Data

아파트 실거래가 예측

Forecast of actual transaction price of apartments

:01

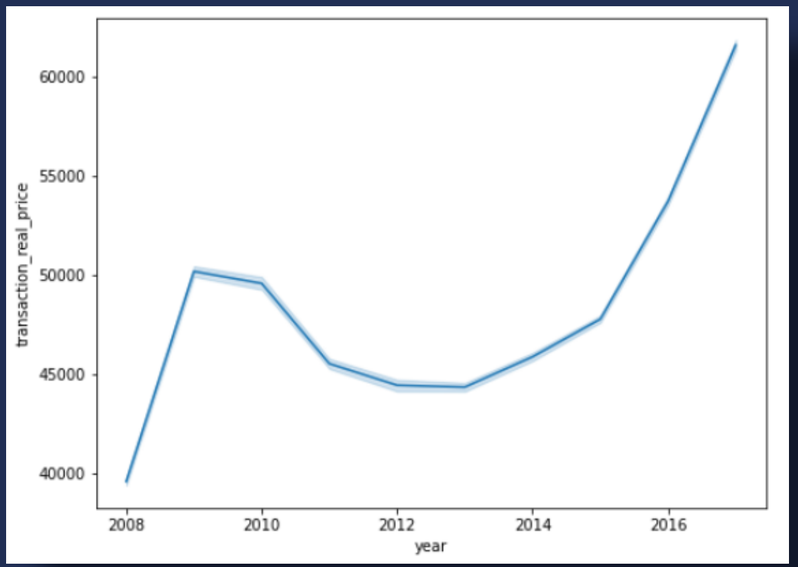
목차

Contents

01 프로젝트 목표와 데이터셋 선정

프로젝트 목표 : 아파트 구매자들을 위한 예측 모델

1. 최근 부동산 시장에 대한 관심이 높아짐
    a. 부동산은 투자의 대상으로 인식
    b. 수도권의 경우 인구 밀집이 높고, 대도시로의 인구 쏠림 현상이 발생하면서 수도권 아파트 매매가에 대한 관심이 높아짐
2. 데이터 분석의 목표와 목적
    a. 목표 : 아파트 매매가에 영향을 미치는 요인들을 분석, 거래가 예측
    b. 목적 :
        i. 아파트 구매자에게 정보 비대칭을 해결
        ii. 중개사와 구매자를 연결하여 부동산정보 서비스 시장의 신뢰도를 높임



01 프로젝트 목표와 데이터셋 선정

프로젝트 목표 : 아파트 구매자들을 위한 예측 모델

1. 데이터는 부동산 기업 앱 '직방'에서 데이터를 제공
2. 2008 ~ 2017년도 서울 지역의 74만 여개의 실거래 데이터
    a. 전용면적, 설립일자, 거래년월,거래날짜,층, 실거래가 등

apartment_id	city	dong	jibun	apt	addr_kr	exclusive_use_area	year_of_completion	transaction_year_month	transaction_date	floor	transaction_real_price
7622	서울특별시	신교동	6-13	신현(101동)	신교동 6-13 신현(101동)	84.82	2002	200801	21~31	2	37500
5399	서울특별시	필운동	142	사직파크맨션	필운동 142 사직파크맨션	99.17	1973	200801	1~10	6	20000
3578	서울특별시	필운동	174-1	두레엘리시안	필운동 174-1 두레엘리시안	84.74	2007	200801	1~10	6	38500
10957	서울특별시	내수동	95	파크팰리스	내수동 95 파크팰리스	146.39	2003	200801	11~20	15	118000
10639	서울특별시	내수동	110-15	킹스매너	내수동 110-15 킹스매너	194.43	2004	200801	21~31	3	120000

