앙상블 (Ensemble) - 실습

목차

[예제] 예측 – 집값 예측

- Bagging_예제.ipynb
- data/house regression.csv
- 데이터를 8:2로 분할하여 train, test 데이터 구축
- Decision Tree, Linear Regression, Decision Tree – Bagging, Linear Regression – Bagging

  → 총 4가지 종류의 모델을 구축
- Test 데이터에 대한 모델 성능 확인 – mean squared error, R-square
- 모델 성능 시각화

예제 - 예측

- RandomForest_예제.ipynb
- house regression.csv
- 데이터를 8:2로 분할하여 train, test 데이터 구축
- 5-CV Grid Search를 통해 최적 파라미터 탐색
- Test 데이터에 대한 모델 성능 확인 – mean squared error, R-square
- Scatter plot 등을 활용해 예측 결과 시각화
- 변수 중요도 산출 및 시각화, 해석

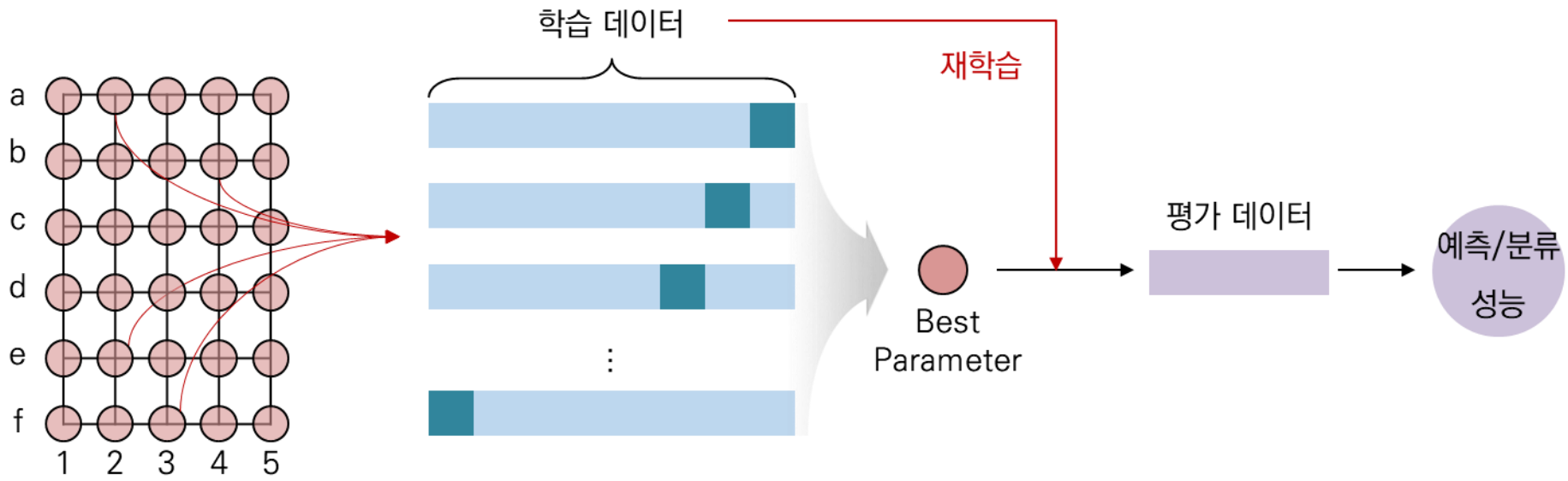
참고: Grid Search

- 목적: 학습 데이터를 기반으로 모델 관점에서 객관적으로 좋은 성능을 보이는 파라미터 탐색

- Grid search with 5-fold cross-validation (CV)

  → 사전에 정의한 모든 파라미터의 조합에 대하여 학습 데이터 내에서 5-CV를 수행

  → 가장 좋은 평가 지표를 보이는 파라미터를 찾아, 모델을 재학습한 후 평가 데이터에 적용



실습 - 분류

- RandomForest_분류_실습.ipynb
- UniversalBank.csv
- 데이터를 8:2로 분할하여 train, test 데이터 구축
- 5-CV Grid Search를 통해 최적 파라미터 탐색
- Test 데이터에 대한 모델 성능 확인 - accuracy
- 변수 중요도 산출 및 시각화, 해석

실습 - 분류

- 모델링에 필요 없는 변수: ID, ZIP Code
- Education – categorical 변수 → dummy 변수 생성 (pd.get_dummies)
- 고객의 인구통계학적 정보를 이용하여 개인신용대출 여부를 판별
- Personal Loan: 대출한 사람 = 1 / 대출하지 않은 사람 = 0

변수	변수설명
ID	Customer ID
Age	Customer's Age in completed years
Experience	# years of professional experience
Income	Annual income of the customer ($000)
ZIPCode	Home Address ZIP code
Family	Family size (dependents) of the customer
CCAvg	Average spending on credit cards of the customer ($000)
Education	Education level. 1: Undergrad 2: Graduate 3: Advanced/Professional
Mortgage	Value of house mortgage if any ($000)
Securities Account	Does the customer have s securities account with the bank?
CD Account	Does the customer have a certificate of deposit account with the bank?
Online	Does the customer use internet banking facilities?
CreditCard	Does the customer use a credit card issued by UniversalBank?
Personal Loan	Did this customer accept the personal loan offered in the last campaign?

실습 - 예측

- RandomForest_예측_실습.ipynb
- cold.csv
- 데이터를 8:2로 분할하여 train, test 데이터 구축
- 5-CV Grid Search를 통해 최적 파라미터 탐색
- Test 데이터에 대한 모델 성능 확인 – mean squared error, R-square
- Scatter plot 등을 활용해 예측 결과 시각화
- 변수 중요도 산출 및 시각화, 해석
- '일요일'과 나머지 요일에 대한 평균 진료 건수를 확인하시오

실습 - 예측

- 문제 상황: 전국의 기상정보 + SNS에서 '감기' 검색량으로 감기환자들이 병원에 방문하여 진료를 받은 건수를 예측하는 모델을 학습
- date 변수에 대한 dummy 변수를 생성
- number_treatment를 예측하는 모델을 구축

EOD