2023-2 DAT 2기
캡스톤 프로젝트 발표

머신러닝을 이용한 상위 5개 브랜드 **아파트 실거래가 예측**

For
Studying &
Project

2023.12.07
Presentation Date

Business Proposal
Learning_Mate_Team. I 백형준(Leader), 고은서, 박서현, 윤희진, 김대현

2023-2 DAT 2기
캡스톤 프로젝트 발표

OutLines.

1. 서론
2. 데이터 소개
3. 변수 선택
4. 모델 실험
5. 결론

2023-2 DAT 2기
캡스톤 프로젝트 발표

1. 서론

Research Background

Topic

"머신러닝을 이용한 상위 5개 브랜드 아파트 실거래가 예측"

Point 1

팬대믹 전후의 부동산 시장 변화

현금 유동성이 증가

Point 2

부동산 실거래가 예측의 필요성

부동산 시장의 불확실성,
가격변동성이 심해짐

Point 3

서울시 각 '구'의 부동산 특성 반영

투자자 및 구매자에게
실질적 부동산 가치 정보 제공

2023-2 DAT 2기
캡스톤 프로젝트 발표

1. 서론

Research Purpose

Topic

"머신러닝을 이용한 상위 5개 브랜드 아파트 실거래가 예측"

Point 1

데이터 분석 & 모델링

아파트 거래 데이터와 경제 지표를 결합하여 서울시 각 '구'별 실거래가 예측 모델 개발

Point 2

ML 모델간의 기법 비교 & 평가

다양한 머신러닝 기법(Linear Regression, Random Forest, XGBoost, CatBoost 등)을 사용하여 정확도 높은 모델 선정.

Point 3

데이터 전처리 & 변수 선택

불필요한 데이터 정제, 클러스터링, 정규화, 타겟 및 원핫 인코딩 등을 통해 최적화된 데이터셋 구성.

Point 4

모델의 성능 평가

결정계수(R^2), 평균제곱오차(MSE), 평균제곱근 편차(RMSE)를 활용해 모델 성능 평가 및 결과 도출.