02 머신러닝 문제 정의

프로젝트 목표에 맞는 머신러닝 학습 모델 설정

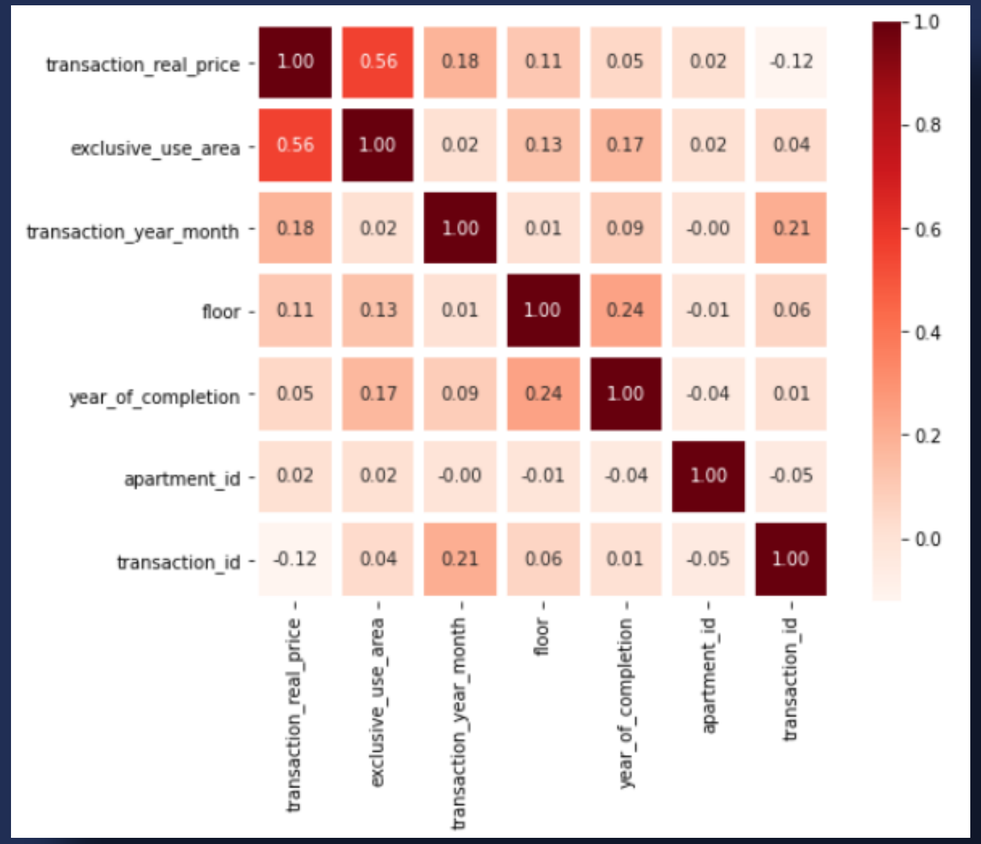
기간 : 2008년 1월 ~ 2017년 12월

머신러닝 문제 정의

- 매매가 예측이 목표이므로 독립 변수로 서울 아파트 매매가로 선정
- 매매가는 선형적으로 변화하는 특성을 보임 선형회귀로 분석해야 함
- RMSE(Root Mean Squared Error) 값으로 각 모델의 성능을 비교 분석
- 사용하는 머신러닝 모델은 선형회귀, 랜덤 포레스트, XGBoost

서울 아파트 매매 가격과 상관관계

- 상관계수를 통해 매매가를 기준으로 각 칼럼들 간의 상관관계를 확인
- 매매가와 연관성이 높은 컬럼 순은 전용면적, 거래년월, 층
- 매매가와 가장 연관성이 낮은 컬럼은 아파트 브랜드

