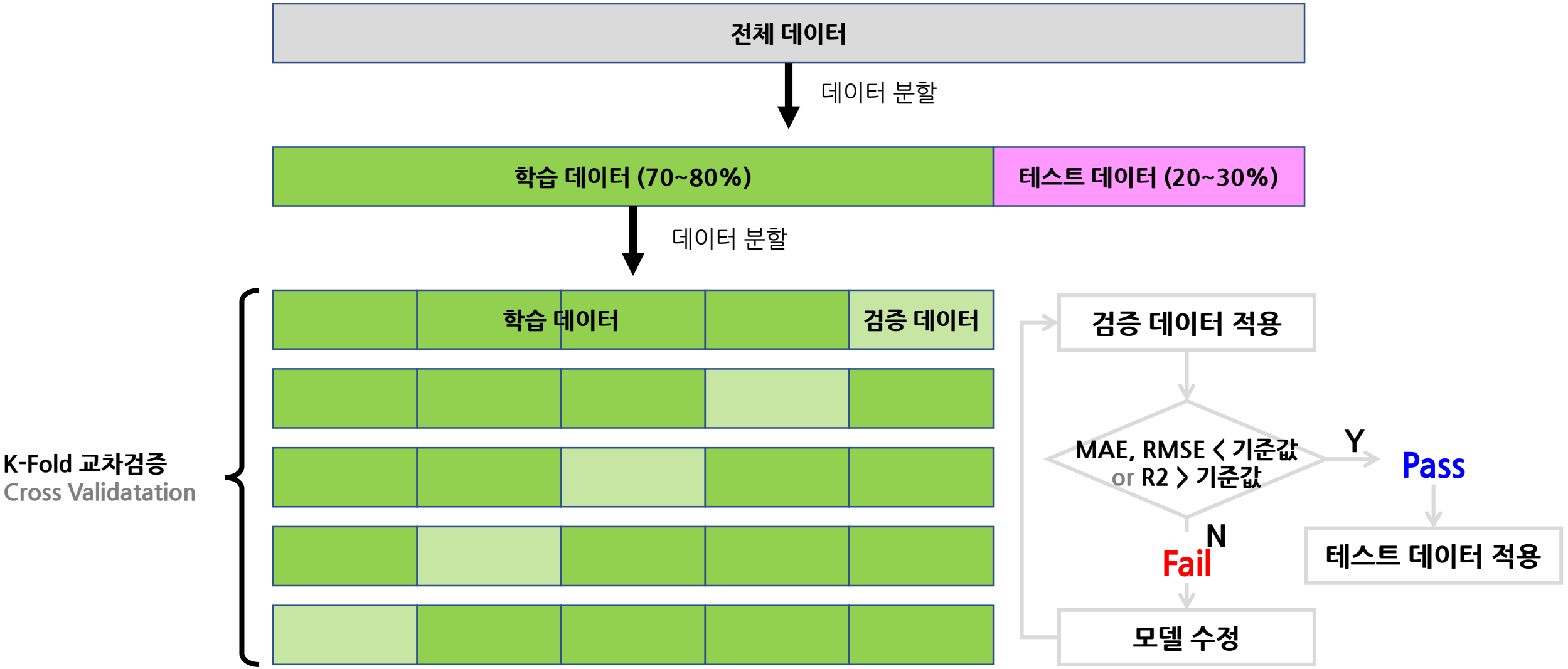
머신러닝 학습 데이터 분할



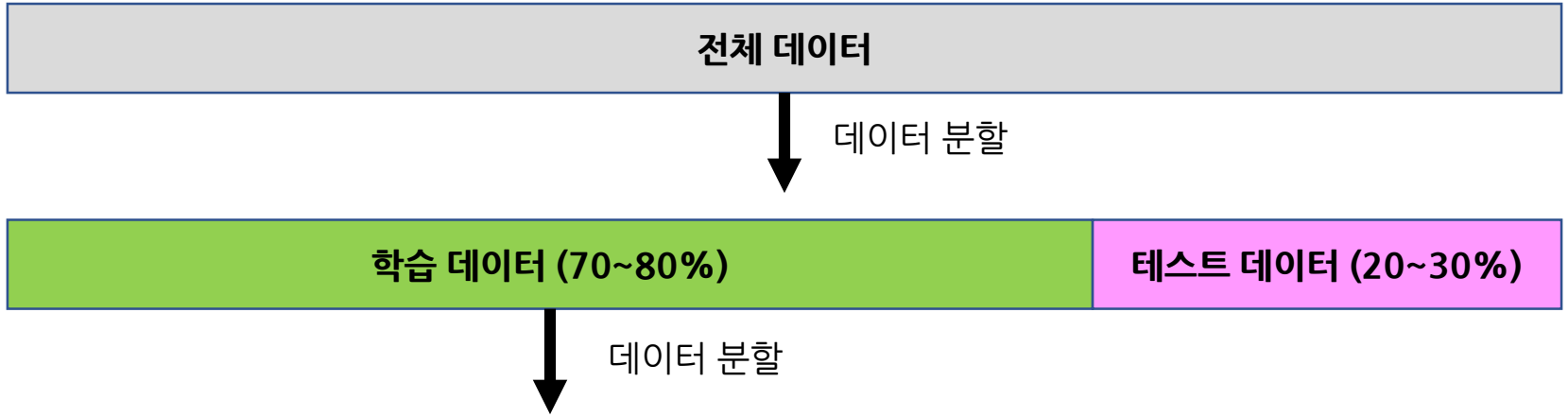
- 학습 데이터를 이용해 머신러닝 모델을 훈련
- 테스트 데이터를 사용해 학습 정도를 평가



