2주차 과제- 머신러닝 프로젝트 기획 요약

2025170941 이준서

큰 그림 -> 데이터 구하기 -> 탐색 & 시각화 -> 데이터 준비 -> 모델 선택 & 훈련 -> 모델 조정 -> 솔루션 제시 -> 모니터링 & 유지 보수

머신러닝으로 문제를 해결할 때, 위 순서대로 기획된다. (python으로 진행)

1. 큰 그림

i)기초 정보 확인

데이터셋 크기와 특성(ex 주택가격 데이터 안에 여러 개의 특징들: 경도, 위도, 침실 수, 인구, 가구 수, 주택 건물 중위연령, 총 방 수, 중위소득, 주택 중위가격, 해안 근접도)

머신러닝 모델의 타깃(ex 주택 중위가격을 제외한 나머지 특성이 주어졌을 때, 주택 중위가격을 예측하는 시스템)

ii)훈련 모델 확인

구역별 주택 중위가격을 타깃으로 예측하는 시스템에 활용될 회귀 모델을 훈련시킨다(ex 지도 학습, 회귀, 배치 학습)

2. 데이터 구하기

함수로 주택가격 데이터를 다운로드 -> pandas 데이터 프레임으로 변환하여 반환

3. 데이터 탐색 & 시각화

i) 탐색

head() 매서드를 통해 샘플 확인 -> info() 매서드를 통해 데이터셋 정보 요약 -> value\_counts() 매서드를 통해 범주형 특성 탐색(ex 해안 근접도), describe() 매서드를 통해 수지형 매서드 탐색 -> 그래프 기본 설정 -> 훈련 셋과 테스트 셋을 구분하기 위해 train\_test\_split() 함수 이용 -> 계층 샘플링을 위해 소득 구간을 5개로 구분 -> 소득 구간을 새로운 범주형 특성으로 추가 -> 소득을 기준으로 계층 샘플링 진행 -> 계층 샘플링의 정확도를 평가하기 위해 무작위 샘플링 결과를 이용 -> 계층 샘플링과 무작위 샘플링의 결과 비교 -> 데이터 되돌리기

ii) 시각화

그래프 기본 설정 -> 지리적 데이터 시각화() grid 키워드 인자 활용 -> 투명도 조절을 통해 밀집도를 선명하게 시각화(alpha 키워드 인자 활용) -> 주택 중위가격을 색상으로 활용 -> 캘리포니아 지도와 합치기 -> 다운로드한 캘리포니아 지도와 위도/경도를 이용한 산포도를 함께 그리기 -> 수치형 특성 간의 상관관계 계산(산점도로 시각화) ->

4. 데이터 준비(전처리)

결측치 처리법 결정 -> 결측치가 존재하는 샘플 확인 -> 결측치 처리(결측치가 포함된 샘플 삭제 or 특성 삭제 or 결측치를 해당 특성의 중앙값/평균값 등으로 대체)

5. 모델 선택 & 훈련

선형 회귀 모델 훈련 -> 예측값과 오차 발생, 예측값의 RMSE가 많이 높음 -> 과소 적합, 선형 회귀 모델 부적절 -> 결정 트리 회귀 모델 훈련 -> RMSE가 아주 적음 -> 과대 적합, 결정트리 회귀 모델 부적절 -> 랜덤 포레스트 회귀 모델 훈련 -> RMSE가 결정트리 보다는 높으나 선형회귀 모델보다는 훨씬 낮음 -> 교차 검증 -> 훈련 과정 중의 모델 성능 평가 진행 -> 랜덤 포레스트 회귀 모델 성능 검증

6. 모델 미세 조정

그리드 탐색(파이프라인에 포함된 전처리와 예측기에 사용되는 하이퍼파라미터 중에서 미세조정에 사용될 하이퍼파라미터가 가질 수 있는 값들의 리스트를 지정) -> 최적의 하이퍼파라미터 조합, 최적의 모델 성능 평가 -> 랜덤 탐색 -> 최적의 하이퍼파라미터 조합, 최적 모델 활용