



부산대학교

PUSAN NATIONAL UNIVERSITY

Feature Engineering and Pipelines

Data Leakage

Scaling, Encoding, and Pipelines

BX3600087-140 / DS3600087-140 / AB3600087-140

기계학습개론

2026학년도 2학기

오용경

데이터사이언스전공

yongkyung.oh@pusan.ac.kr

Section 1 • Feature Engineering and Pipelines

Problems in Raw Data

원자료에서 먼저 확인할 네 가지

모델에 넣는 두 종류의 열

DEFINITION

수치형 변수 (numerical variable)

값이 수이고 크기와 차이에 의미가 있는 열

DEFINITION

범주형 변수 (categorical variable)

값이 어느 범주에 속하는지를 나타내는 열

원자료에서 먼저 확인할 것

상태	문제	처리 방향
결측값(missing value)	계산할 값이 없음	지우거나 값을 채움
이상치(outlier)	다른 관측과 크게 떨어짐	원인과 기준을 확인
특성마다 다른 값 범위	큰 값이 계산을 좌우	척도 조정(scaling)으로 범위를 맞춤
범주 값	수치 연산의 입력이 아님	수로 부호화(encoding)

원자료의 몇 행

mpg	horsepower	weight	acceleration	origin
18	130	3504	12	1
25	?	2046	19	1
40.9	?	1835	17.3	2
15	165	3693	11.5	1

?는 결측값이다. 전체 397행에서 weight는 1,613에서 5,140까지, acceleration은 8에서 24.8까지로 범위가 크게 다르고, origin의 1, 2, 3은 제조 지역을 나타내는 범주다

Source: James et al. (2023), data: Auto.

전처리가 하는 일

DEFINITION

전처리(preprocessing)

원자료를 모델이 계산할 수 있는 입력으로 바꾸는 처리

입력의 문제	전처리의 역할
결측값	계산 가능한 값으로 채움
값 범위	중심과 퍼짐을 맞춤
범주	수치 열로 표현

Source: UBC CPSC 330 (2026), Lecture 5.

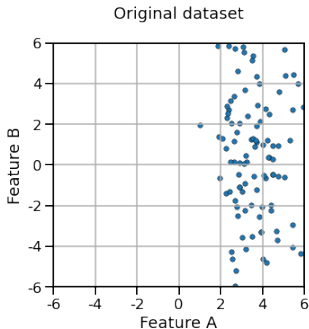
결측값을 채우는 기준

대체(imputation)는 결측값을 정한 규칙으로 채우는 처리

방법	채우는 값	주의할 점
수치형 열 채우기	평균 또는 중앙값	학습 자료에서 계산
범주형 열 채우기	가장 자주 나온 범주	학습 자료에서 계산
행 삭제	값을 채우지 않음	관측 수가 줄고, 예측할 행에는 쓸 수 없음

Source: UBC CPSC 330 (2026), Lecture 5.

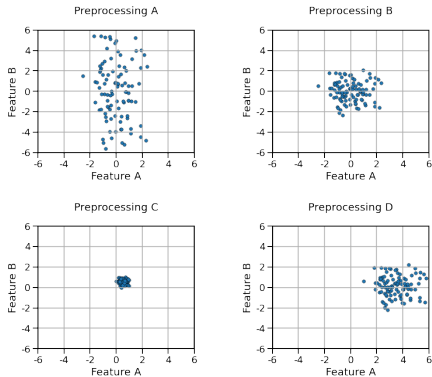
두 특성의 범위가 다를 때



두 특성 중 A는 양수 구간에 몰려 있고, B는 더 넓게 퍼져 있다

Source: scikit-learn MOOC (2022), Numerical Pipeline Quiz.

어느 결과가 표준화인가



Source: scikit-learn MOOC (2022), Numerical Pipeline Quiz.

평균과 표준편차를 맞춘다

표준화(standardization)는 학습 자료에서 각 특성의 평균이 0, 표준편차가 1이 되게 맞춘다

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \quad \mu_j = \frac{1}{n_{\text{train}}} \sum_{i \in \text{train}} x_{ij}, \quad \sigma_j = \sqrt{\frac{1}{n_{\text{train}}} \sum_{i \in \text{train}} (x_{ij} - \mu_j)^2}$$

x_{ij} 는 i 번째 관측의 j 번째 특성으로, $x_j^{(i)}$ 로도 쓴다

앞 그림에서는 중심과 퍼짐이 함께 맞은 B가 표준화 결과

척도 조정이 영향을 주는 계산

모델이 쓰는 계산	값 범위가 다를 때
관측 사이의 거리	값이 큰 특성이 거리를 지배
그래디언트를 이용한 갱신	축마다 적절한 이동 폭이 크게 달라짐
값의 순서만 쓰는 계산	순서가 그대로라 영향이 없다

Source: scikit-learn MOOC (2022), Predictive Modeling Pipeline.