머신러닝 학습 데이터 분할

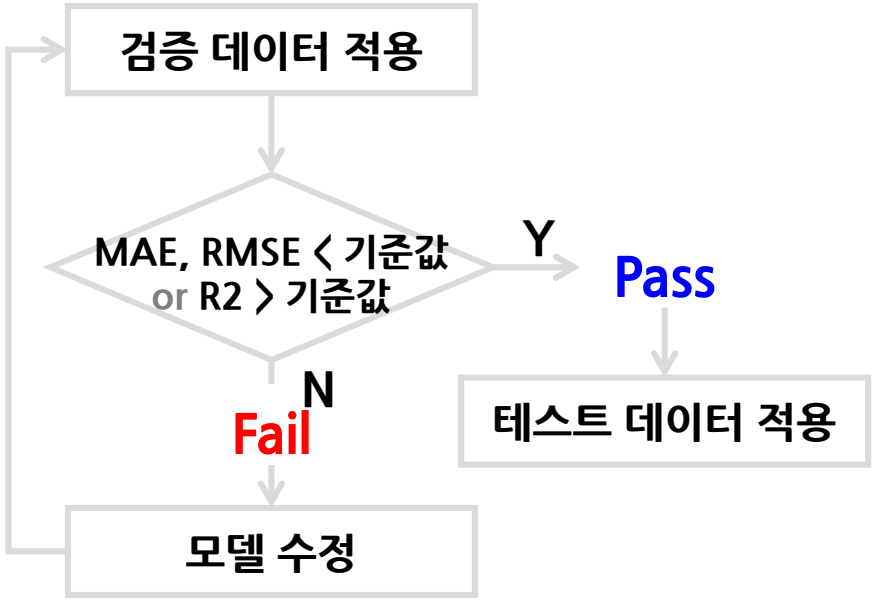


머신러닝 학습 데이터 분할 : 시계열

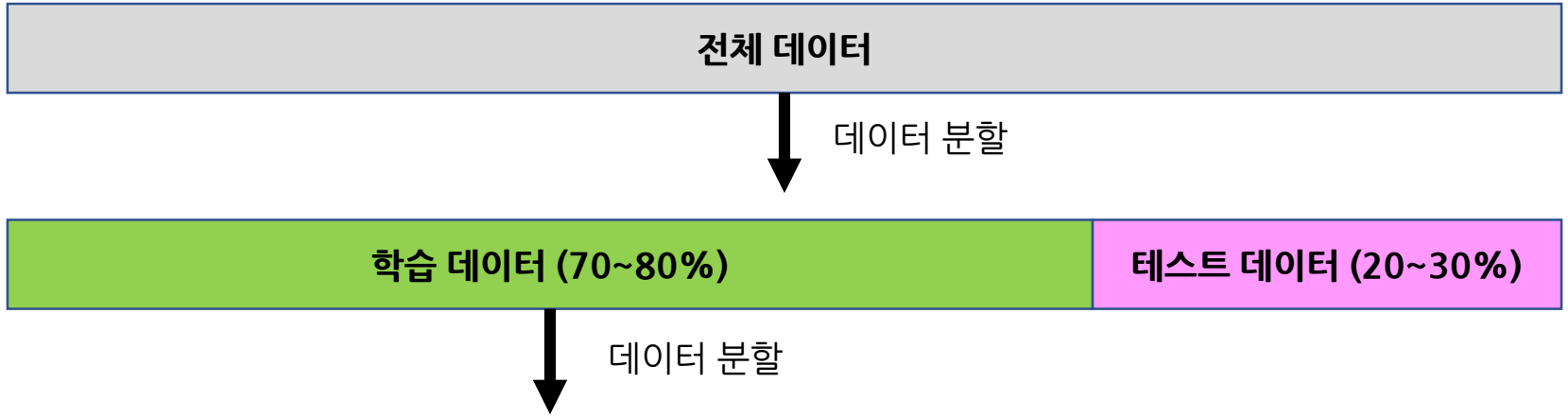


Nested Cross Validation

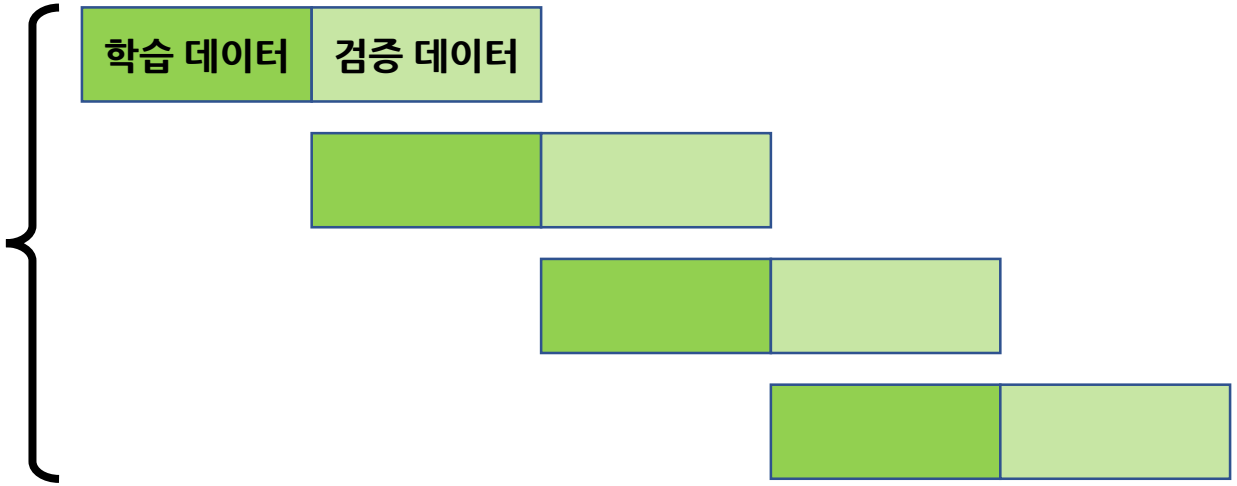
시계열 교차검증
Time Series
Cross Validatation



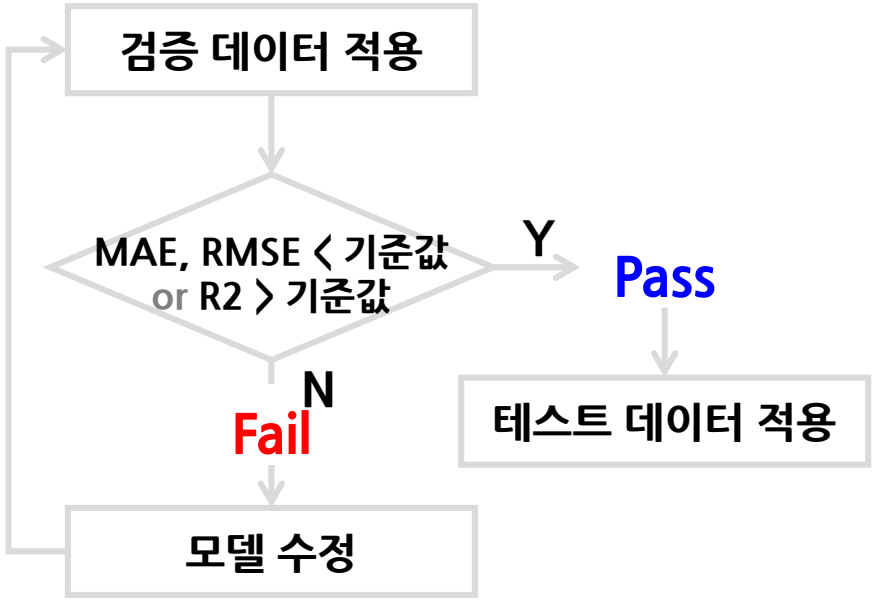
머신러닝 학습 데이터 분할 : 시계열