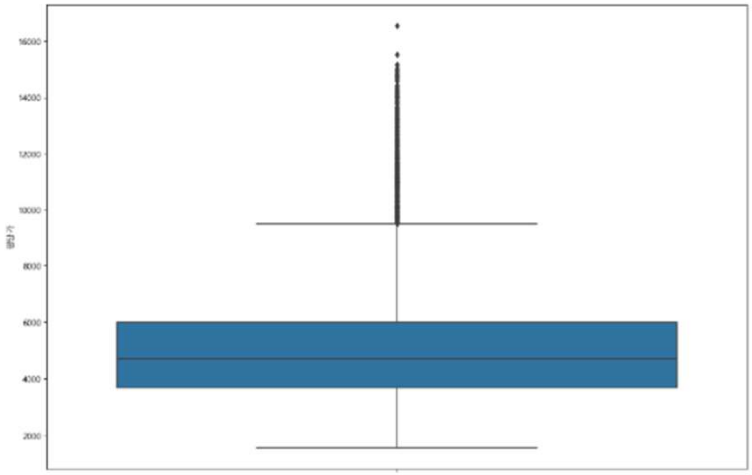
2023-2 DAT 2기
캡스톤 프로젝트 발표
2. 데이터 소개

전처리 과정

사용할 수 있는 데이터로 변환하기 위한 전처리 과정

External_data_total

Real_Estate_Top5_2020-2022



Point 1 매매가격 데이터, 경제 지표 데이터, 주가 데이터와 병합

Point 2 이상치 제거

2023-2 DAT 2기
캡스톤 프로젝트 발표
2. 데이터 소개

전처리 과정

사용할 수 있는 데이터로 변환하기 위한 전처리 과정

Point **불필요한 데이터 제거 및 결측치 확인**

```
data_1.drop(labels=['해제사유발생일', '등기신청일자', '거래유형', '중개사소재지', '번지', '본번', '부번'], axis=1, inplace=True)
data_1.dropna(axis=0, inplace=True)
data_1 = data_1.drop_duplicates(keep='first')
data_1.reset_index(drop=True, inplace=True)
```

```
시군구                                0
단지명                                0
전용면적(㎡)                           0
계약년월                               0
계약일                                0
거래금액(만원)                         0
층                                  0
건축년도                               0
도로명                                0
건설수주_건축(단위 백만원)                    0
건설수주_주택(단위 백만원)                    0
매매가격지수(아파트)                         0
경상수지(백만불)                           0
무역수지(백만불)                           0
대출금액(아파트)(억원)                       0
대출잔액(아파트)(억원)                       0
서울_신규_분양세대(단위: 세대)                    0
아파트 동(호)수_(단위: 호)                    0
생산자물가지수 총지수                         0
전규모(1인이상) 전체임금총액[원]                    0
소비자물가지수 총지수                         0
소비자물가지수 주택, 수도, 전기 및 연료              0
가계대출 (연리%)                          0
경기종합지수(2020=100)                     0
경제활동인구_ 실업률(단위: %)                    0
경제활동인구_고용률(단위: %)                    0
경제활동인구_취업자(단위: 천명)                   0
국제 주요국 주가지수(KOSPI)                  0
예금은행 대출금리(신규취급액 기준)_대출평균(연%)          0
예금은행 대출금리(잔액 기준)_총대출(연리%)            0
주택매매가격지수(KB)_서울                     0
소비자물가지수_총지수(가중치:1000?)                0
원화의 대미달러, 원화의 대위안/대엔 환율(원/달러(종가)_원    0
경기종합지수                               0
S&P 500_close                      0
S&P 500_change_rate                0
NASDAQ_close                       0
NASDAQ_change_rate                 0
DOW Jones_close                    0
DOW Jones_change_rate              0
KRX300_close                       0
KRX300_change_rate                 0
KOSPI_close                        0
KOSPI_change_rate                  0
KOSDAQ_close                       0
KOSDAQ_change_rate                 0
dtype: int64
```

전처리 과정

문자열 데이터의 수치화 & 정규화 진행

원핫인코딩 / 타겟인코딩

타겟인코딩된 데이터 < 매칭구 >

강남구	강동구	강북구	강서구	관악구	광진구	구로구	금천구	노원구	도봉구	동대문구	동작구	마포구	서대문구	서초구	성동구	성북구	송파구	양천구	영등포구	용산구	종로구	중구	중랑구
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

