5.9 회귀 실습 – 자전거 대여 수요 예측

로그 변환 후 예측 평가 결과

- 모델들이 count 를 바로 예측하는 대신 **log 변환 → 예측 → 역변환** 방식 사용
- 출력 지표:
    - **RMSLE**: 수요 예측에 적합 (실제 비율 차이 반영)
    - **RMSE / MAE**: 절대 오차 기반 평가

>> 일반적인 RMSE 보다 RMSLE 가 더 중요한 상황 → 자전거 대여량처럼 큰 값이 드물고 작은 값이 많을 때 적합

모델별 성능 비교 결과

- Linear Regression, Ridge, Lasso 비교

- 규제 모델(Ridge/Lasso)로 과적합 줄임

- Lasso 는 불필요한 피처 계수 0 → 변수 선택 효과

앙상블 모델 결과

트리 기반 모델들은 일반적으로 선형 모델보다 성능 좋음

특히 XGBoost / LightGBM 이 높은 성능 가능

>> 데이터 패턴이 비선형 → 트리 모델에서 더 좋은 성능

계수/피처 중요도 시각화 결과

- 대여량에 가장 영향을 많이 준 변수들 출력

- 주로 중요했던 피처: hour_XX, workingday, season, weather, 특정(출퇴근) 시간대
>> 자전거 대여는 시간대 특징이 매우 중요
>> 출퇴근 시간에 대여량 급증 → 실제 패턴 반영 성공

5.10 회귀 실습 – 캐글 주택 가격 : 고급 회귀 기법

Kaggle의 **House Prices** 데이터를 기반으로 **회귀 모델을 활용하여 주택 가격을 예측**

데이터 전처리 흐름

1) SalePrice 분포 확인 → **정상분포가 아님(비대칭, 왜도 존재)**

2) SalePrice와 왜도가 큰 변수들 : $\log_{1+}x$ 변환

3) 결측치 처리 (너무 결측 많은 컬럼 제거, 수치형 → 평균 대체, 범주형 → 원-핫 인코딩)

모델 학습 및 평가

1) Baseline Regression

- RMSE로 평가
- 로그 변환 후 RMSE 개선 확인

2) 하이퍼파라미터 튜닝

- GridSearchCV 사용
- Ridge, Lasso 최적 alpha 탐색
- 교차검증(CV) 기반 평균 RMSE 계산

3) 트리 기반 모델

- 설정한 파라미터로 훈련 후 예측
- Ridge/Lasso보다 비선형 데이터에서 강함

4) 모델 블렌딩(혼합)

- Ridge + Lasso → 가중 평균
- XGB + LGBM → 가중 평균

서로 다른 모델 장점 결합 → RMSE 감소

최종 결론

1. 전처리 → 선형 모델 → 트리 기반 모델 → 스태킹 앙상블 순서로 진행.

2. 로그 변환 + 원핫인코딩 + 모델 스태킹이 가장 높은 성능을 보임

3. 단일 모델보다 조합 모델의 성능 우수성을 확인