Blocking Time Series Split



시계열 교차검증
Time Series
Cross Validatation

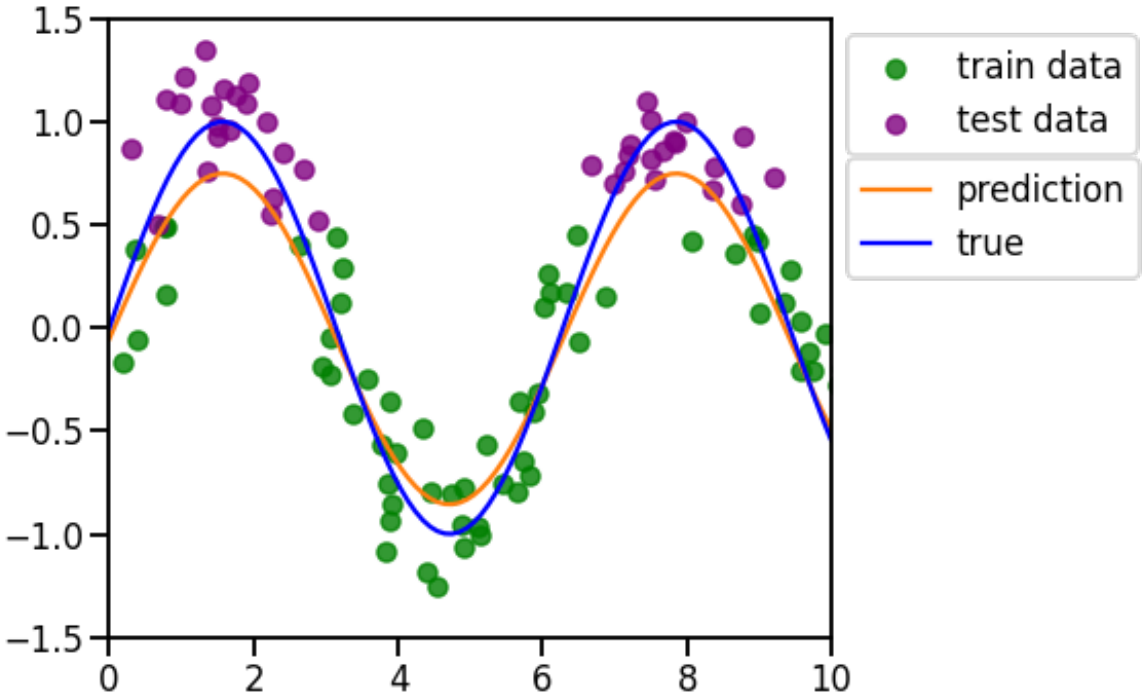


오류 3 학습, 테스트 데이터 분할

- 테스트 범위를 학습하지 못함 : extrapolation
- 사람으로 치면 : 안 배운 범위에서 시험보기

ML model : $y = a\ sin(x)+b$

$a = 0.8023,\ b = -0.0536$



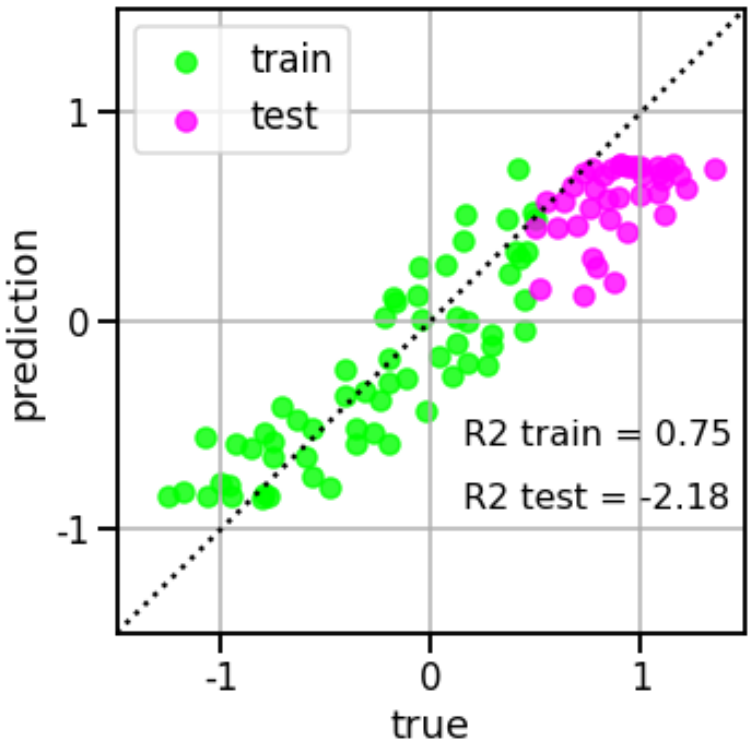