범주를 수로 표현하는 두 방법

부호화(encoding)는 범주를 모델이 계산할 수 있는 값으로 바꾼다

방법	표현	사용 조건
순서형 부호화(ordinal encoding)	낮음=0, 중간=1, 높음=2	순서 있는 범주
one-hot encoding	서울=(1,0), 부산=(0,1)	순서 없는 범주

순서가 없는 범주에 0, 1, 2를 주면 없던 순서와 간격이 생긴다

학습 자료에 없던 범주

부호화는 학습 자료에서 확인한 범주 목록을 저장한다

처리 방식	결과	주의할 점
새 범주를 오류로 처리	변환이 멈춤	입력 거부 후 원인 확인
새 범주를 모두 0으로 처리	서로 다른 새 범주가 같은 값으로 표현된다	처리 규칙을 미리 정한다
드문 범주를 하나로 묶는다	특성 수가 지나치게 늘어나는 것을 막는다	묶을 기준을 미리 정한다

Source: scikit-learn MOOC (2022), Categorical Pipeline.

Checkpoint

Problems in Raw Data

- 수치형 열과 범주형 열은 어떻게 다른가
- 표준화 뒤 각 특성의 평균과 표준편차는 얼마인가
- 순서가 없는 범주에 ordinal encoding을 쓰면 어떤 관계가 생기는가

Section 2 • Feature Engineering and Pipelines

Preprocessing Leakage

처리 규칙을 학습 자료에서만 정하는 이유

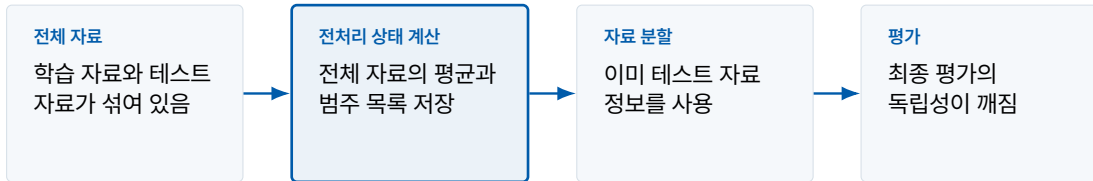
처리 규칙도 자료에서 배운다

채울 값, 중심, 퍼짐, 범주 목록은 자료를 보고 정한다

처리	저장할 상태	계산에 쓰는 관측
대체	평균, 중앙값 또는 최빈값	학습 자료의 해당 열
표준화	평균과 표준편차	학습 자료의 해당 열
부호화	범주 목록	학습 자료의 해당 열

Source: UBC CPSC 330 (2026), Lecture 5.

잘못된 순서



자료를 나누기 전에 계산한 전체 처리 상태에는 테스트 자료가 섞여 있다

DEFINITION

데이터 누수(data leakage)

예측할 때 사용할 수 없거나 평가에 남겨 둔 자료의 정보가 학습 과정에 들어가 최종 성능을 과대 추정할 수 있게 만드는 현상

평가에서

테스트 성능이 비정상적으로 높아질 수 있다

배포에서

같은 정보가 없어 성능이 재현되지 않는다

자료에서 상태를 계산하는 변환

변환	보는 범위	상태 계산
섭씨를 화씨로 바꿈	한 값	아니오
두 열의 비율을 만듦	한 행	아니오
평균으로 결측값을 채움	여러 행	예
평균과 표준편차로 척도 조정	여러 행	예

Source: UBC CPSC 330 (2026), Lecture 5.

예측 시점 뒤에 정해지는 특성

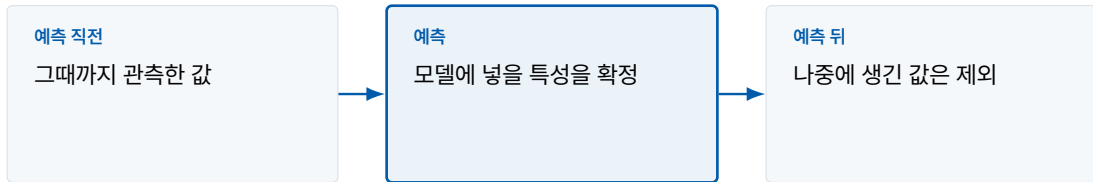
결과가 정해진 뒤에야 알 수 있는 값은 입력으로 쓸 수 없다

예측 문제와 특성	누수가 생기는 이유
입원 기간 예측에 퇴원일 사용	퇴원일을 알아야 실제 입원 기간이 정해짐
구매액 예측에 환불 여부 사용	환불은 구매 뒤에 관측되는 결과

