서울특별시 아파트 가격 예측 모델링

OVERVIEW

Project Outline

- Introduction
- EDA(Exploratory Data Analysis)
- Modeling
  - Pipeline(Linear Regression, Randomforest, XGBBoost, LGBM)
  - RandomizedCV
  - Evaluation
- XAI(eXplainable AI)
  - Global(Feature Importance, Permutation Importance with ELI5)
  - Local(SHAP)
- Conclusion

01 Background

: 부동산 적정가 산출의 필요성

1) 각종 규제정책에도 불구하고 부동산 가격은 지속적으로 증가하는 추세

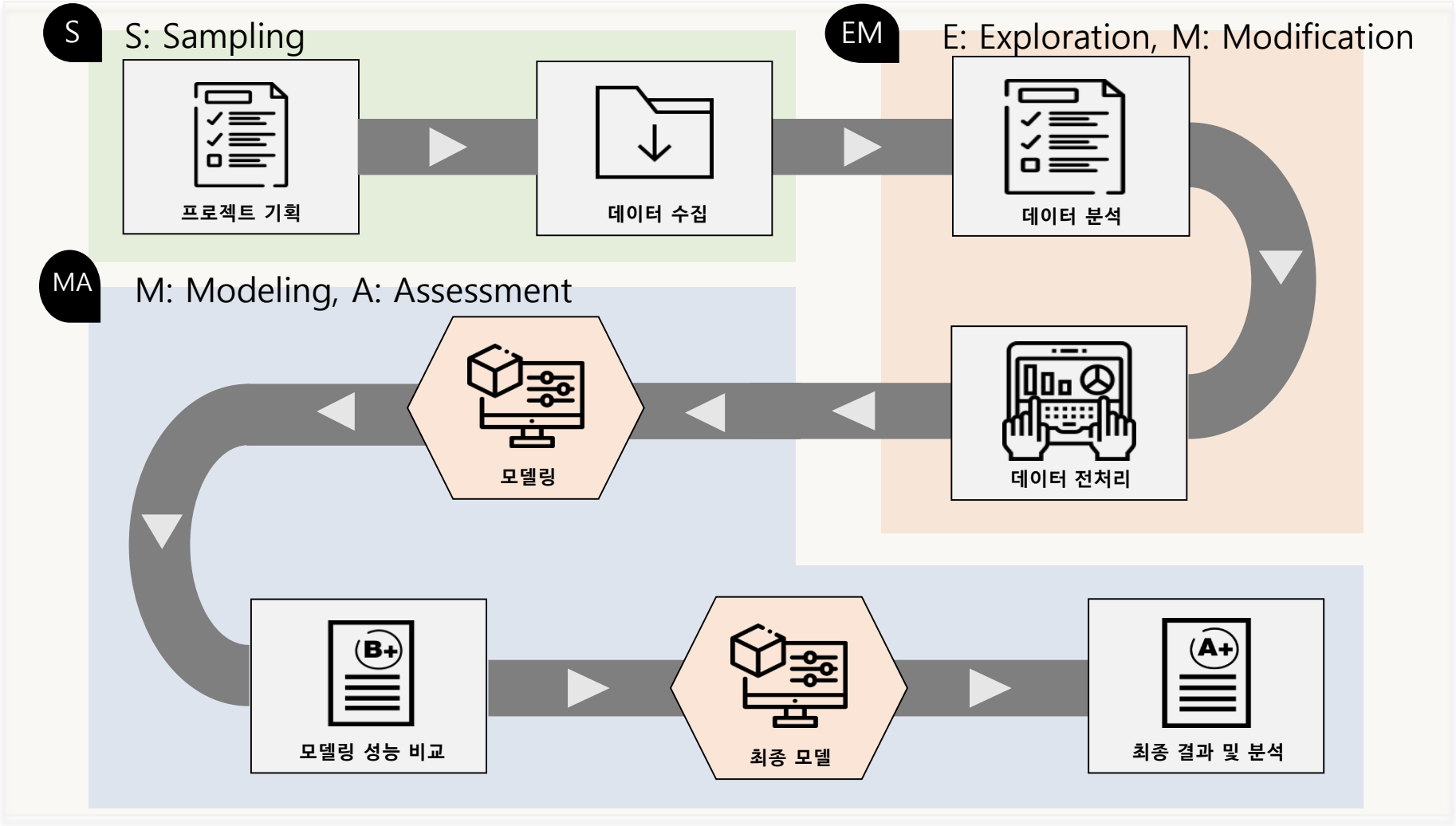
2) 서울특별시의 아파트 가격은 69주 연속 상승

3) 적정 아파트 가격 산출 모델링으로 투자 판단을 돕고 부작용을 방지하고자 함

02 Goal

: R2 Score 95% 이상의 서울아파트 거래가격 예측 모델 구현 및 분석

03 SEMMA


S
S: Sampling
EM
E: Exploration, M: Modification
MA
M: Modeling, A: Assessment
프로젝트 기획
데이터 수집
데이터 분석
데이터 전처리
모델링
모델링 성능 비교
최종 모델
최종 결과 및 분석
B+
A+

01 데이터 수집

❖ 데이터셋 7종

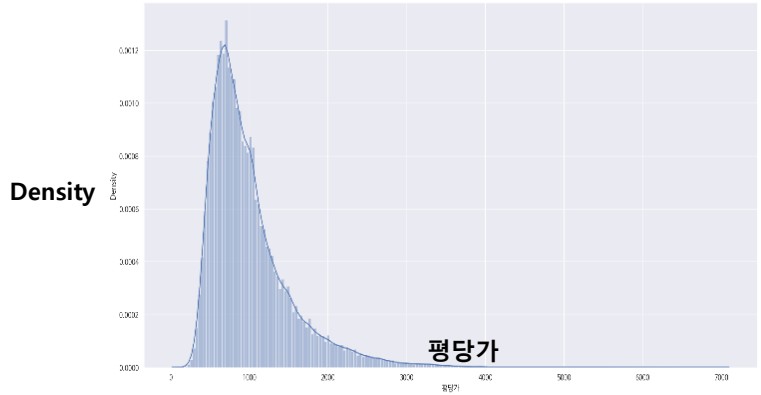
구분	수집 지표	출처
서울시 아파트 매매	지역코드, 법정동, 거래일, 아파트, 지번, 전용면적, 층, 건축년도, 거래금액, 거래년도	국토교통부 아파트매매 실거래자료
주가지수	KS11(코스피), KQ11(코스닥), DJI (다우존스), IXIC (나스닥), VIX (뉴욕주식시장), CSI300(상하이/심천 상위 300 주가지수), SSEC (상하이), DE30(독일), FCHI(프랑스) NG/GC/HG/CL(선물가지수)	E-나라지표
외국인증권투자		FinanceData.KR
인구	인구, 세대당 인구, 세대, 서울시 전입 인구	서울시 열린데이터 광장
금리	국고채 3년(평균) , 국고채 5년(평균) , 국고채 10년(평균) , 회사채 3년(평균) , CD 91물(평균)	E-나라지표
물가	소비자물가 , 농축수산물 , 공업제품 , 공공서비스 , 근원물가	E-나라지표
자동차등록		서울시 열린데이터 광장