- R2 < 0 의 의미?

"평균값으로 찍은 것보다 예측을 못한다."

$$R^2 = 1 - \frac{RSS}{TSS}$$

RSS = sum of squares of residuals

TSS = total sum of squares



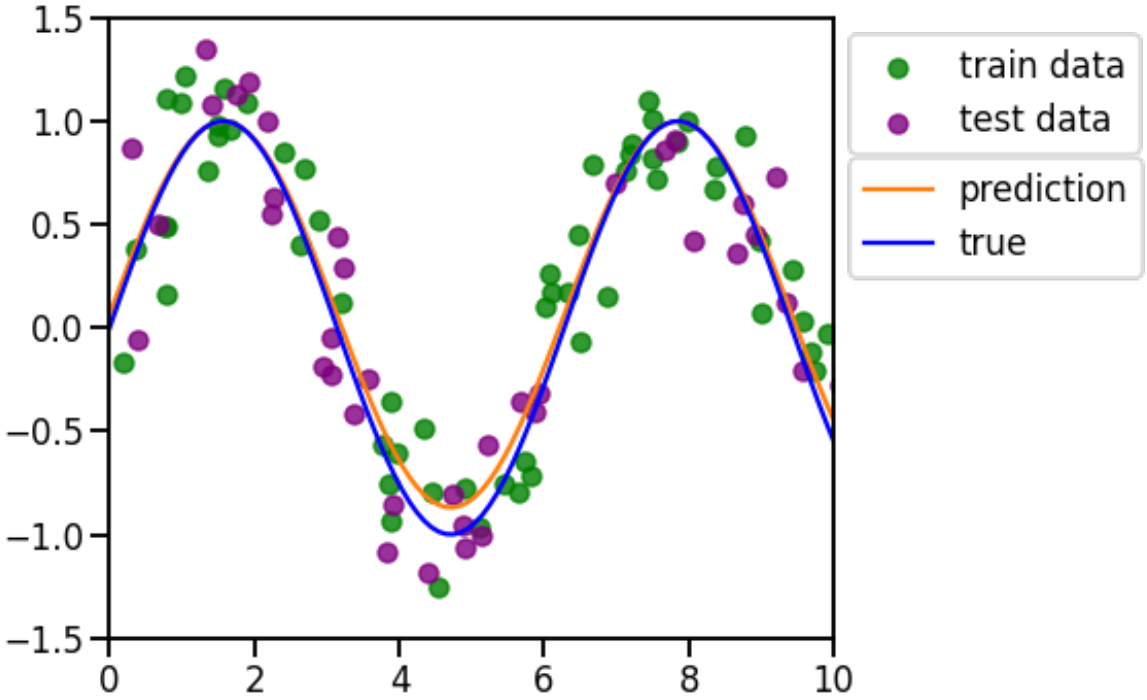
오류 3 학습, 테스트 데이터 분할

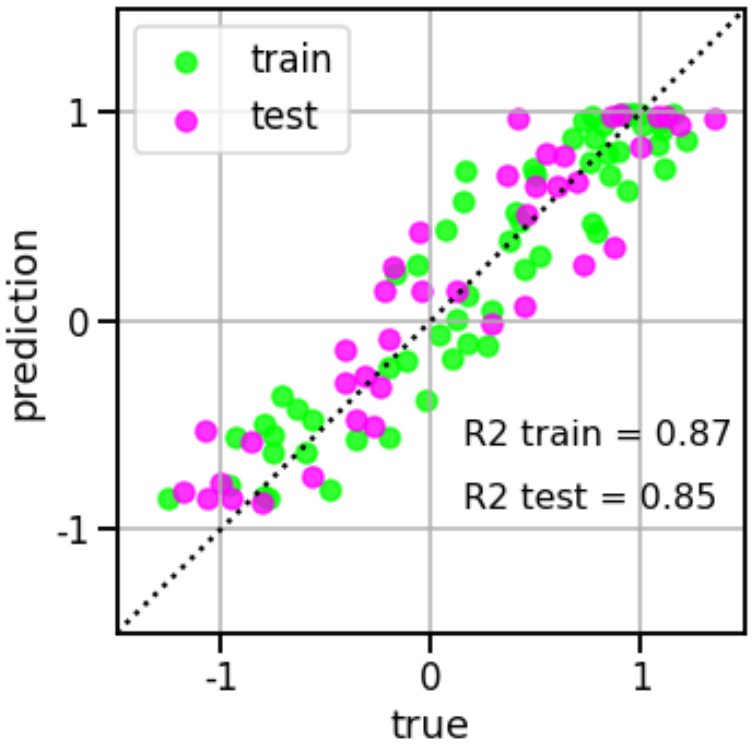
- **테스트 범위를 학습해야 함 = 데이터를 골고루 나누어야 함 : interpolation**