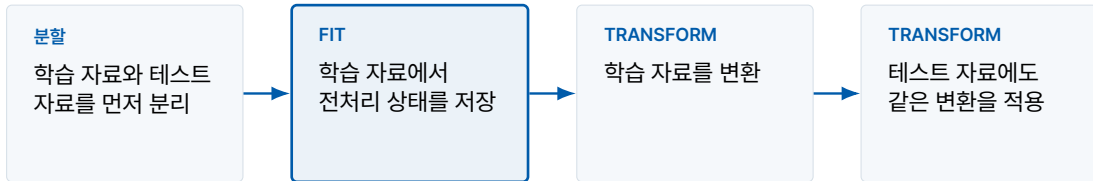
예측 시점에 알 수 없는 값은 특성에서 제외한다

예측 시점을 먼저 정한다

각 특성을 예측 시점에 관측할 수 있는지 확인한다



올바른 순서



Adapted from scikit-learn MOOC (2022), Transformer API diagrams.

fit과 transform

DEFINITION

fit

학습 자료에서 변환에 쓸 상태를 저장하는 단계

DEFINITION

transform

저장한 상태를 바꾸지 않고 자료에 같은 변환을 적용하는 단계

테스트 자료에는 transform만 적용하고 전처리 상태를 다시 fit하지 않는다

Transformer가 저장하는 상태

fit으로 저장한 상태를 학습 자료와 테스트 자료에 함께 사용한다



Adapted from scikit-learn MOOC (2022), Transformer API diagrams.

데이터 누수를 의심할 결과

관찰된 결과	먼저 확인할 것
테스트 성능이 지나치게 높다	분할 전에 전처리를 fit했는가
배포 뒤 성능이 크게 낮아진다	예측 시점 뒤의 특성을 썼는가
특성 하나만으로 정답이 거의 드러난다	예측할 값에서 파생된 값인가

누수가 있으면 오류 없이 실행돼도 성능 추정치가 비정상적으로 높아질 수 있다

Checkpoint

Preprocessing Leakage

- 대체와 표준화를 학습 자료에서만 fit해야 하는 이유
- 한 행만 보는 변환과 여러 행을 보는 변환의 차이
- 테스트 자료에 적용하지만 다시 fit하지 않는 단계

Pipelines

전처리와 모델의 순서를 한 단위로 실행하는 방법

순서를 손으로 관리할 때의 오류

작업	생길 수 있는 오류
단계를 따로 실행	테스트 자료에 fit을 잘못 호출
실험마다 순서를 변경	결과가 어느 실행 순서에서 나왔는지 알기 어렵다
새 자료에 다시 적용	학습 때와 다른 변환을 사용

DEFINITION

파이프라인(pipeline)

전처리 단계와 모델을 순서대로 묶어 fit과 predict를 한 단위로 실행하는 객체

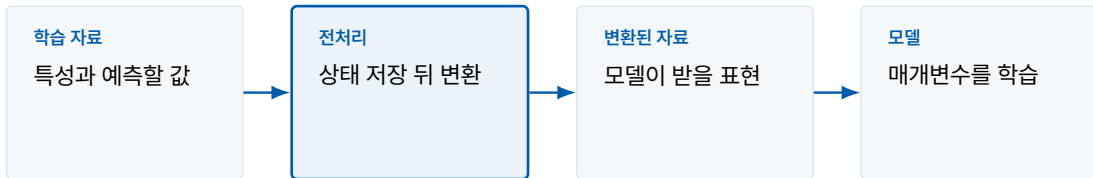
FIT

각 전처리 상태를 정한 뒤 모델을 학습

PREDICT

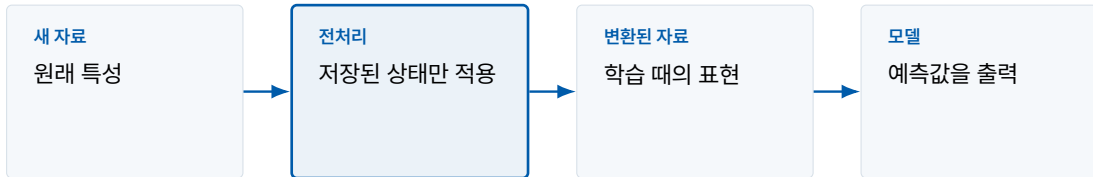
저장된 전처리를 적용한 뒤 예측값을 계산

pipeline.fit의 순서



Adapted from scikit-learn MOOC (2022), Pipeline fit diagram.