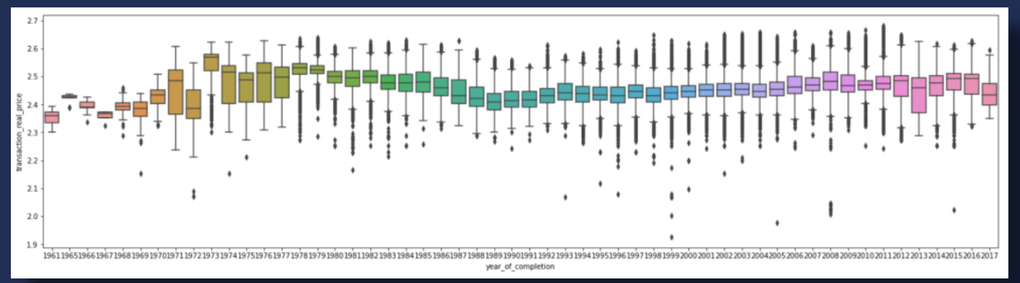


03 EDA & 전처리

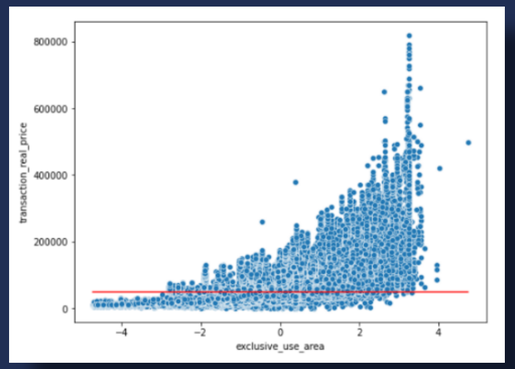
데이터의 결측치, 이상치, 중복값을 확인

데이터 분석 및 정제

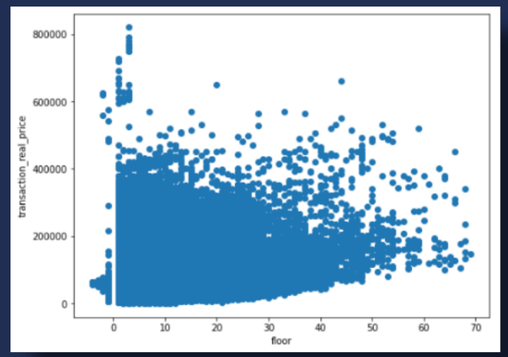
1. 전용면적과 매매가의 관계 : 선형 관계 - 전용면적이 클수록 매매가 증가
2. 층과 매매가의 관계 : 층이 높을수록 매매가가 높아 지는 것은 아님
3. 거래년도와 매매가 데이터 확인
   a. 매우 낮은 매매가는 지하층, 지하층 전체 데이터 확인 결과
   모든 지하층이 낮은 매매가는 아니었습니다.



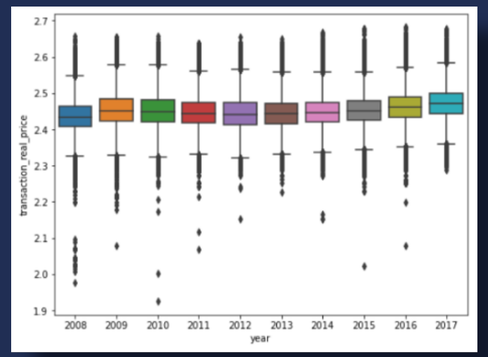
거래년도 별 매매가격



매매가와 전용면적의 연관성



매매가와 층의 연관성



매매가와 거래년도의 연관성

03 EDA & 전처리

데이터의 결측치, 이상치, 중복값을 확인

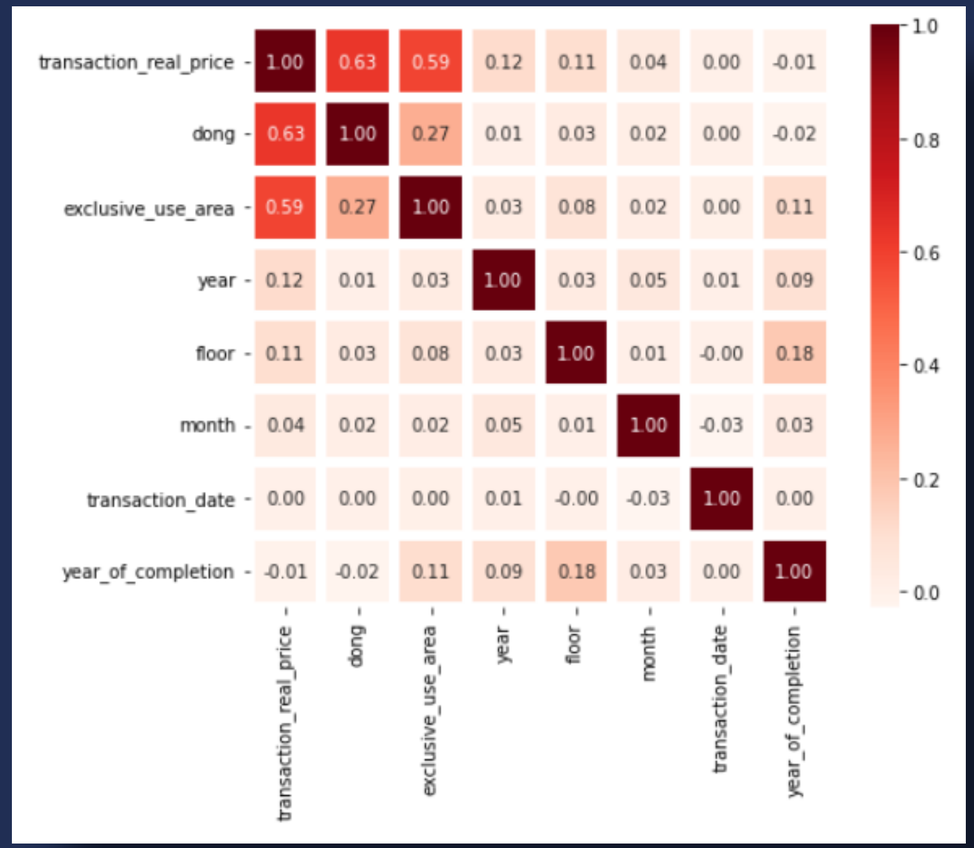
기간 : 2008년 1월 ~ 2017년 12월

데이터 정제

- 층, 매매가격, 전용면적을 정규화
- 사용하지 않는 컬럼 아파트 브랜드, 지번 주소 등 삭제

매매가 상관관계

- 매매가와 연관성이 높은 컬럼 지역, 평수, 거래년도 순



04 모델 학습 및 검증

여러가지 머신러닝 모델을 학습하여 비교

기간 : 2008년 1월 ~ 2017년 12월

교차검증(Cross Validation)

