Homework #2 머신러닝 프로젝트의 과정 요약_2023170855 최정연

관련 데이터 선정 및 모델 선택

데이터 활용법 확인

-데이터셋의 크기와 특성, 머신러닝 모델의 target 확인

-훈련모델 확인

데이터 다운로드 및 적재 (pandas)

데이터 탐색 및 시각화

-샘플 확인 및 자료형 파악, 그래프를 통한 시각화

-훈련셋과 테스트셋 구분 (계층샘플링 활용; 소득 구간에 대한 비율 일치시키기)

-데이터 시각화, 상관계수 계산

데이터 정제와 전처리

-정제; 결측치 처리 (샘플 삭제, 결측치 특성 삭제, 중앙값 또는 평균값 등으로 결측치 대체)

-사이킷런 API (추정기fit, 변환기fit+transform, 예측기fit+predict 세 클래스의 인스턴스로 생성됨)

-simpleimputer 변환기: 결측치 처리

-입력 데이터셋(housing)과 타깃 데이터셋(housing_label) 지정

-one hot encoder 변환기: 범주형 데이터 전처리

-MinMax Scaler, Standard Scaler 변환기: 수치형 데이터 스케일링 (치우쳐진 데이터 해결)

-Function Transformer 변환기; 로그변환기, 비율계산변환기 (fit 사용x, transform)

-군집 변환기 (위도 경도 상의 데이터를 군집화로 새 데이터 생성)

파이프라인; 전처리부터 모델학습 등의 과정을 묶어서 하나의 객체로 사용하도록 처리

모델 선택과 훈련 (선형회귀모델, 결정트리회귀모델, 랜덤포레스트회귀모델, ... 선택 / 교차검증)

모델 미세 조정

-그리드 탐색; 하이퍼파라미터의 조합 탐색으로 최적의 조합 선정

-랜덤 탐색; 하이퍼파라미터 무작위 선택으로 훈련

최적 모델 저장 및 활용