[5.9장 회귀 실습 – 자전거 대여 수요 예측]

선형회귀와 트리 기반 회귀 비교 – 캐글의 자전거 대여 수요 이용

주요 칼럼 (결정 값: count, 대여 횟수)

- datetime: hourly date + timestamp
- season: 1 = 봄, 2 = 여름, 3 = 가을, 4 = 겨울
- holiday: 1 = 토, 일요일의 주말을 제외한 국경일 등의 휴일, 0 = 휴일이 아닌 날
- workingday: 1 = 토, 일요일의 주말 및 휴일이 아닌 주중, 0 = 주말 및 휴일
- weather:
  - 1 = 맑음, 약간 구름 낀 흐림
  - 2 = 안개, 안개 + 흐림
  - 3 = 가벼운 눈, 가벼운 비 + 천둥
  - 4 = 심한 눈/비, 천둥/번개
- temp: 온도(섭씨)
- atemp: 체감온도(섭씨)
- humidity: 상대습도
- windspeed: 풍속
- casual: 사전에 등록되지 않는 사용자가 대여한 횟수
- registered: 사전에 등록된 사용자가 대여한 횟수
- count: 대여 횟수

데이터 클렌징 및 가공

datatime 칼럼만 object형이므로 4개의 속성으로 분리 후 월 일 시간 분 초로 변환

판다스는 문자열을 datetime 타입으로 변환하는 apply(pd.todatetime) 메서드 제공

캐글요구 성능 평가 방법 – RMSLE

주의할 점: 오버플로/언더플로 오류가 발생하기 쉬움

해결책: log()보다는 log1p()를 이용 (1+log()값) 넘파이의 expm1()함수로 쉽게 원래의 스케일로 복원 가능

로그 변환, 피처 인코딩과 모델 학습/예측/평가

결괏값이 정규 분포로 되어있는지 확인 & 카테고리형 회귀 모델의 경우 원-핫 인코딩으로 피처를 인코딩하는 것

- LinearRegression 객체 이용하여 회귀 예측
- 판다스 DatFrame의 hist() 이용하여 count 정규 분포 이루는지 체크
- 왜곡된 값을 바꾸는 방법은 일반적으로 로그 변환
- 로그 변환된 값과 원-핫 인코딩된 피처 데이터 세트를 그대로 이용해 랜덤 포레스트, GBM, XGBoost, LightGBM 순차적으로 성능 평가
- 넘파이 ndarray 로 변환

[5.10장 회귀 실습 – 캐글 주택 가격: 고급 회귀 기법]

데이터 사전 처리 (Preprocessing)

- 정규 분포 형태로 변환하기 위해 로그 변환을 적용시킴
- Null 값인 숫자형 피처만 평균값으로 대처해준다
- 문자형 피처를 모두 원-학 인코딩으로 변환해준다 → get_dummies() 이용

선형 회귀 모델 학습/예측/평가

로그 변환된 RMSE 측정

- 라쏘 회귀는 최적 하이퍼 파라미터 튜닝을 해준 후에 수행
- 라쏘 모델의 경우는 릿지에 비해 동일 피처여도 회귀 계수 값이 매우 작음
- 피처 데이터 세트의 분포도를 확인 → 왜곡된 피처 체크
- 사이파이 stats 모듈의 skew() 함수를 이용하여 칼럼의 데이터 세트의 왜곡된 정도 추출
- 주의할 점: skew() 를 적용하는 숫자형 피처에서 원-핫 인코딩된 카테고리 숫자형 피처는

제외

- 이상치 데이터 분석 통해 이상치를 찾고 GrLivArea 속성이 회귀 모델에서 차지하는 영향도가 크기 때문에 이 이상치를 개선하는 것을 목표로 함

회귀 트리 모델 학습/예측/평가

- 시간이 오래 걸리는 것을 감안하여 하이퍼 파라미터 설정을 미리 적용한 상태로 5 폴드 세트에 대한 평균 RMSE 값을 구함

회귀 모델의 예측 결과 혼합을 통한 최종 예측

- 개별 값을 혼합해서 최종 회귀 값 예측
- 앞 모델의 40퍼와 뒤 모델의 60퍼를 더해 최종 회귀 값으로 예측
- get_rmse_pred() 함수 생성

스태킹 앙상블 모델을 통한 회귀 예측

- 개별적인 기반 모델과 개별 기반 모델의 예측 데이터를 학습 데이터로 만들어서 학습하는 최종 메타 모델이 필요
- 스태킹 모델의 핵심은 여러 개별 모델의 예측 데이터를 각각 스태킹 형태로 결합하여 최종 메타 모델의 학습용 피처 데이터 세트와 테스트용 피처 데이터 세트를 만드는 것
- get_stacking_base_datasets() 함수 이용
- 회귀에서 특히 효과적으로 사용됨