- **사람으로 치면 : 배운 범위에서 시험보기**

ML model : $y = a\ sin(x)+b$

$a = 0.9344,\ b = 0.0649$



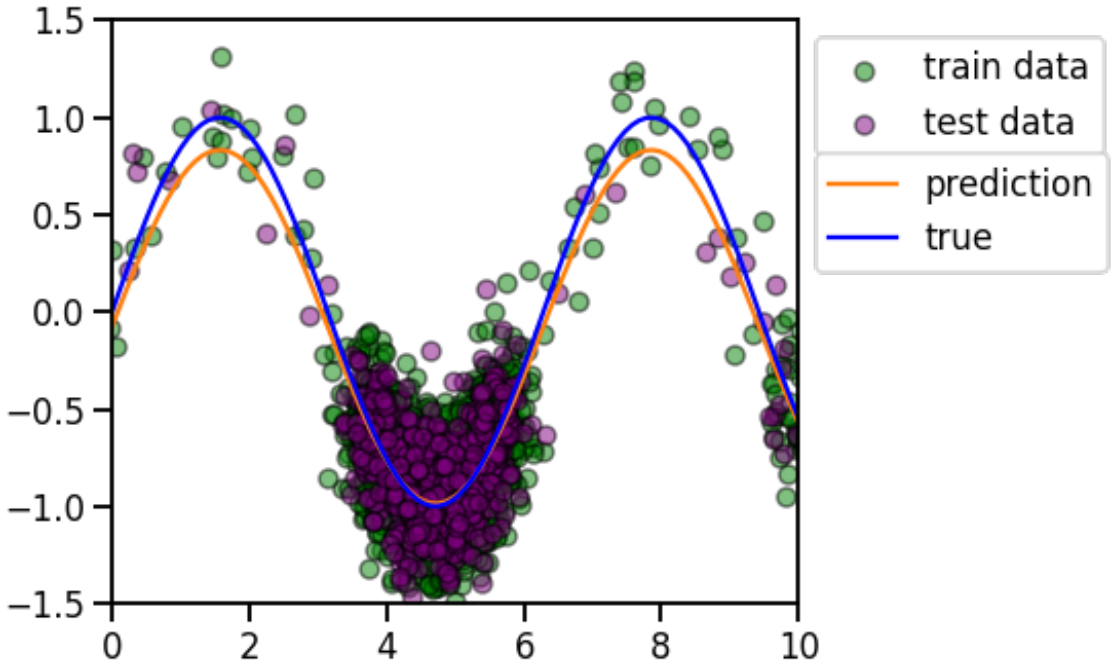


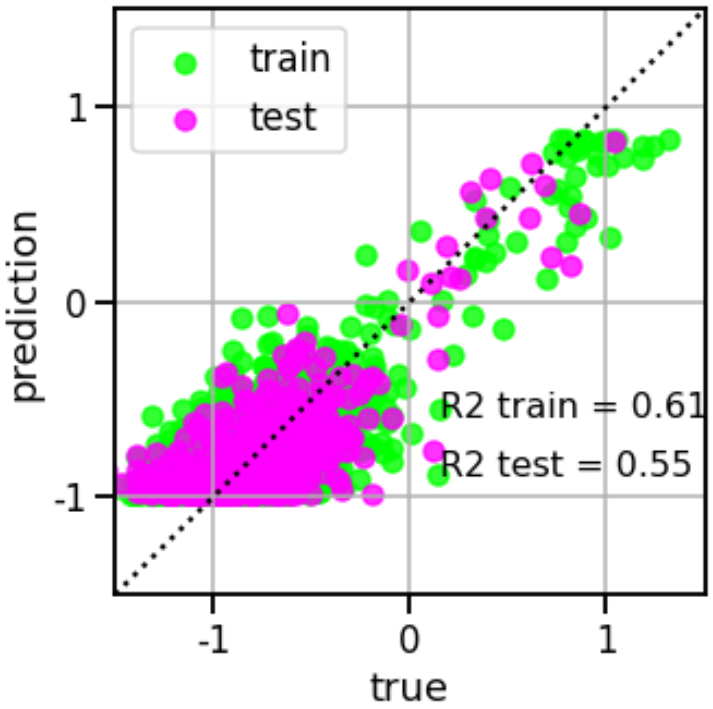
오류 4 데이터 불균형

- Y < -0.5에 데이터의 99%가 존재하는 상황.
- 학습이 다수 데이터에 치중. 소수 데이터 예측력 저하

ML model : $y = a\ sin(x)+b$

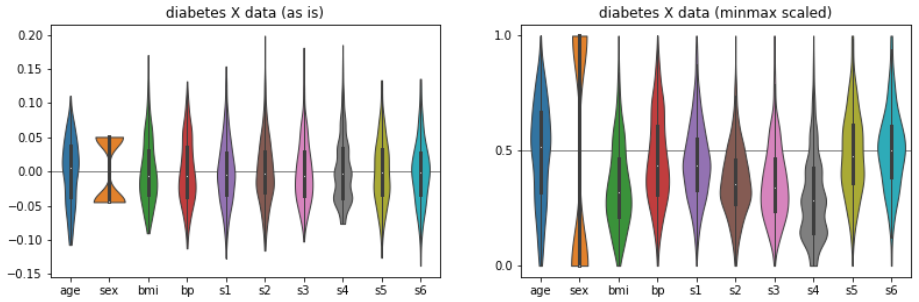
$a = 0.9096,\ b = -0.0766$

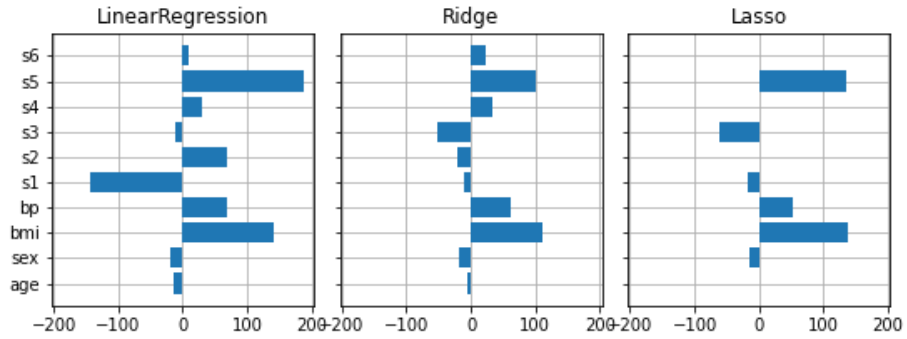


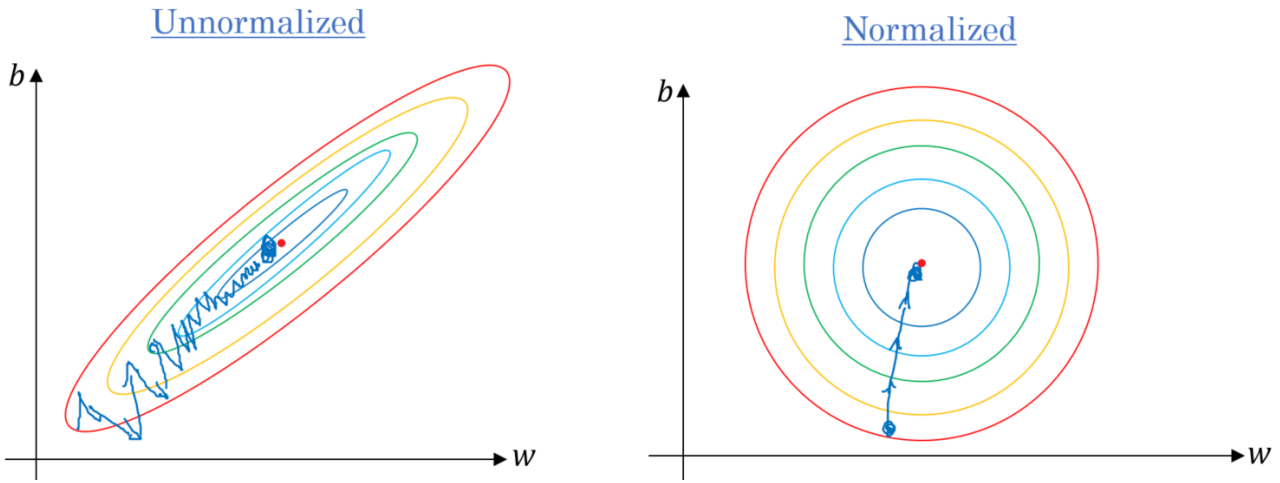


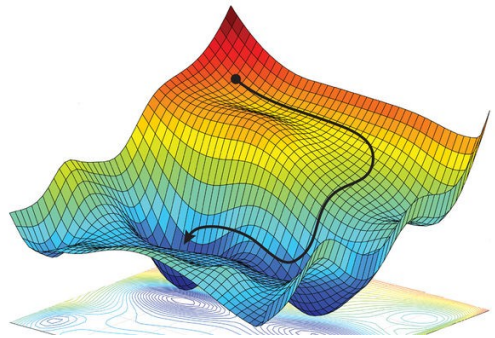
오류 5 스케일링 Data Scaling

- (데이터 분석) 단위 불일치에 따른 데이터 대소 오류
  - ex) 1m > 50 cm