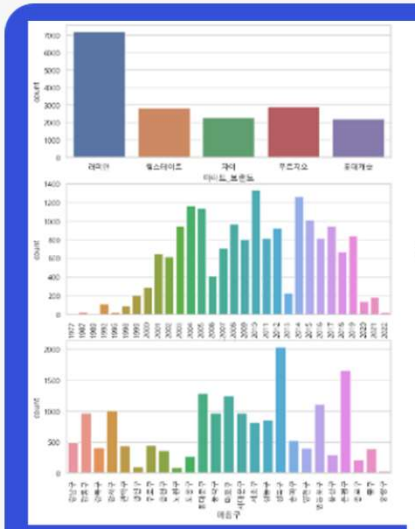
정규화

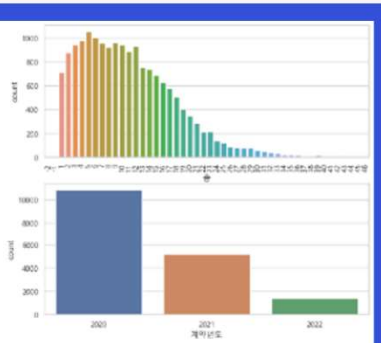
Sklearn 패키지, StandardScaler 라이브러리

2023-2 DAT 2기
캡스톤 프로젝트 발표
2. 데이터 소개

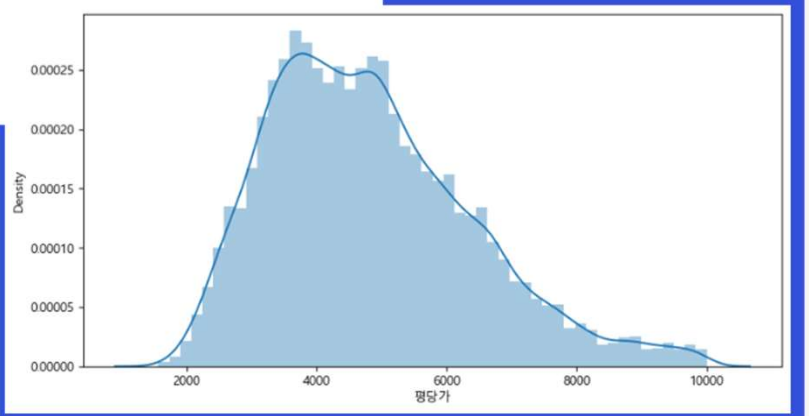
데이터 EDA

데이터 시각화를 통해 데이터 분포 및 특징을 확인하여
Countplot을 통해 범주형 데이터의 분포 확인



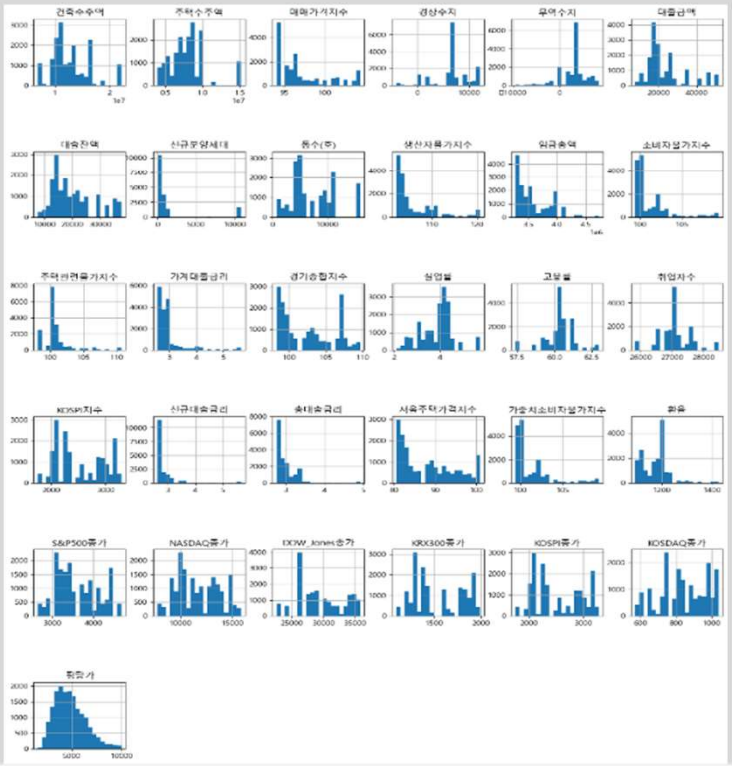


아파트 브랜드,
층수,
건축년도,
계약 연도,
매칭구
(좌측상단 부터)



평당가에 대한
Displot
(연속형 데이터)