pipeline.predict의 순서



predict를 호출할 때 전처리 상태와 모델의 매개변수는 바뀌지 않는다

학습 자료로 fit하고 테스트 자료로 predict한다

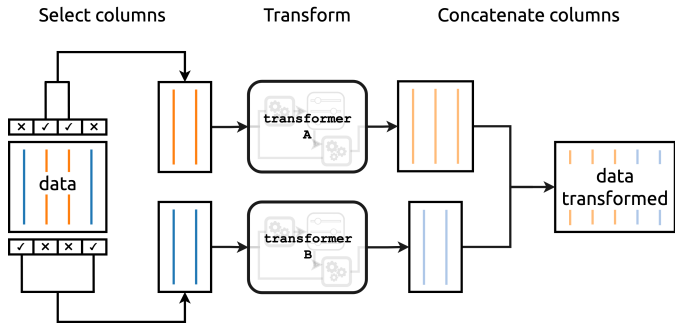
Auto 자료 397행으로 결측값 채우기, 표준화, 선형 회귀를 한 파이프라인으로 묶는다(import 생략)

```
1 auto = pd.read_csv("Auto.csv", na_values="?")
2 X = auto[["horsepower", "weight", "acceleration"]]
3 y = auto["mpg"]
4 X_train, X_test, y_train, y_test = train_test_split(
5     X, y, test_size=0.25, random_state=0)
6
7 pipe = make_pipeline(SimpleImputer(strategy="median"),
8                      StandardScaler(), LinearRegression())
9 pipe.fit(X_train, y_train)
10 y_pred = pipe.predict(X_test)
11 print(pipe[0].statistics_) # [ 95. 2865. 15.5]
```

pipe.fit은 학습 자료 297행에서만 중앙값, 평균, 표준편차를 구해 저장하고 모델을 학습한다. pipe.predict는 테스트 자료 100행을 저장된 값으로 transform만 한 뒤 예측하므로, 테스트 자료의 결측값도 학습 자료의 중앙값 95로 채운다. 선형 회귀는 파이프라인 끝의 예측 모델 자리에 둔 예다

열 종류마다 다른 처리

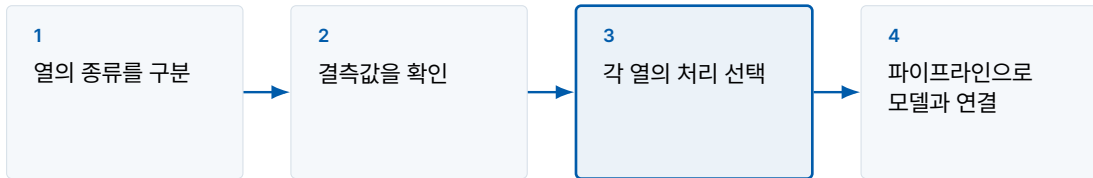
ColumnTransformer는 열을 나누어 변환한 뒤 다시 합친다



수치형 열과 범주형 열은 서로 다른 변환을 거친 뒤 하나의 특성 행렬로 합쳐진다

전처리 파이프라인을 정하는 순서

열의 종류와 결측값을 확인한 뒤 필요한 처리만 선택한다



Source: scikit-learn MOOC (2022), Predictive Modeling Pipeline.

순서가 어긋나는 세 가지 경우

잘못된 실행	결과
전체 자료를 척도 조정한 뒤 분할	테스트 자료의 분포가 전처리 상태에 반영
테스트 자료에서 <code>transformer.fit</code> 호출	테스트 자료마다 다른 입력 표현을 생성
테스트 성능을 보고 전처리 변경	테스트 자료를 모델 선택에 반복 사용

세 경우 모두 오류 없이 실행될 수 있지만 테스트 성능의 의미가 달라진다

Checkpoint

Pipelines

- `pipeline.fit`에서 전처리와 모델은 어떤 순서로 학습되는가
- `pipeline.predict`에서 다시 `fit`하지 않는 것은 무엇인가
- `ColumnTransformer`가 열을 나누어 처리하는 이유

Section 4 • Feature Engineering and Pipelines

Feature Engineering

문제의 구조를 새 특성으로 표현하는 방법

DEFINITION

특성 공학(feature engineering)

기존 열을 변환하거나 결합해 모델이 사용할 새 특성을 만드는 일

전처리

입력 형식, 결측값, 값 범위를 다룸

특성 공학

문제에서 중요한 관계를 새 열로 표현

한 행에서 만드는 특성

원래 열	새 특성	표현하는 관계
출발 시각	시간대	출퇴근 시간의 차이
방 수, 가구 수	가구당 방 수	주거 공간의 여유
날짜	요일	주중과 주말의 차이

Adapted from UBC CPSC 330 (2026), Lectures 5 and 13.