- 2018년 1월부터 2021년 2월까지 월 기준으로 데이터 정렬
- 평당가(거래금액/전용면적), 거래 횟수, KRX(KQ11+KS11), K-means Clusting에 따른 군집화 라벨 컬럼 추가

02 데이터 분석

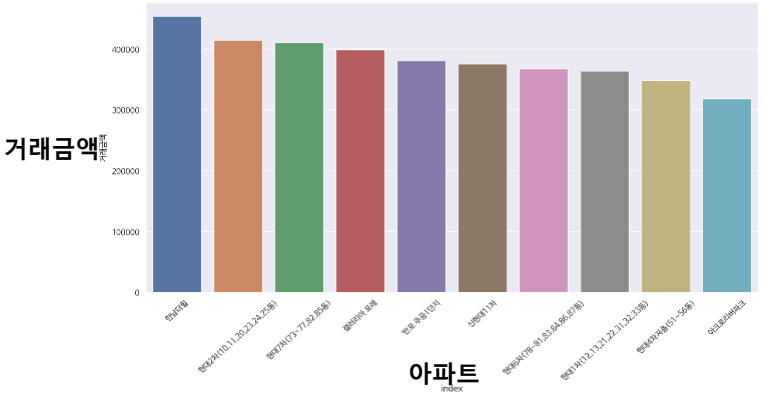
❖데이터 시각화

1. 평당가 분포도



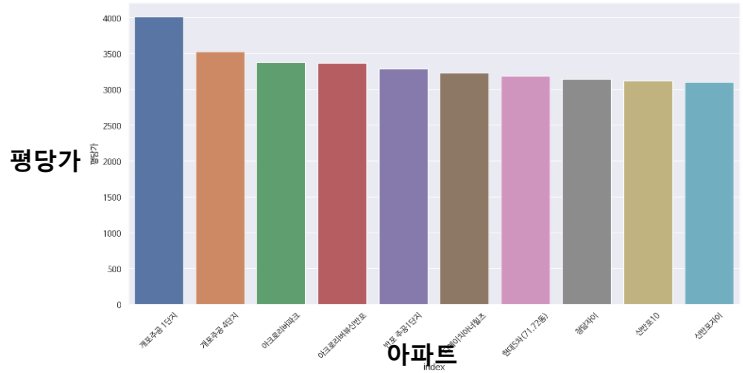
- 적은 수의 아파트가 높은 평당가로 거래되는 right_skewed 분포