- (선형회귀) 인자 중요도 파악
  - ex) $Y = a_0X_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + \cdots$
- (신경망) 비용함수cost function 왜곡으로 인한 수렴 곤란









오류 6 데이터 유출 Data Leakage

- **실제 상황에서 발생할 수 없는, 미래를 엿보는 일이 발생. 당장은 기분이 좋지만 실전 성능 저하.**
  - "어? 저는 잘 나오던데요?"

1. 타겟 유출 Target Leakage

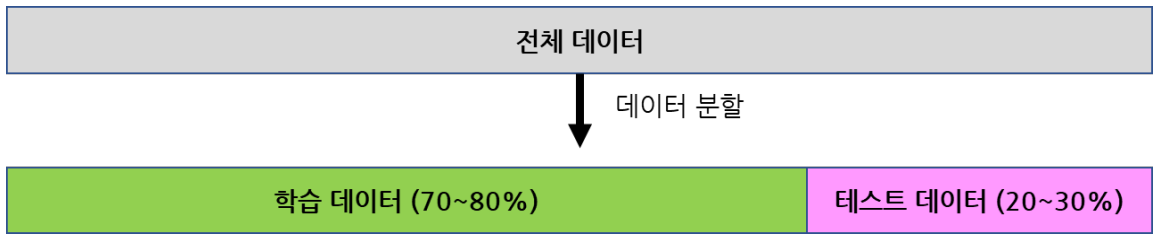
- 예측 시점에서 사용할 수 없는 데이터를 학습에 사용
- ex) 감염 후에 먹는 항생제 복용 여부를
  폐렴 감염 예측 인자로 사용

got_pneumonia	age	weight	male	took_antibiotic_medicine	...
False	65	100	False	False	...
False	72	130	True	False	...
True	58	100	False	True	...