학습 자료의 통계로 만드는 특성

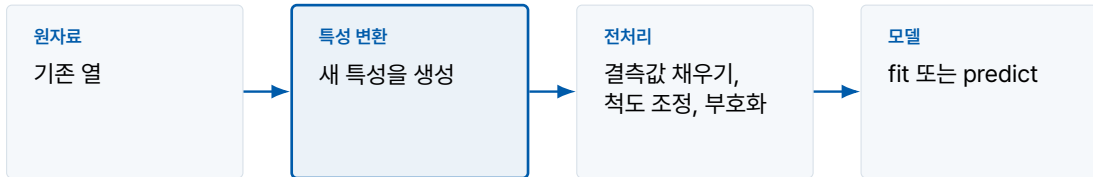
여러 행을 함께 보는 특성 변환도 학습 자료에서만 상태를 정한다

새 특성	필요한 상태	상태를 정할 자료
학습 자료 평균과의 차이	학습 자료의 평균	학습 자료
범주 출현 빈도	학습 자료의 빈도표	학습 자료
표준화된 비율	학습 자료의 평균과 표준편차	학습 자료

예: 학습 자료 $x = (1, 2, 3)$ 의 평균은 2라 테스트 값 4의 새 특성은 $4 - 2 = 2$ 다. 테스트 값까지 넣은 전체 평균 2.5를 쓰면 1.5가 된다

새 특성이 자료의 통계를 쓰면 그 계산도 파이프라인의 fit에 포함한다

특성 변환도 파이프라인 안에 둔다



같은 파이프라인을 쓰면 학습과 예측에서 같은 특성이 만들어진다

특성을 늘릴 때

새 특성은 모델이 표현할 수 있는 관계를 늘릴 수 있다

변화	현재 자료에서의 결과	새 자료에서의 결과
유용한 특성 추가	학습 오차 감소	테스트 오차도 감소 가능
불필요한 특성 다수 추가	학습 오차 감소	분산이 커져 테스트 오차가 늘 수 있다
테스트 자료로 특성 선택	테스트 자료에도 과적합	최종 성능을 과대 추정

검증 자료(validation data)는 특성·모델 선택에, 테스트 자료는 선택 후 최종 평가에 쓴다

자료를 나눈 뒤에는 역할을 바꾸지 않는다

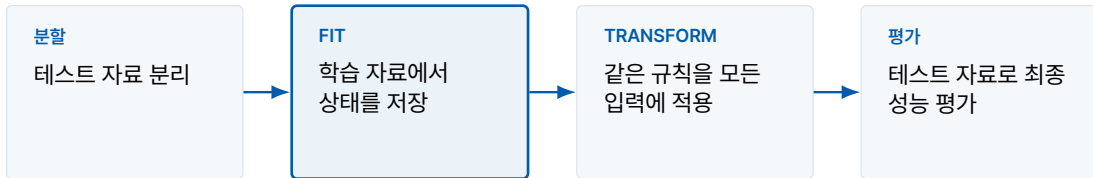
같은 성능 수치도 계산한 자료에 따라 쓰임이 달라진다

자료	정하는 것	하지 않는 일
학습 자료	전처리 상태와 모델 매개변수	최종 성능을 보고하지 않음
검증 자료	특성·모델 후보의 선택	전처리 상태를 다시 fit하지 않음
테스트 자료	선택을 마친 모델의 최종 성능	결과를 보고 선택을 바꾸지 않음

테스트 결과를 본 뒤 선택을 바꾸면 그 자료는 더 이상 최종 평가 자료가 아니다

전처리에서 지킬 순서

처리 규칙을 어디서 정했는지가 테스트 성능의 의미를 결정한다



전처리 규칙과 모델 매개변수는 학습 자료에서만 정한다

Checkpoint

Feature Engineering

- 전처리와 특성 공학은 무엇을 다르게 다루는가
- 여러 행의 통계를 쓰는 특성 변환은 어디에서 fit해야 하는가
- 특성을 고르는 동안 테스트 자료를 남겨 두는 이유

What Do We Choose When We Fit a Linear Model?

Next Lecture: Linear Regression Revisited

- 선형 모델의 계수와 절편
- 손실함수와 최소제곱
- 추정과 예측