2. 아파트별 거래 금액 순위



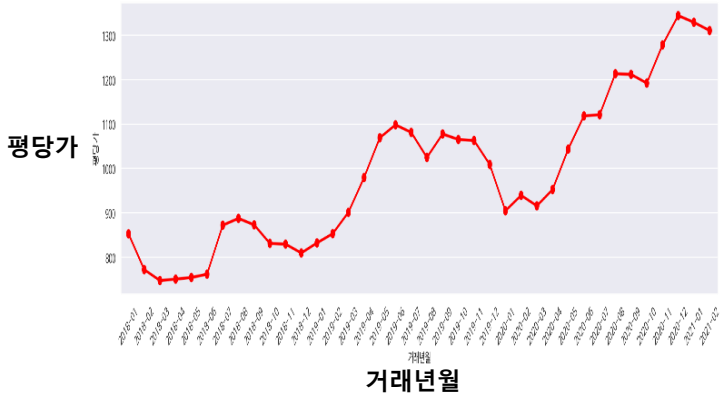
- 용산구 위치 한남더힐, 강남구 위치 현대아파트 등이 상위권

3. 아파트별 평당가 순위



- 강남구 개포 주공, 서초구 아크로리버파크 등이 높은 평당가를 기록

4. 월별 평당가

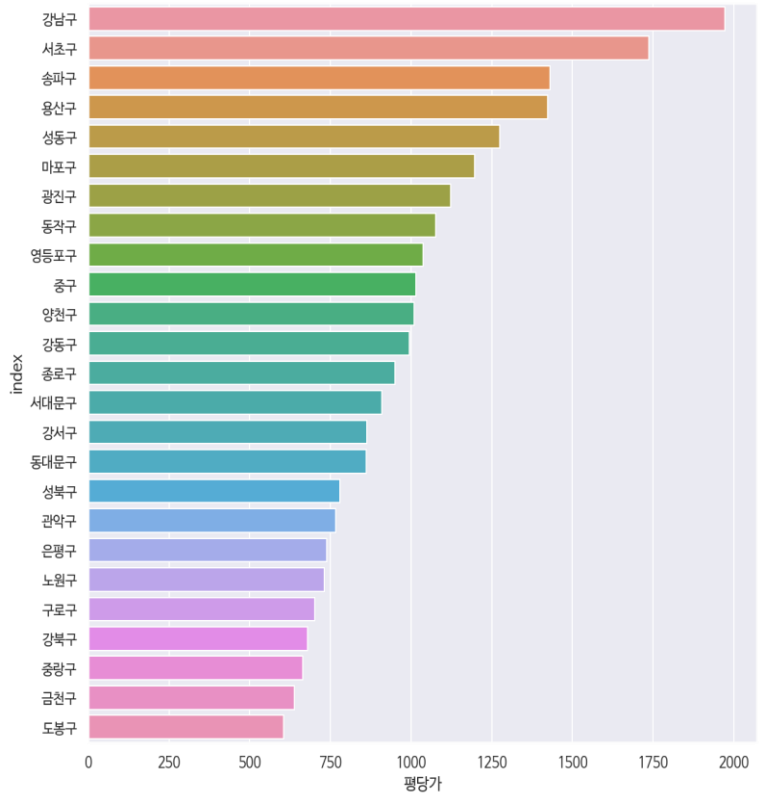


- 2018년부터 2021년 상반기까지 평당가는 전반적으로 상승하고 있음

02 데이터 분석

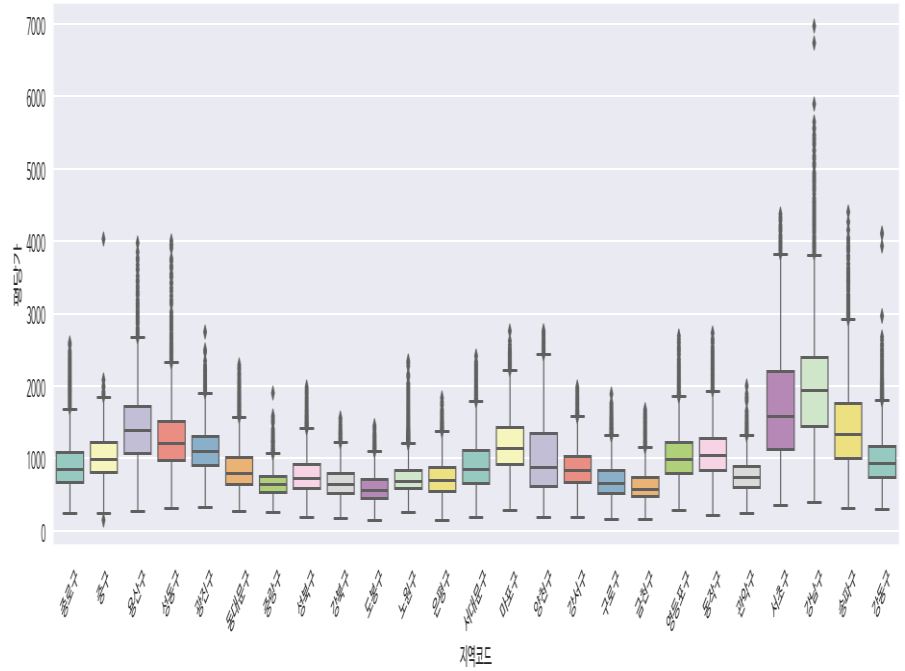
❖데이터 시각화

5. 구별 평당가 순위



- 구별 순위는 강남구, 서초구, 송파구, 용산구, 성동구 순으로 높음

6. 아파트별 평당가 순위

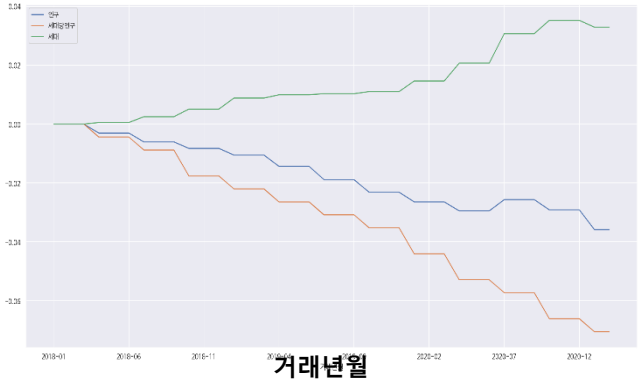


- 순위가 높은 강남구, 서초구, 송파구 등은 평당가의 이상치 또한 높음

02 데이터 분석

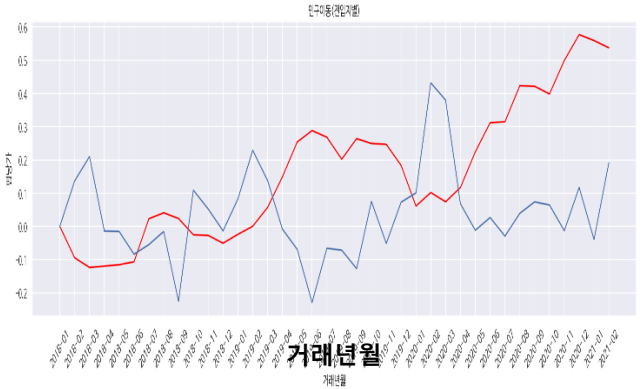
❖ 데이터 시각화

7. 인구(녹색선) vs 세대(푸른선) vs 세대별 인구(주황선)



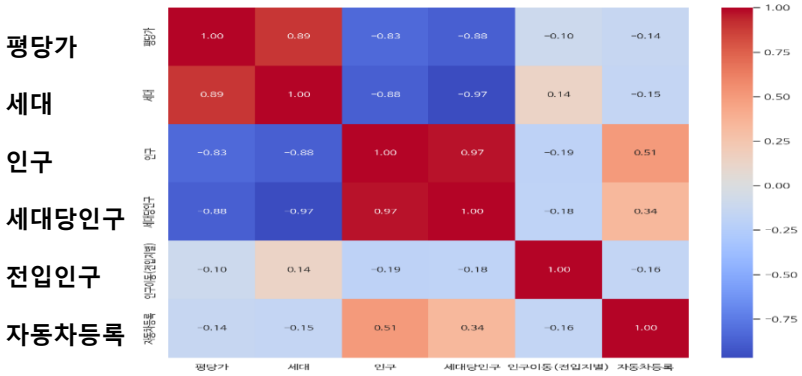
- 인구는 지속적으로 감소하는 반면 세대수는 증가하는 추세 (1인 가구 증가로 판단)

8. 전입인구(푸른선) vs 평당가(붉은선)



- 평당가가 상승할 시기 전입인구가 줄어드는 양상을 보임(가격 부담에 따른 영향)