2. 학습-테스트 오염 Train-Test Contamination

- 테스트용으로 완전히 분리해야 하는 데이터가 학습데이터에 반영됨
- ex) 데이터 표준화 : 평균 m, 표준편차 σ 계산시 테스트 데이터 포함

$$Z = \frac{x - m}{\sigma}$$



그 외 많은 실수들 : https://bit.ly/3j3vnYL

#	실수 사례
Pega	Scaling을 하지 않고 선형회귀나 PCA, 딥러닝하기
Pega	Tree 모델에 기를 쓰고 Scaling하기
Pega	validation set 대신 test set으로 hyperparameter tuning하고 예측값 만들기
Pega	데이터 분포를 확인하지 않고 학습하기 → 극심한 불균형에 똑같은 값만 예측.
Pega	test set을 분리하지 않고 scaling이나 encoding하기
Pega	분포를 조정하려고 y값에 log를 씌운 채로 prediction 결과를 report 하기
Pega	분포를 조정하려고 y값에 log를 씌운 채로 예측, 여기서 나온 오차범위를 오차범위로 report
고속터미널	NLP에서 텍스트 데이터 내부에 정답 토큰이 있는 상태에서 정답 토큰을 예측해서 Accuracy 95%
이동진	PCA로 차원축소를 할 때, 1. 전체 데이터로 fit하고 전체 데이터 transform 하기. 2. 훈련 데이터 / 테스트 데이터 각각 fit하고 transform 하기. 3. 훈련 데이터로 fit하고 테스트 데이터에 transform하기
Danny To Eun Kim	Scaling 할때 train set 과 test set에 다른 스케일을 적용한 경험이 있습니당
이동진	시계열 데이터 훈련/검증/테스트 데이터 나눌 때, 랜덤하게 나누기
우현우(Hub1)	Unbalanced data에서, Tran: Test(또는 validation)으로 나눈 뒤, oversampling(또는 undersampling)을 Train에만 해야하는데.. Test(or validation) set에도 하는 경우
Pega	데이터 시각화할 때 엉뚱한 범위 잡아놓고 "왜 그림이 안그려지지?"
안중호	여자가 92%인 데이터에서 성별을 예측하려고 애쓰기
안중호	머신러닝 할 필요 없이 룰베이스드로 풀리는 문제에 굳이 머신러닝을 적용해서 풀려고 노력하기
안중호	성능을 올리기 위해 하이퍼파라미터를 튜닝하는것이 아니라 랜덤시드를 튜닝하고 있기
우현우(Hub1)	(CV; Cross Validation말고 70:30 나눠서 하는 경우) trainset으로 만든 model로 testset에 적용해서 나온 성능 1개 가지고 "이것의 성능은 XX입니다~"라고 확실한 것처럼 말하기 (성능 말고, coefficient도 마찬가지)
Hassan	ROC-AUC Score 올리려고 결과 반올림 하기
Hassan	고객사 DB 복사해온 회사 개발 서버에 시스템 설치하기
Hassan	하이퍼 파라미터 튜닝까지 다 해놓고 모델 저장 안하고, 예측값만 저장하기
Hassan	GBM모델만 써놓고 앙상블 ^^
Hassan	예측한 변수를 다시 X에 추가하기
김현우	k means 시드 한번만으로 학습하기?? 보통 여러번해서 나온 결과를 평균내는데 처음 배울때 저런 실수를 했습니다
전영태	입력데이터에 test 타겟데이터를 넣고 학습하기
튜뷰튜뷰	로지스틱회귀분석에서 최종 모델 만들고 예측할때 결과값을 exp(y)그대로 써서 '1'일 확률이 96%라고 잘못 해석했을때 (사실은 49.6%...ㅎㅎ)
고현웅	pytorch 이용시 성능이 너무 나쁘게 나와서 하루 종일 삽질했는데 알고보니 eval()을 안찍어서였던 경험이 있네요 ㅋㅋㅋ
박군	분석에서 컬럼명을 활용안하고 컬럼인덱스를 썼다가 DB가 조금씩 바뀔때마다 코드 전체가 꼬여서 사업진행이 밀림
익명의 기린	디버깅용 100% 피팅을 시도하는데 Batch Norm을 켜놔서 계속 Regularize 당함.
game	이탈 예측 모형을 만들 때, 적용 시점을 고려하지 않은 train 데이터 구성 (신규 유저 이탈 예측이 목표인데 train에는 10개월치 데이터가 들어간 사례)
김성현	순서대로 레이블링 된 데이터를 셔플 안해서 몇 번이나 삽질
Pide	딥러닝)Inference로 결과 뽑는데, 결과가 이상해서(예상한 점수가 안나와서) 봤더니 transform을 valid가 아닌 다른 녀석을 썼던 경험이 있네요
황준원	데이터가 제대로 전처리되는지 확인하지 않고 넣어서 잘못된 성능을 뽑기
지홍	x-y 시차를 고려하지 못함(ex. 2개월차 시점의 고객상태를 예측하는데 3개월차 시점의 데이터를 변수로 사용)
지니	CV(cross validation) 을 딥러닝의 epoch 개념과 헷갈려서 모델이 5번(5-fold) 업데이트 되겠지? 헷갈렸습니다 ㅎ
김지은	테스트 데이터 검증할때 즉, 레이블 렌덤화를 테스트 데이터에만 적용해야되는데 학습할 데이터도 레이블 랜덤화후 학습시킴.
Pega	DOE가 너무 잘 된 데이터에 Boosting model 적용했다가 예측모델의 isosurface가 계단모양으로 출력됨

