Airbnb 데이터를 이용한 회귀분석

(https://www.kaggle.com/stevezhenghp/airbnb-price-prediction)

Air B&B 를 통해 대여한 주택의 1 일 이용 가격을 설명하는 모형을 만들고자 한다.
데이터는 "airbnb.xlsx" 파일에 담겨있다. 이 데이터에 내장된 변수는 다음과 같다.

변수명	변수명	변수명	변수명
log_price	cancellation_policy	host_response_rate	neighbourhood
property_type	cleaning_fee	host_since	number_of_reviews
room_type	city	instant_bookable	review_scores_rating
amenities	description	last_review	thumbnail_url
accommodates	first_review	latitude	zipcode
bathrooms	host_has_profile_pic	longitude	bedrooms
bed_type	host_identity_verified	name	beds

변수 변환 혹은 제거 리스트 (언급되지 않은 변수는 그대로 유지)

1. "number_of_reviews"가 11개 이상인 데이터만 추출하시오.
2. "property_type"은 'House', 'Aprtment', 'Other' 등의 3범주로 변환하시오. I_house, I_apartment 등의 2개 더미변수를 생성하고, "property_type"은 삭제하시오.
3. "room_type"은 'share room'=1, 'private room'=2, 'entire home/apt'=3 으로 정수형으로 변환하시오.
4. "amenities"는 amenities에 포함된 편의시설의 갯수로 정의하시오.
5. "bed_type"은 'Real Bed'인 경우는 1, 그 외의 경우는 0으로 더미변수화 하시오,
6. "cancellation_policy"는 5개의 순서가 존재하는 범주형이므로, 이를 1,2,3,4,5의 정수형으로 변환하시오. (flexible=1, moderate=2, strict=3, super_strict_30=4, super_strict60=5).
7. "cleaning fee" 는 더미변수화 하시오.
8. "description" 변수는 문자열의 길이로 정의하시오. (더 긴 소개문을 제공한 곳은 더 비싼지 여부 확인해보기 위해)
9. "host_identity_verified" 변수는 더미변수화 하시오.
10. "instant_bookable" 변수는 더미변수화 하시오.
11. "latitude"와 "longitude"를 이용하여 "도심의 중심위치로부터의 거리" 라는 변수를 추가하시오.
12. 로그가격비(log_price_ratio)' 변수를 생성하시오. 여기서, 가격비는 아래와 같다.
    - $\textit{로그가격비} = log(\frac{\textit{원가격}}{\textit{도시별 평균가격}} \times 100)$
    - 여기서 '원가격' $= e^{\log\_price}$, '도시별 평균가격'은 같은 도시내의 '원가격'의 평균값을 의미한다.
13. 'id', 'first_review', 'host_has_profile_pic', 'host_since', 'last_review', 'latitude',

'longitude', 'city_lat', 'city_long', 'price', 'avg_price_by_city', 'name', 'neighbourhood', 'thumbnail_url', 'zipcode', 'city', 'log_price' 변수를 삭제하시오.

14. 결측치가 있는 데이터는 삭제하시오.

- 로그가격비(log_price_ratio)"를 종속변수로 하여 아래와 같은 회귀분석을 수행하시오.

분석리스트

1. *선형회귀분석 (statsmodels OLS)*
2. *DecisionTreeRegressor*
3. *MLPRegressor*
4. *SVR (linear)*
5. *SVR (rbf)*
6. *BaggingRegressor*
7. *RandomForestRegressor*
8. *AdaBoostRegressor*
9. *GradientBoostingRegressor*

분석순서

1. Variable selection 을 수행한다. 변수선택은 랜덤포레스트의 변수중요도를 이용한다, 변수중요도가 거의 없는 변수들을 제거하고 나서 진행한다.
2. 데이터를 train:test = 5:5의 비율로 분할한다.
3. 모형은 train 데이터로 학습하고, test데이터로 평가한다.
4. 선형회귀분석 (statsmodels OLS) 과 DecisionTreeRegressor 는 결과해석을 시도한다. Tree는 depth 2단계 까지만 해석한다. 그 외 방법은 해석은 하지 않는다.
5. *MLPRegressor 와 SVR 방법은 표준화를 사용하여 학습한다.*
6. 모형의 예측력 평가시 기준은 MAE와 예측R^2를 사용한다.
7. Lineplot을 그리는데, x축은 방법이름, y축은 MAE인 그래프로 그린다.
8. Lineplot을 그리는데, x축은 방법이름, y축은 예측R^2인 그래프로 그린다.

방법별 세부 옵션

1. 모든 방법을 사용할때, (옵션이 있다면) random_state=0 을 사용하여, 재현이 가능하도록 한다.
2. DecisionTreeRegressor는 'ccp_alpha 옵션으로 튜닝한다.
3. MLPRegressor: hidden layer의 갯수를 1개와 2개를 시도하고, 각 layer내 node의 개수를 튜닝한다. 그 외 옵션은 디폴트로 한다.
4. SVR: C값을 튜닝한다. 그 외 옵션은 디폴트로 한다.
5. RandomForestRegressor 는 max_feature 옵션으로 튜닝한다.
6. GradientBoostingRegressor 는 max_depth 옵션으로 튜닝한다.
7. 모든 앙상블 방법은 n_estimators=100 으로 한다.