9. 평당가와 인구/자동차 등록 간의 상관관계

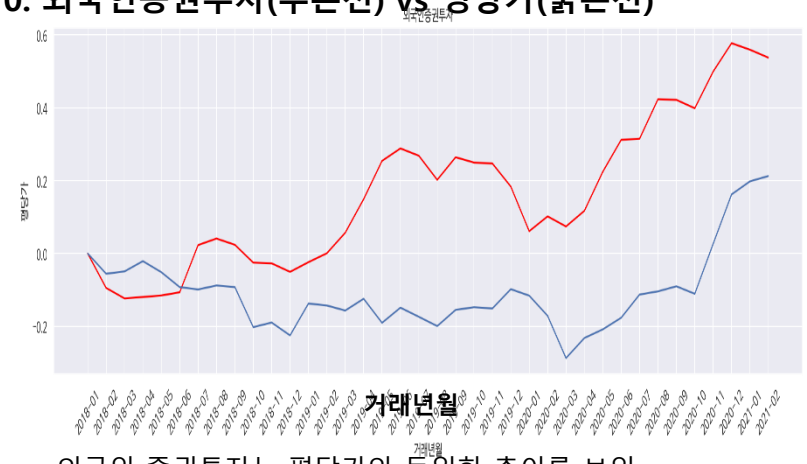


- 세대는 평당가와 양의 상관관계, 인구는 음의 상관관계를 보이며 전입인구와 자동차등록수는 높은 상관성을 보이지 않음

02 데이터 분석

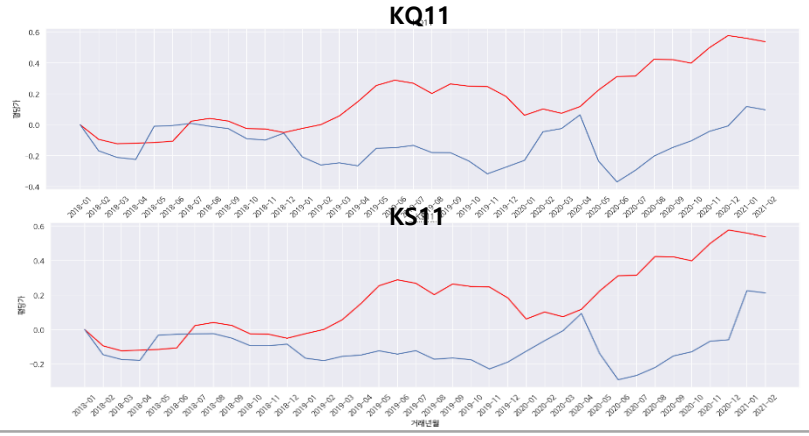
❖ 데이터 시각화

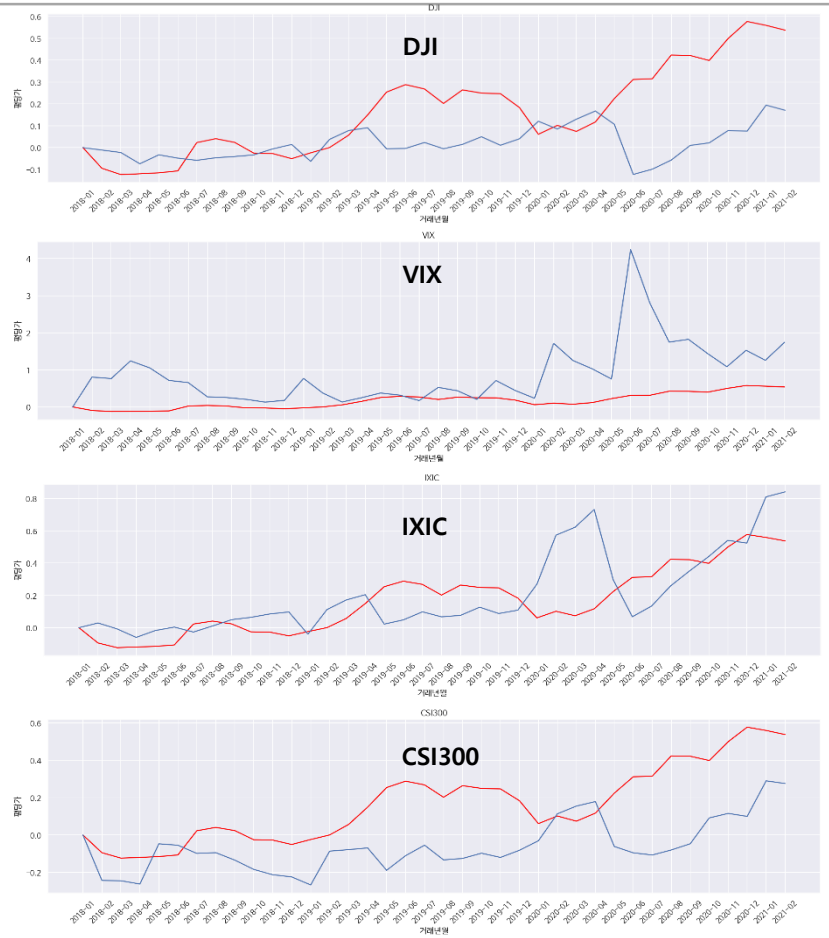
10. 외국인증권투자(푸른선) vs 평당가(붉은선)



- 외국인 증권투자는 평당가와 동일한 추이를 보임

11. 주가지수(푸른선) vs 평당가(붉은선)