#	실수 사례
Linejin	matplotlib으로 line형 그래프를 그릴 때, 깔끔한 선형으로 나와야 하는데 번개마냥 들쭉날쭉 나왔었습니다. 찾아보니 그 이유로 x축에 해당할 값이 일정한 step이 아니였었던게 문제였습니다.
이지운(metr0jw)	데이터를 scailing하는 과정에서 각 column에 대한 평균, 편차를 이용했어야 하는데 실수로 전체의 평균, 편차를 가지고 scailing을 해 결과값이 이상하게 나왔던 경험이 있네요
최원우	python에서의 데이터 복사는 default가 call by reference! (필요에 따라 call by value로 deepcopy하는걸 잊지말자!)
정씨	Pre-Training 할 때 쓴 데이터의 일부를 Fine-Tuning 때 다시썼습니다.
전씨	전체데이터 augmentation 하고 그 중에서 train/validation/testset으로 나눔.. test 정확도 95% ㅋ.. (실제는 70%도 힘든케이스)
김종수	Time Series dataset를 rolling window로 구성했는데, 이를 train/valid/test set으로 나눌 때 단순하게 random split을 했다가 test set과 train/valid set이 섞여서 overfitting된 문제가 있었습니다.
이씨	String형식의 label을 integer로 바꿨다가, 최종 제출할때 그대로 integer로 내서 0점 받기(혹은 integer 순서를 대회에서 제공하는 것과 맞지 않게 바꾸거나)
이제영	스타일 트랜스퍼를 도전 중, 논문의 노말라이제이션 방법을 시도 시, 나눌 때 0인 경우가 있었는데, api에서 그런 경우 자동으로 미지의 값으로 학습 시켜서 결과가 안나와서 삽질 했었습니다..
이제영	라벨링 하는게 귀찮아서, 반쯤 가려진 오브젝트들은 걸러버리니 데이터의 절반이 라벨이 없어서 혼났던 기억이 있습니다..
Pega	GridSearchCV 결과 최적 파라미터로 내가 입력한 범위의 맨 끝값이 나왔는데, 범위를 연장하고 다시 GridSearchCV하지 않고 그 값을 그대로 사용함. 나중에 알고보니 최적값은 저 너머에 있었음.
Pega	line plot 그릴때 데이터 셔플된채 그림. 그림이 랜덤모션!
김종헌	쉽게 분류가 가능한 데이터를 다수 넣어서 정확도가 높다고 좋아했는데 애매한 데이터를 중심으로 분류해보니 성능이 매우 떨어지더군요. 애매한 데이터를 분류하려면 애매한 데이터 중심으로 모델을 만들어야 원래 목적에 부합되는 성능을 확인할 수 있습니다.
김종헌	데이터가 많으면 무조건 좋을 줄 알았던 시절이 있었습니다. 하지만 타겟을 잘 설명할 수 있는 핵심 피쳐들을 잘 선택해야 모델 성능이 좋습니다.
윤성국	mini batch를 나눠놓고 계속 첫번째 mini batch만 학습시키기
김정태	with torch.no_grad()이거 안찍어서 3일동안 찾았네요ㅋㅋㅋㅋㅋ
Pega	GPU 메모리 꽉 차있는 걸 모르고 자꾸 학습 시도하기
뛰어라고양이	엄청나게 큰 csv 파일 아무 생각없이 엑셀로 열기
뛰어라고양이	파일 이름 저장 규칙 생각없이 지정 안했다가 다 지우고 다시 저장하는 불상사를 반복하기. (0001.jpg로 하려했으나 잊어버리고 1.jpg, 10.jpg 순으로 정렬되는 불상사)
이명규	인재에 가까운 실수인데, 객체인식 태스크의 confidence thresholding 과정을 통일된 가이드라인 없이 진행하는 바람에 라벨러와 개발자 간 쓸데없는 기싸움을 했던 기억이 납니다.
김지현	이미지의 경우 bgr, rgb와 float, int 데이터 형에 관한 문제가 빈번했습니다.
송민수	이미지 저장할때 float형은 0~1사이로 normalizing해줘야되는데 0~255값으로 했다가 흰색이미지만 나와서 계속뻘짓하기
임용택	loss.backward와 optimizer.step()의 순서를 반대로 써 1주일내내 이유를 밝힌다고 논문을 뒤져봤읍니다ㅠㅠ
mcp	중복 처리를 안해서 test와 train에 같은게 존재하여 성능이 높게 나옴
mcp	categorical value를 input으로 쓸 때 test에 처음 나오는 categorical value는 맞출수 없어 성능의 한계가 존재하나 더 개선하려 삽질했습니다.
mcp	dataset 사이즈가 작은 경우 최적화를 너무 열심히하다보면 운 좋게 cv 성능이 모두 좋게 나오는 확률이 존재함. 테스트 성능은 매우 떨어짐
Pega	상관관계를 인과관계로 해석하기
Pega	Pearson correlation coefficient를 "영향력"으로 해석하기
백관구	데이터에 결측값(nan, -999, -, 0 등)이 어떻게 기입되어 있는지 모르고 사용
patrick	classification할 때, 클래스별 레이블링을 1부터 시작함
김현민	간단하게 시작해야하는데 첨부터 어려운 모델로 시작하기 (ex. ANN으로 대충 상황만 파악해야하는데 첨부터 Resnet 사용)

머신러닝 오류

오늘 살펴본 부분

- 머신러닝 = 도메인+통계+코딩
- **머신러닝 오류 = 도메인 오류 + 통계 오류 + 코딩 오류**
- Keep in mind : “어떤 일을 하는 모델, 어떤 데이터에 대응하는 모델을 만들어야 할까?”