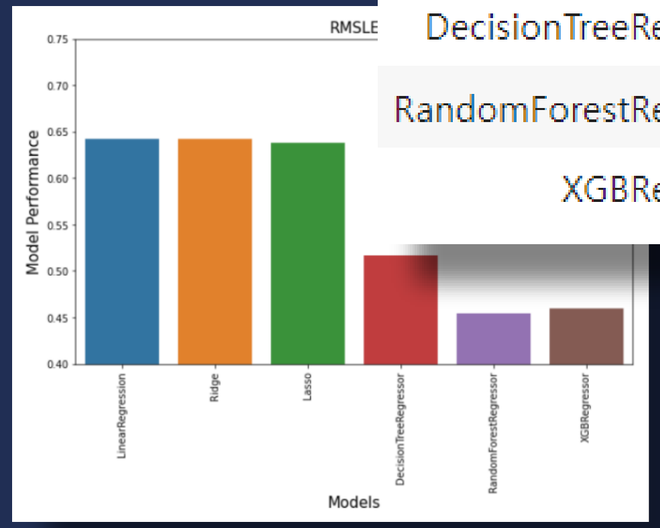
- Train 데이터를 Train,validation,Test로 분리하여 과소 적합 해소

RMSE 측정값 비교

머신러닝 모델의 RMSE 성능 비교

- 릿지(Ridge), 라쏘(Lasso) , 선형회귀, 의사결정나무(DecisionTree), 랜덤포레스트(Random Forest) , XGBoost 의 RMSE값 비교
- 랜덤포레스트의 RMSE 값이 0.454로 최소
- Random Forest 모델을 최종 모형으로 선정

Model	Score
LinearRegression	0.642496
Ridge	0.642495
Lasso	0.638354
DecisionTreeRegressor	0.516240
RandomForestRegressor	0.454641
XGBRegressor	0.460051



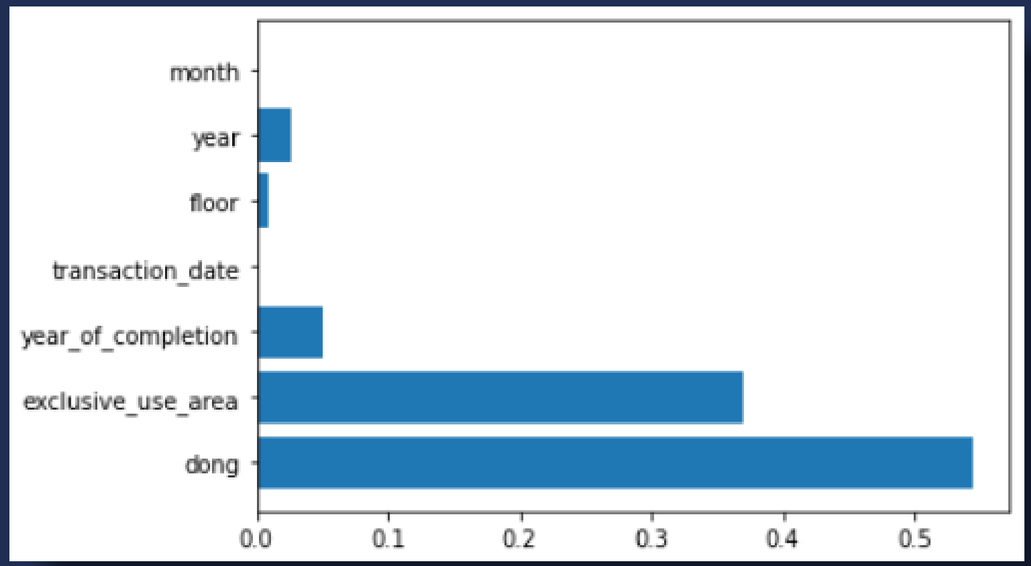
04 모델 학습 및 검증

데이터의 결측치, 이상치, 중복값을 확인한다.

기간 : 2008년 1월 ~ 2017년 12월

특성중요도

- RandomForestRegressor 사용하여 Train,validation 학습
  - n_estimators=10
  - max_depth=9
  - min_samples_split=50
  - min_samples_leaf=5
  - n_jobs=-1)
- Test 데이터로 예측
  - 동, 전용면적, 완공연도 순으로 중요도가 높음
- RandomForest RMSE값
  - Train_RMSE : 0.198
  - Valid_RMSE : 0.196



처음 가설은 전용면적, 거래년월, 층이 영향이 높을 것으로 예상 했으나
실제 분석 결과 지역과 평수, 설립연도가 영향력이 높았습니다.

AI_17_이다인
github.com/leedain0301