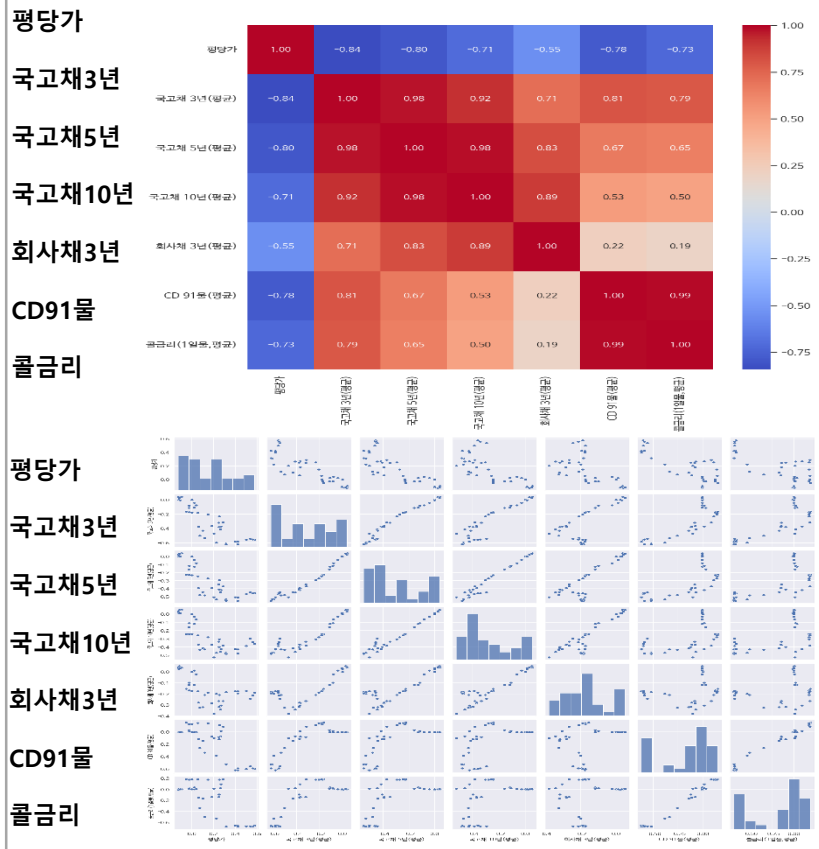


- 평당가와 주가지수는 상관계수가 높지 않으나 양의 상관관계를 보임

02 데이터 분석

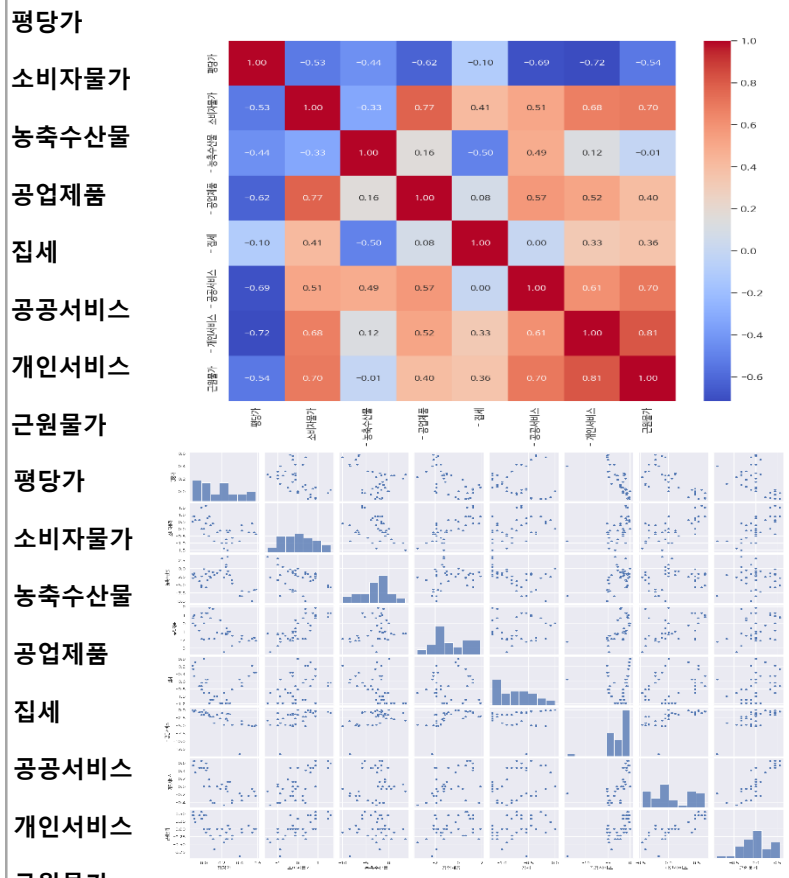
❖ 데이터 시각화

12. 금리



- 평당가와 금리는 높은 음의 상관관계를 보임(낮은 금리로 인해 증가한 유동성이 부동산 투자로 이어지는 것으로 보임)