HistoGram

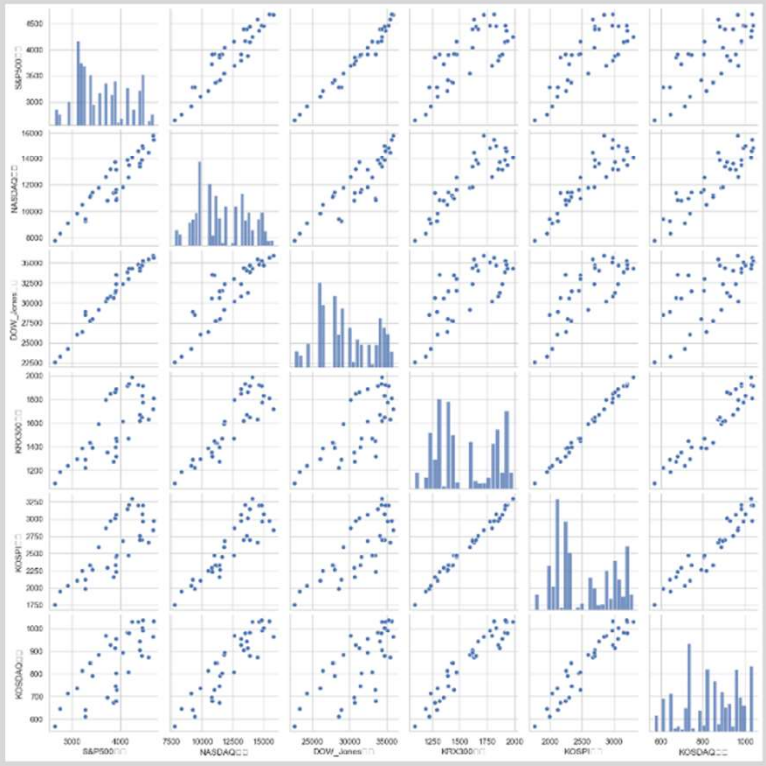


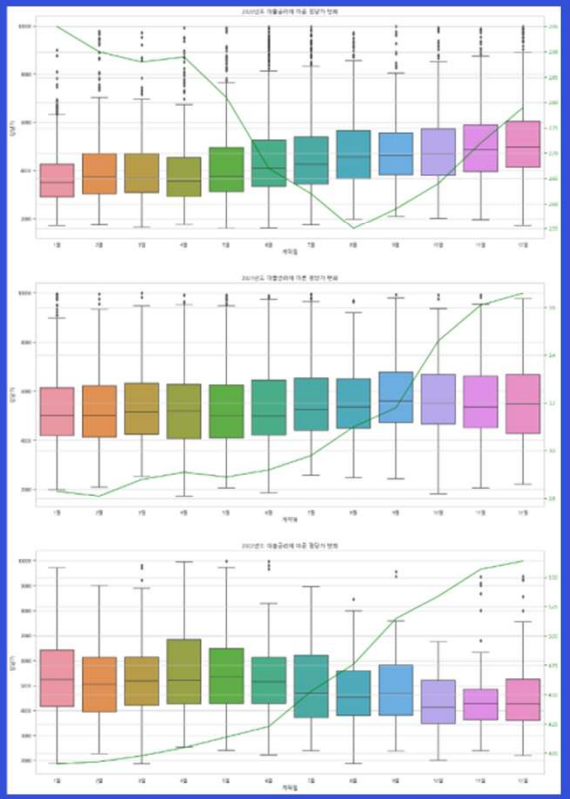
2023-2 DAT 2기
캡스톤 프로젝트 발표
2. 데이터 소개

데이터 EDA

데이터 시각화를 통해 데이터 분포 및 특징을 확인하여
Countplot을 통해 범주형 데이터의 분포 확인

Pairplot - 종속 변수간의 상관관계 확인





계약 월과
평당가에
대한
Box plot

2023-2 DAT 2기
캡스톤 프로젝트 발표
3. 변수 선택

변수선택_모든데이터

시군구와 같은 범주형 변수가 속하기 때문에 범주형 변수 처리를 해줘야 한다

전처리 | 상관관계 0.9 이상 변수 제거, 다중공선성 변수 제거

기본다중회귀분석

Parametric Coefficients (모수계수)	Estimate (추정값)	Std.Error (표준오차)	t value (t 값)	Pr(>|t|) (p-value)
(Intercept)	0.476	0.012	39.019	<2e-16
아파트_브랜드_롯데 캐슬	-0.468	0.021	-22.677	<2e-16
아파트_브랜드_자이	-0.220	0.021	-10.576	<2e-16
아파트_브랜드_푸르지오	-0.538	0.019	-29.005	<2e-16
아파트_브랜드_힐스테이트	-0.436	0.019	-23.021	<2e-16
cluster.1	-0.541	0.015	-36.866	<2e-16
cluster.3	0.340	0.024	14.129	<2e-16
cluster.4	-0.685	0.023	-29.351	<2e-16
cluster.5	-0.977	0.040	-24.155	<2e-16
층	0.058	0.006	8.990	<2e-16
건축년도	0.372	0.007	56.262	<2e-16
임금총액	0.063	0.006	9.849	<2e-16
나스닥변동률	-0.173	0.011	-15.613	<2e-16
다우존스변동률	0.158	0.011	14.276	<2e-16

Residual standard error: 0.8293 on 17385 degrees of freedom	Multiple R-squared: 0.3128 Adjusted R-squared: 0.3123	F-statistic: 608.8 on 13 and 17385 DF, p-value: < 2.2e-16

Allsubset Regression

개수	롯데 캐슬	자이	푸르지오	힐스테이트	cluster1	cluster3	cluster4	cluster5	층	건축년도	임금총액	나스닥변동률	다우존스변동률
1										*			
2					*					*			
3					*			*		*			
4					*		*	*		*			
5			*		*		*	*		*			
6			*	*	*		*	*		*			
7	*		*	*	*		*	*		*			
8	*		*	*	*	*	*	*		*			
9	*		*	*	*	*	*	*		*	*		
10	*		*	*	*	*	*	*		*		*	*
11	*	*	*	*	*	*	*	*		*		*	*
12	*	*	*	*	*	*	*	*		*	*	*	*
13	*	*	*	*	*	*	*	*	*	*	*	*	*

1	2	3	4	5	6	7	8	9	10	11	12	13
0.128	0.179	0.207	0.236	0.252	0.267	0.284	0.291	0.295	0.301	0.305	0.309	0.312

