2025.3.24

머신러닝 세션 2주차 요약

머신러닝 프로젝트 진행 과정

1. 문제 정의

2. 데이터 수집

3. 데이터 탐색 및 시각화

데이터를 훈련셋과 테스트셋으로 나누고 훈련셋들을 복사하여 복사본을 시각화 한다.

4. 데이터 준비

결측치가 있는 샘플을 확인하고 처리한다.

처리방법 : 결측치 특성 포함 샘플 삭제, 결측치 포함 특성 삭제, 결측치를 해당 특성의 중앙값, 평균값으로 대체

비율 변환기, 로그 변환기, 군집 변환기, 기본 변환기 등을 사용하여 데이터를 전처리한다.

범주형 데이터는 원핫인코딩 기법을 사용하여 수치화한다.

원-핫 인코딩 : 범주 수 만큼의 새로운 특성을 추가하여 해당되는 범주와 관련된 특성값은 1, 나머지 특성값은 0으로 하는 행렬로 표현

전처리 과정과 이어 나올 모델을 하나의 파이프라인으로 묶어 정의한다.

5. 모델 선택 및 훈련

선형 회귀 모델, 결정트리 회귀 모델, 랜덤포레스트 회귀 모델 등 다양한 모델을 사용하여 훈련한 후 RMSE를 비교하여 최적의 모델을 선택한다.

RMSE : 예측값과 실제값 사이의 차이를 평가하는 대표적인 회귀 평가 지표

6. 모델 조정

cross_val_score() 함수를 이용하여 교차검증하여 모델 성능을 평가한다.

그리드 탐색을 통해 최적의 하이퍼파라미터를 찾아낸다.

찾아낸 하이퍼파라미터 조합에 대해 3-겹 교차 검증을 진행한다.

7. 솔루션 제시

8. 시스템 론칭, 모니터링 유지 보수