13. 물가

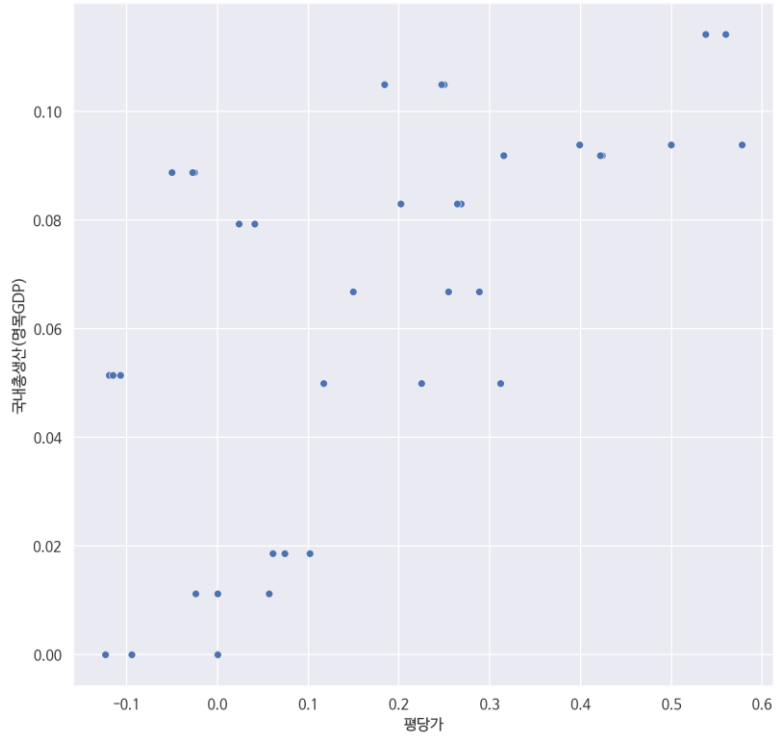


- 평당가와 물가는 음의 상관관계를 보임

02 데이터 분석

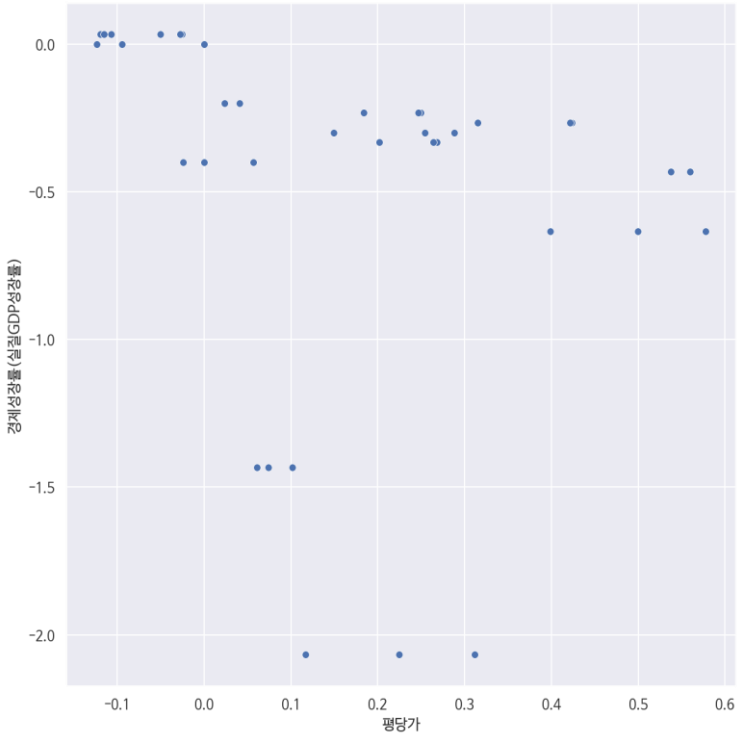
❖ 데이터 시각화

14. 명목 GDP



- 명목GDP는 평당가와 음의 상관관계

15. 실질 GDP

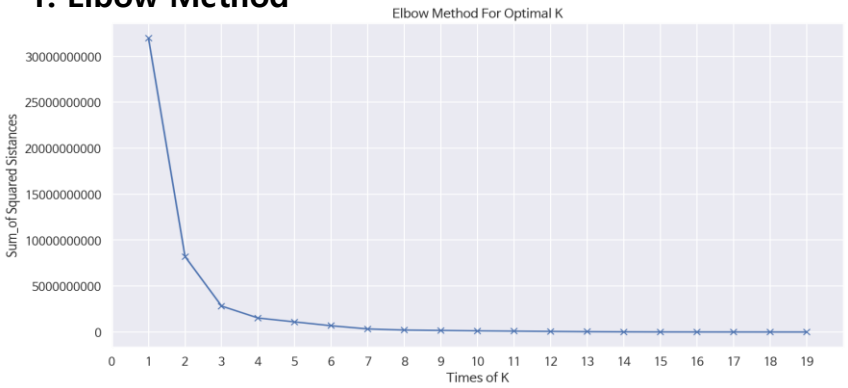


- 경제성장률(실질)GDP는 평당가와 상관관계가 뚜렷하지 않음

02 데이터 분석

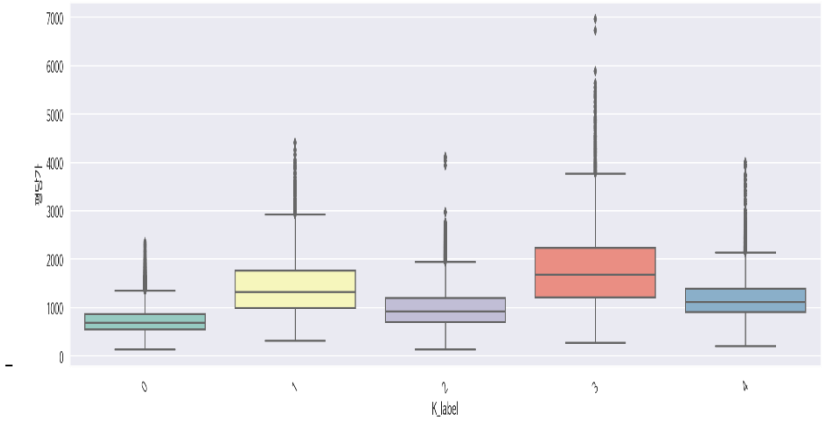
❖ K-means Clustering

1. Elbow Method

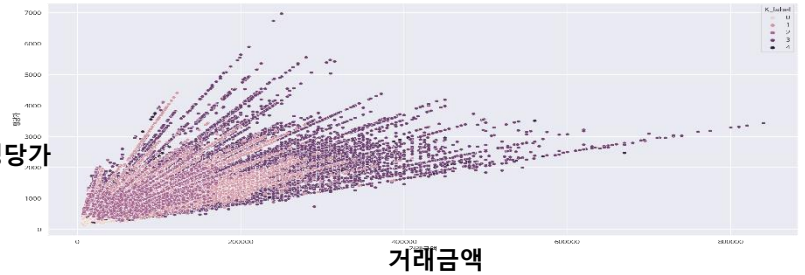


- 거래금액, 평당가를 기준으로 최적 K를 찾기 위한 Elbow Method 실시 : 최적 K = 5

2. K에 따른 평당가 분포 확인



3. 평당가와 건축금액 산점도(구분=K)