2023-2 DAT 2기
캡스톤 프로젝트 발표
3. 변수 선택

변수선택_모든데이터

시군구와 같은 범주형 변수가 속하기 때문에 범주형 변수 처리를 해줘야 한다

Stepwise Selection

Parametric Coefficients (모수계수)	Estimate (추정값)	Std.Error (표준오차)	t value (t 값)	Pr(>|t|) (p-value)
(Intercept)	0.215	0.009	23.878	< 2e-16
아파트.브랜드_힐스테이트	-0.213	0.017	-12.429	< 2e-16
cluster.1	-0.514	0.014	-35.767	< 2e-16
cluster.3	0.343	0.024	14.514	< 2e-16
cluster.4	-0.540	0.022	-23.981	< 2e-16
cluster.5	-0.162	0.006	-25.748	< 2e-16
층	0.054	0.006	8.580	< 2e-16
건축년도	0.357	0.006	56.211	< 2e-16
주택수주액	0.049	0.008	6.092	1.14E-09
경상수지	0.083	0.014	5.963	2.52E-09
무역수지	-0.140	0.016	-8.540	< 2e-16
신규분양세개	0.079	0.010	7.748	9.87E-15
동수(호)	-0.152	0.012	-12.870	< 2e-16
임금총액	0.047	0.008	5.638	1.75E-08
고용률	0.050	0.010	5.140	2.77E-07
총대출금리	-0.206	0.012	-16.709	< 2e-16
환율	-0.054	0.014	-3.836	0.000126
나스닥변동률	-0.075	0.014	-5.472	4.51E-08
다우존스변동률	0.035	0.013	2.753	0.005913

Residual standard error: 0.8156 on 17380 degrees of freedom	Multiple R-squared: 0.3355 Adjusted R-squared: 0.3348	F-statistic: 487.4 on 18 and 17380 DF, p-value: < 2.2e-16

Forward Selection

Parametric Coefficients (모수계수)	Estimate (추정값)	Std.Error (표준오차)	t value (t 값)	Pr(>|t|) (p-value)
(Intercept)	0.222	0.010	21.834	<2e-16
아파트.브랜드_자이	0.006	0.020	0.309	0.757
아파트.브랜드_힐스테이트	-0.219	0.018	-11.959	<2e-16
cluster.1	-0.530	0.015	-35.060	<2e-16
cluster.3	0.336	0.025	13.546	<2e-16
cluster.4	-0.559	0.024	-23.581	<2e-16
cluster.5	-0.167	0.007	-25.362	<2e-16
층	0.055	0.007	8.367	<2e-16
건축년도	0.366	0.007	53.990	<2e-16
임금총액	0.061	0.007	9.249	<2e-16
나스닥변동률	-0.174	0.011	-15.236	<2e-16
다우존스변동률	0.161	0.011	14.109	<2e-16

Residual standard error: 0.8548 on 17387 degrees of freedom	Multiple R-squared: 0.2697 Adjusted R-squared: 0.2693	F-statistic: 583.8 on 11 and 17387 DF, p-value: < 2.2e-16

2023-2 DAT 2기
캡스톤 프로젝트 발표
3. 변수 선택

변수선택_범주형 변수 제거

시군구와 같은 범주형 변수가 속하기 때문에 범주형 변수 처리를 해줘야 한다

전처리 | 상관관계 0.9 이상 변수 제거, 다중공산성 변수 제거

기본다중회귀분석

Parametric Coefficients (모수계수)	Estimate (추정값)	Std.Error (표준오차)	t value (t 값)	Pr(>|t|) (p-value)
(Intercept)	-3.73E-15	7.08E-03	0	1
아파트.브랜드	1.79E-01	7.16E-03	25.046	< 2e-16
층	5.45E-02	7.15E-03	7.625	2.61E-14
건축년도	3.38E-01	7.16E-03	47.159	< 2e-16
주택수주액	7.18E-02	7.76E-03	9.246	< 2e-16
임금총액	4.98E-02	7.30E-03	6.824	9.20E-12
나스닥변동률	-1.95E-01	1.26E-02	-15.502	< 2e-16
다우존스변동률	1.49E-01	1.25E-02	11.904	< 2e-16
클러스터링_구	3.17E-01	7.14E-03	44.416	< 2e-16

Residual standard error: 0.8351 on 13910 degrees of freedom	Multiple R-squared: 0.3029 Adjusted R-squared: 0.3025	F-statistic: 755.6 on 8 and 13910 DF, p-value: < 2.2e-16

Allsubset Regression

개수	아파트.브랜드	층	건축년도	주택수주액	임금총액	나스닥변동률	다우존스변동률	클러스터링_구
1			*					
2			*					*
3	*		*					*
4	*		*		*			*
5	*		*			*	*	*
6	*		*	*		*	*	*
7	*	*	*	*		*	*	*
8	*	*	*	*	*	*	*	*

