2 주차 세션 내용 요약

캘리포니아 주택 가격 예측 모델 만들기

목표:
캘리포니아 주의 각 구역에 대한 **중위 주택 가격(median_house_value)**을 예측하는 **머신러닝 모델**을 훈련시키는 것.

활용할 데이터:
인구, 소득, 방 개수, 침실 개수, 해안 근접도 등 **주택 관련 특성들**

예측 대상:
median_house_value (중위 주택 가격)

실습 코드 내용 요약

1. 환경 준비

Python 3.7 이상, scikit-learn 1.0.1 이상

무작위성 제어를 위한 np.random.seed(42) 설정

2. 데이터 수집

GitHub 에서 캘리포니아 주택 데이터(tgz 형식)를 다운로드 → housing.csv 로드

load_housing_data() 함수로 데이터프레임 반환

3. 데이터 탐색

데이터 구조, 결측치, 통계 요약

ocean_proximity 는 **범주형**

나머지는 **수치형**

4. 시각화

히스토그램: 수치형 특성들의 분포 파악

위도/경도 기반 산점도: 지역별 분포 시각화

색상과 원 크기로 인구 수와 가격 표시

캘리포니아 지도와 시각화 겹쳐서 실제 지역 감각 제공

5. 훈련/테스트셋 분리

단순 무작위 샘플링 vs. 계층 샘플링 비교

소득 기반 계층 샘플링: 대표성 유지

6. 상관관계 분석

median_income 이 주택 가격과 가장 높은 양의 상관관계

scatter_matrix()와 산점도 시각화로 관계 분석

7.데이터 전처리 및 정제

결측치 처리

total_bedrooms 에 168 개 결측치 → 중앙값(SimpleImputer(strategy="median"))으로 대체

수치형 & 범주형 분리

수치형: 표준화(StandardScaler), 정규화(MinMaxScaler), 로그변환

범주형: OneHotEncoder 사용

사용자 정의 변환기

ClusterSimilarity: 위도/경도를 KMeans 로 군집화 후 유사도 추가 특성 생성

8.전처리 파이프라인 구성

Pipeline: 여러 전처리 작업을 하나로 연결

ColumnTransformer: 수치형, 범주형 등 **타입별 전처리** 적용

새로운 특성: 가구당 인구수, 침실 비율, 로그 변환, 군집 유사도

최종 전처리 결과: 24 개의 특성

9.모델 선택과 훈련

모델들

선형회귀: 과소적합 (RMSE ≈ 68,000)

결정트리: 과대적합 (RMSE ≈ 0)

랜덤 포레스트: 우수한 성능 (RMSE ≈ 17,521)

교차 검증 결과

cross_val_score()로 모델 성능 비교

랜덤포레스트 평균 RMSE: **~46,938**

10.모델 튜닝 (하이퍼파라미터 탐색)

Grid Search

군집 수(n_clusters), max_features 튜닝

최적 조합: n_clusters=15, max_features=6

평균 RMSE: **~43,616**

Random Search

무작위 10 개 조합 테스트

최고 성능: n_clusters=45, max_features=9

평균 RMSE: **~42,107**

특성 중요도 분석

가장 영향 큰 특성:

'log__median_income', 'ocean_proximity_INLAND', 'bedrooms__ratio'

11.모델 저장 및 로딩

최종 모델 저장: joblib.dump()

모델 재사용: joblib.load() → .predict() 호출