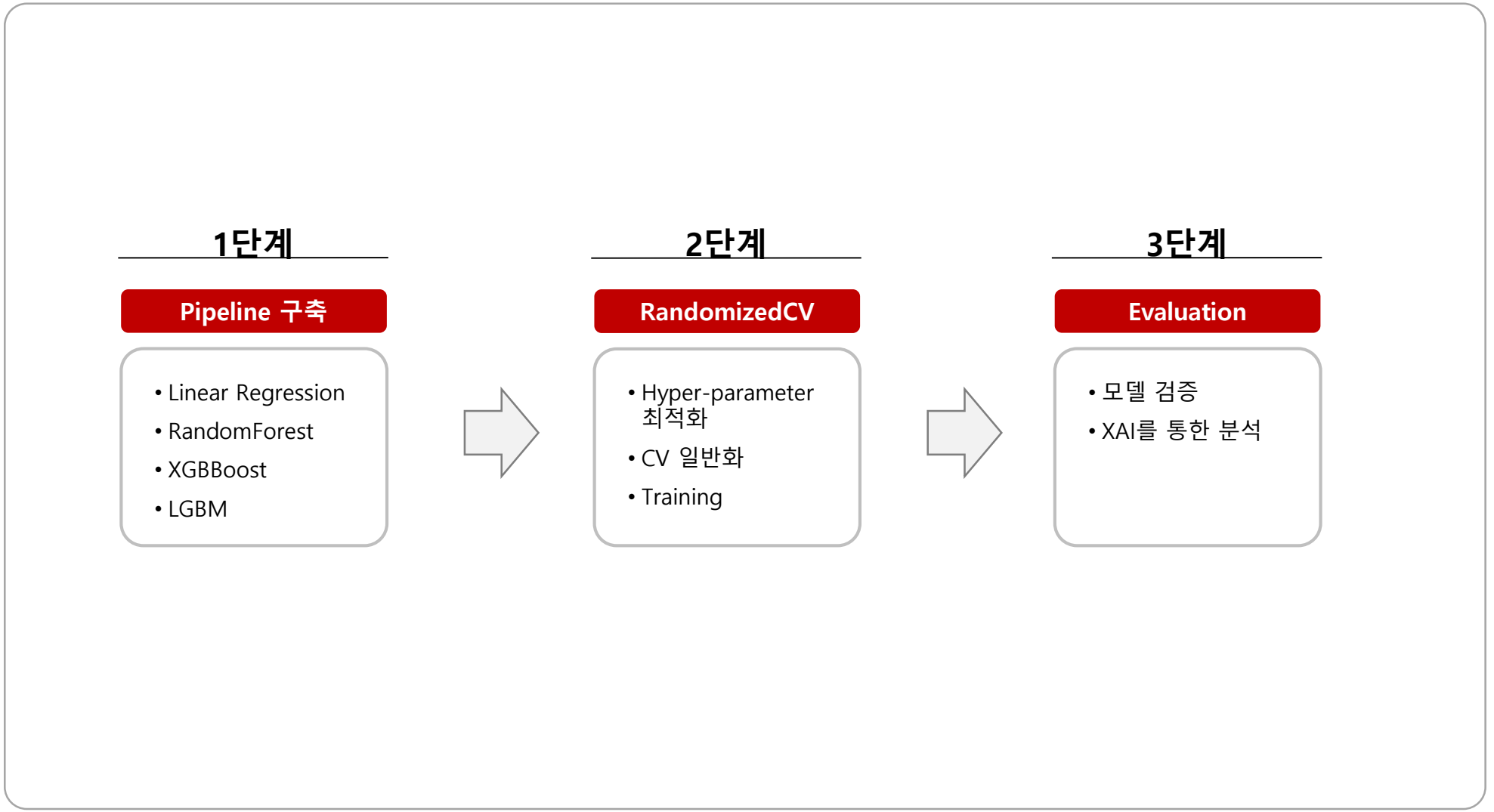


4. 건축년도에 따른 거래금액 및 평당가 산점도(구분=K)

- 재개발이 가능성이 높은 1990년대 이전 아파트에서 거래금액과 평당가가 높아짐

Modeling

01 Process



Modeling

02 Results

1. 성능비교

구분	Training		Validation		Test	
Score	R2	MAE	R2	MAE	R2	MAE
Linear Regression	0.55	24823	0.55	24679	0.56	29947
RandomForest	0.99	1811	0.97	4799	0.98	3799
XGBBoost	0.85	14154	0.84	14110	0.83	19214
LGBM	0.93	9736	0.93	9883	0.92	13852

- RandomForest 성능이 가장 높게 나왔지만 Hyper-parameter Tuning 후의 LGBM 모델 성능이 더 높을 것이라 판단, 최종 모델을 LGBM으로 선정 후 Randomized CV 진행

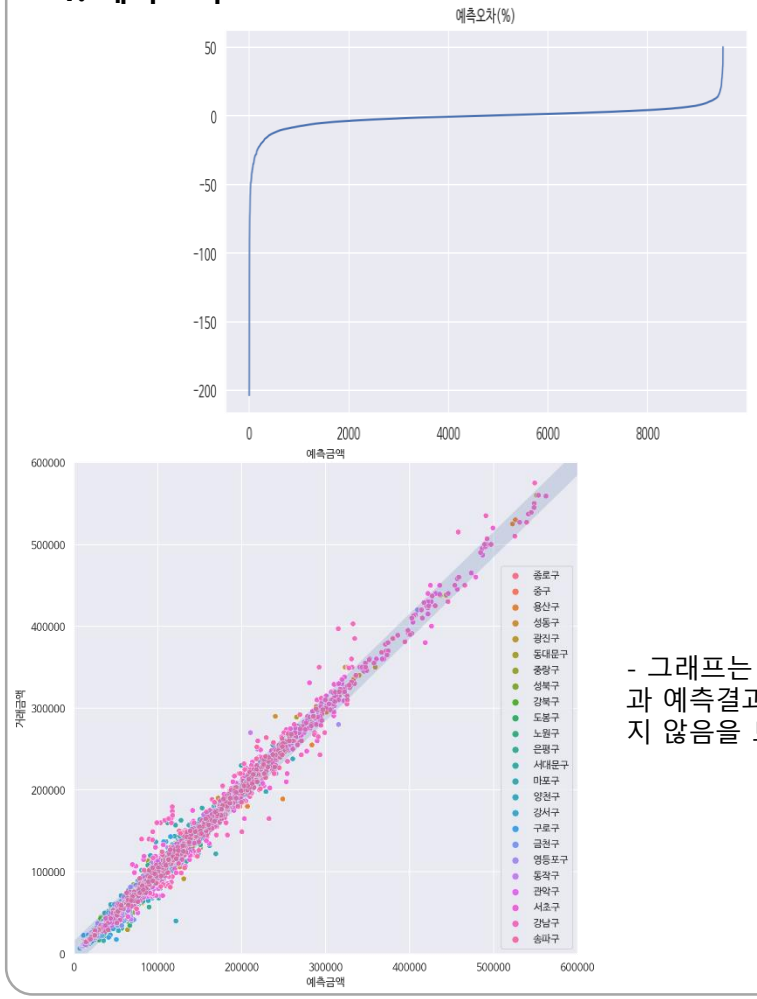
2. RandomizedCV Fine Tuning 후 LGBM 성능 평가

구분	Training		Validation		Test	
Score	R2	MAE	R2	MAE	R2	MAE
LGBM	0.99	2449	0.97	4163	0.99	3854