1	2	3	4	5	6	7	8
0.129	0.247	0.281	0.285	0.292	0.297	0.300	0.303

2023-2 DAT 2기
캡스톤 프로젝트 발표
3. 변수 선택

변수선택_범주형 변수 제거

시군구와 같은 범주형 변수가 속하기 때문에 범주형 변수 처리를 해줘야 한다

Stepwise Selection

Parametric Coefficients (모수계수)	Estimate (추정값)	Std.Error (표준오차)	t value (t 값)	Pr(>|t|) (p-value)
(Intercept)	-1.14E-14	6.72E-03	0	1
층	5.46E-02	6.79E-03	8.034	1.02E-15
건축년도	3.28E-01	6.82E-03	48.144	< 2e-16
주택수주액	8.73E-02	8.45E-03	10.341	< 2e-16
대출금액	-2.69E-02	9.08E-03	-2.957	0.003111
동수(호)	-4.55E-02	9.48E-03	-4.801	1.60E-06
임금총액	1.76E-02	9.18E-03	1.915	0.055551
실업률	-8.63E-02	1.40E-02	-6.17	7.03E-10
고용률	-2.29E-02	1.61E-02	-1.419	0.155922
총대출금리	-1.34E-01	9.14E-03	-14.675	< 2e-16
나스닥변동률	-2.69E-02	8.16E-03	-3.298	0.000976
계약연도	1.75E-01	1.09E-02	16.055	< 2e-16
클러스터링_구	3.12E-01	6.78E-03	46.034	< 2e-16
아파트.브랜드	1.76E-01	6.80E-03	25.868	< 2e-16

Residual standard error: 0.7925 on 13905 degrees of freedom	Multiple R-squared: 0.3726 Adjusted R-squared: 0.372	F-statistic: 635.2 on 13 and 13905 DF, p-value: < 2.2e-16

Forward Selection

Parametric Coefficients (모수계수)	Estimate (추정값)	Std.Error (표준오차)	t value (t 값)	Pr(>|t|) (p-value)
(Intercept)	-8.50E-15	7.08E-03	0	1
층	5.45E-02	7.15E-03	7.625	2.61E-14
건축년도	3.38E-01	7.16E-03	47.159	< 2e-16
주택수주액	7.18E-02	7.76E-03	9.246	< 2e-16
임금총액	4.98E-02	7.30E-03	6.824	9.20E-12
나스닥변동률	-1.95E-01	1.26E-02	-15.502	< 2e-16
다우존스변동률	1.49E-01	1.25E-02	11.904	< 2e-16
클러스터링_구	3.17E-01	7.14E-03	44.416	< 2e-16
아파트.브랜드	1.79E-01	7.16E-03	25.046	< 2e-16

Residual standard error: 0.8351 on 13910 degrees of freedom	Multiple R-squared: 0.3029 Adjusted R-squared: 0.3025	F-statistic: 755.6 on 8 and 13910 D F, p-value: < 2.2e-16

2023-2 DAT 2기
캡스톤 프로젝트 발표
3. 변수 선택

Stepwise & Forward 변수 추출

변수 추출 방법	최종 변수
Re-step	['아파트 브랜드', **'층', '건축년도'**, '건축수주액', **'대출금액', '동수(호)', '임금총액'**, '실업률', '고용률', '총대출금리', '나스닥변동률' ,'계약연도','구']
Re-forward	['아파트 브랜드', **'층', '건축년도'**, '건축수주액', **'임금총액'**,'나스닥변동률', '다우존스변동률' ,'계약연도','구']
Ex-step	['아파트 브랜드', **'층', '건축년도'**, '건축수주액', '매매가격지수','무역수지', **'대출금액'**, '신규분양세대', **'동수(호)', '임금총액'**,'고용률', '신규대출금리' ,'계약연도','구']
Ex-forward	['아파트 브랜드', **'층', '건축년도'**, '주택수주액', '경상수지', **'대출금액','동수(호)', '임금총액'**,'S&P500 변동률','계약연도','구']

Point

Re가 붙은 것은 모든 데이터를 기반으로 분석한 결과.
ex가 붙은 것은 범주형 변수를 제외하고 분석한 결과에 추후, 범주형변수를 추가해 모델에 반영한 것.