- Hyper-parameter Tuning 후 Test R2 0.99/ MAE 3854로 최고 성능을 기록

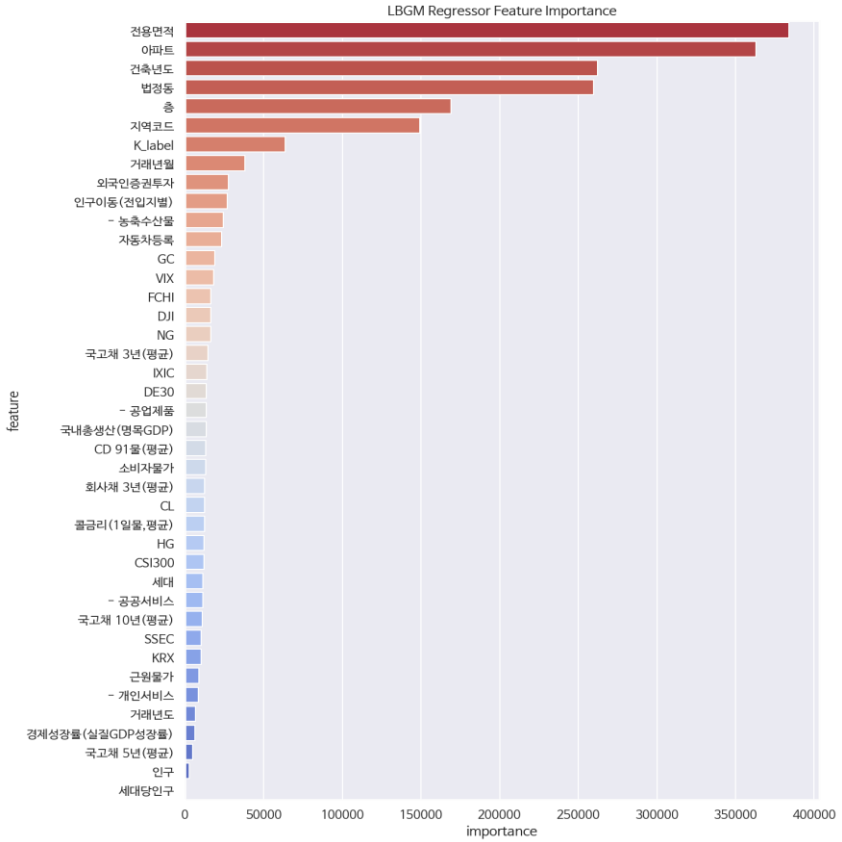
01 예측오차 및 특성 중요도

1. 예측오차



- 그래프는 실제 거래금액과 예측결과의 오차가 크지 않음을 보여줌

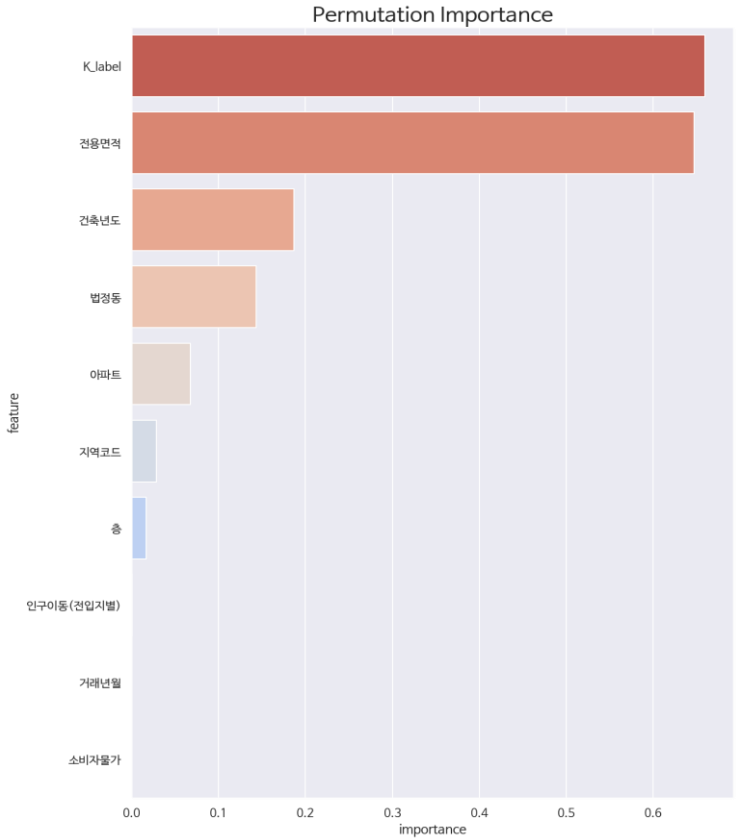
2. Feature Importance



- 특성중요도에 의하면 전용면적, 아파트, 건축년도, 법정동, 층이 거래가격 예측에 중요한 변수로 작용함

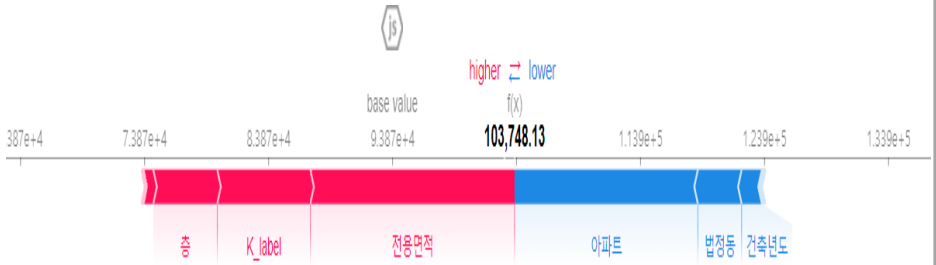
02 ELI5 / SHAP

3. Permutation Importance



- EL5를 사용한 순열중요도에 의하면 군집화된 라벨, 전용면적, 건축년도, 법정동, 아파트명, 지역코드, 층 순이 중요변수로 작용하였으며 나머지 변수는 예측에 큰 영향을 주지 못함

4. SHAP



- SHAP을 사용하여 예측 결과에 대한 분석을 할 수 있음
- 해당 row는 전용면적, 라벨 K, 층이 가격을 상승시키는 주요 변수였고 아파트, 법정동, 건축년도는 가격을 하락시키는 주요 변수로 작용하였음

Conclusion

01 결론

1) 서울특별시 아파트 거래가격 예측 모델(LGBM) Evaluation 결과 - R2 Score 0.99, MAE 3854

2) 가격에 영향을 미치는 주요 변수는 아파트 매매 데이터(전용면적, 아파트, 지역, 층 등)였으며 유동성 지표와 인구 등의 비중은 적음

3) ELI5, SHAP 라이브러리를 사용하여 예측 결과를 Global, Local 단위로 분석할 수 있음

02 기대효과

: ML 알고리즘을 통해 적정 부동산뿐 아니라 각종 금융상품의 가격을 오차율 1% 내로 예측할 수 있을 것으로 예상되며, LGBM과 같은 경량화된 모델로 앱 배포가 가능할 것으로 보임