2023-2 DAT 2기
캡스톤 프로젝트 발표
3. 변수 선택

타겟인코딩 / One-hot encoding 값 비교

Variable

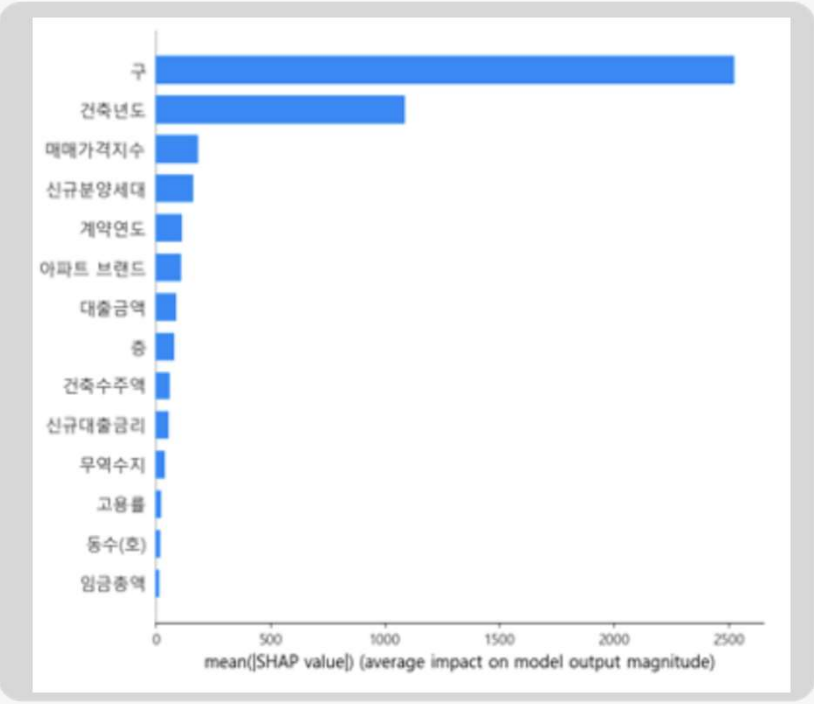
'아파트 브랜드', '층', '건축년도', '건축수주액', '매매가격지수', '무역수지', '대출금액', '신규분양세대', '동수(호)', '임금총액', '고용률', '신규대출금리', '계약연도', '구'

모델명	Target Encoding RMSE	Target Encoding R^2	One-hot encoding RMSE	One-hot encoding R^2
LinearRegression	955.0854	63%	892.6104	67%
Lasso	955.0245	63%	891.9576	68%
Ridge	955.0871	63%	892.4305	67%
XGBRegressor	751.1644	77%	802.2604	74%
LGBMRegressor	751.0502	77%	**761.0450**	**76%**
RandcomForestRegressor	769.9768	76%	907.1259	66%
CatBoostRegressor	**731.2412**	**78%**	799.5713	74%

2023-2 DAT 2기
캡스톤 프로젝트 발표
4. 모델 부분

Target Encoding Feature Importance

- 최소 RMSE 값: 731
- Target 변수에 크게 영향을 미치는 요인은 구, 건축년도, 매개가격지수 순서
- '구', '건축년도'의 변수가 평당가의 높은 영향을 미치는 것을 알 수 있음.
- 단, 변수 안의 어떤 값이 영향을 미쳤는지 해석이 불가능



Point

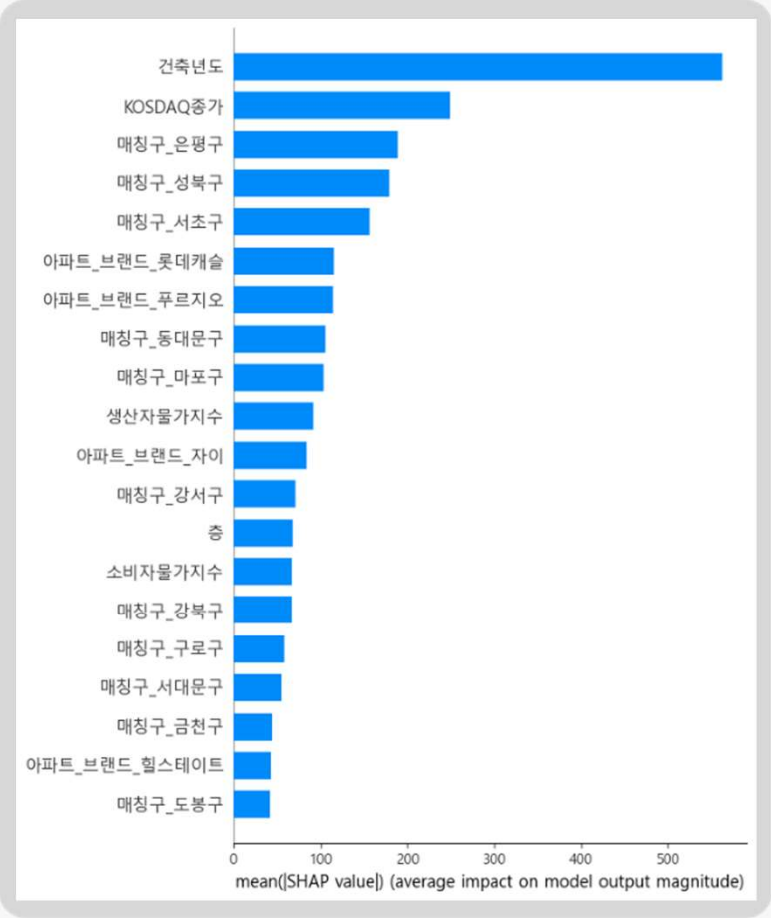
그러나, Target Encoding은 어떤 특정 변수가 높은 확률로 타겟 값에 영향을 미치는지 알 수 없기 때문에, 특정 변수 안에 어떤 값이 영향을 미쳤는지 해석할 수 없음.

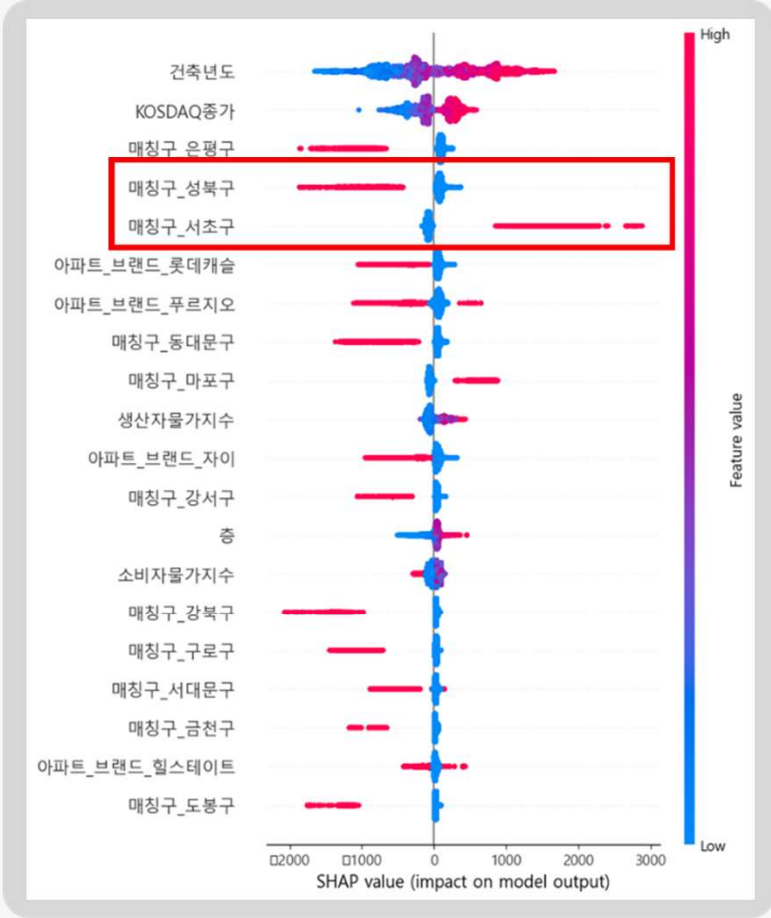
⬇

이를 보완하고자 One-Hot Encoding 기법을 함께 적용

2023-2 DAT 2기
캡스톤 프로젝트 발표
4. 모델 부분

One-Hot Encoding Feature Importance





2023-2 DAT 2기
캡스톤 프로젝트 발표

5. 결론

Conclusion

Point 1	건축년도는 '평당가'에 일관된 영향을 미치지 않으며, SHAP 값이 넓게 분포함을 보여줍니다.
Point 2	KOSDAQ 종가의 높은 SHAP 값은 평당가에 긍정적인 영향을 주며, 평당가격과 KOSDAQ 지수 간의 상관관계를 나타냅니다.
Point 3	강서구, 강북구 등의 지역은 평당가가 낮을 가능성을 나타내는 반면, 서초구와 마포구는 평당가가 높을 가능성을 보여줍니다.
Point 4	아파트 브랜드와 관련하여, '롯데캐슬', '푸르지오', '자이'에 비해 '힐스테이트'가 모델 예측에 더 큰 기여를 하는 것으로 나타났습니다.

Main Point	모델의 기여도가 높다는 것은 특정 특성이 평당가격에 중대한 영향을 미친다는 의미이며, 이는 특성의 값이 크면 평당가격도 높아질 것임을 의미합니다.

2023-2 DAT 2기
캡스톤 프로젝트 발표
5. 결론

제안 및 보완사항

Point 1	데이터의 범위와 다양성 부족
Point 2	월별 데이터의 한계
Point 3	적은 양의 데이터
Point 4	Feature 선정을 위한 도메인 지식 부족

2023-2 DAT 2기
캡스톤 프로젝트 발표

Q & A