

# 부동산 허위 매물



# INDEX

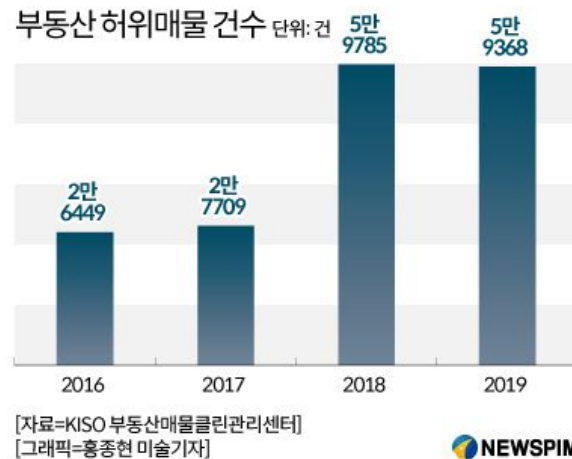
1. 분석 개요
2. EDA 및 데이터 전처리
3. 모델링 및 모델 학습
4. 모델 평가
5. 분석 결론

# 01 | 분석 개요

## 분석 배경

### ● 개요

- 부동산 시장은 우리의 삶과 밀접하게 연결되어 있지만 허위매물 문제는 여전히 많은 소비자들에게 불편과 손해를 초래하고 있음
- **지속되는 부동산 허위 매물 문제**를 해결하기 위해 부동산 **허위 매물 분류 모델**을 개발함
- 허위매물을 효과적으로 탐지하고 제거하는 기술은 부동산 시장의 투명성을 높이고, 소비자와 중개업자 모두에게 신뢰할 수 있는 거래 환경을 제공하는 데 기여할 수 있음



# 01 | 분석 개요

## 데이터 소개

- 변수 소개

- 아래의 변수들로 분류 모델 개발

- 변수 :

- 매물확인방식, 보증금, 월세, 전용면적, 해당층, 총층, 방향, 방수, 욕실수, 주차가능여부, 총주차대수, 관리비, 중개사무소, 제공플랫폼, 게재일, 허위매물여부

- 변수 타입 정리 :

- 수치형 변수 : 보증금, 월세, 전용면적, 해당층, 총층, 방수, 욕실수, 총주차대수, 관리비
- 범주형 변수 : 매물확인방식, 방향, 주차가능여부, 중개사무소, 제공플랫폼
- 날짜형 변수 : 게재일
- 타겟 변수 : 허위매물여부

# 01 | 분석 개요

## 데이터 소개

- train data

- 총 2452 개의 데이터

ID	매물확인방식	보증금	월세	전용면적	해당층	총층	방향	방수	욕실수	주차가능여부	총주차대수	관리비	중개사무소	제공플랫폼	게재일	허위매물여부
TRAIN_0000	현장확인	402500000.0	470000			15.0	서향	1.0	1.0	가능	40.0	96	t93Nt6I2I0	B플랫폼	2024-10-09	0
TRAIN_0001	현장확인	170500000.0	200000		3.0	4.0	남동향	2.0	1.0	불가능		0	q39iV5J4E6	D플랫폼	2024-12-26	0
TRAIN_0002	전화확인	114000000.0	380000		2.0	3.0	동향	1.0	1.0	불가능		0	b03oE4G3F6	A플랫폼	2024-11-28	0
TRAIN_0003	현장확인	163500000.0	30000	36.3	3.0	9.0	남동향	2.0	1.0	가능	13.0	10	G52Iz8V2B9	A플랫폼	2024-11-26	0
TRAIN_0004	현장확인	346000000.0	530000		3.0	3.0	동향	2.0	1.0	불가능		0	N45gM0M7R0	B플랫폼	2024-06-25	1
TRAIN_0005	전화확인	153000000.0	530000	29.5		3.0	남향	2.0	1.0	가능	1.0	0	Q42YF3Y0I2	A플랫폼	2024-09-12	0
TRAIN_0006	현장확인	348500000.0	400000		2.0	3.0	북동향	1.0	1.0	불가능		0	A72Mx9C8U2	D플랫폼	2024-08-23	0
TRAIN_0007	현장확인	139500000.0	590000		2.0	3.0	동향	2.0	1.0	불가능		0	d22DX4Y4P8	B플랫폼	2025-03-03	0
TRAIN_0008	현장확인	120500000.0	440000	31.55	1.0	2.0	북향	2.0	2.0	가능	18.0	3	G52Iz8V2B9	B플랫폼	2024-05-23	0

- 다수의 범주형 데이터 및 결측치 존재 → 적절한 전처리 필요

# 01 | 분석 개요

## 도메인 조사 및 활용 전략

- 부동산 데이터 도메인 조사

- 매물 확인 방식 활용 :

- 매물의 변수는 현장확인, 전화확인, 서류확인 총 3가지의 값으로 존재.  
따라서 매물의 확인 방식에 따라 어떤 방식이 더 허위 매물이 많은지, 그 차이는 어느정도인지 살펴보고 결과에 대한 이유도 생각해봐야함.

- 보증금 & 월세 활용 :

- 보증금 & 월세가 비슷한 유형 매물에 대해서 과도하게 적거나 많은 것(특히, 적은 것)은 광고용 미끼 허위 매물일 가능성이 높음.  
이상치 탐지 등을 통해 보증금 & 월세에 접근할 필요가 있음.

- 보증금 월세 비율 활용 :

- 보증금과 월세는 일반적으로 보편적인 비율을 따름. 이를 분석하기 위해 보증금\_월세 비율 변수를 생성 및 분석  
보증금\_월세의 비율이 평균에 비해 매우 높거나 매우 낮은 값, 즉 이상치를 보일 때 허위매물일 가능성이 높다고 생각함.  
따라서 보증금 월세 비율에 따른 허위매물의 수를 비교해 보아야 함.

# 01 | 분석 개요

## 도메인 조사 및 활용 전략

### ● 부동산 데이터 도메인 조사

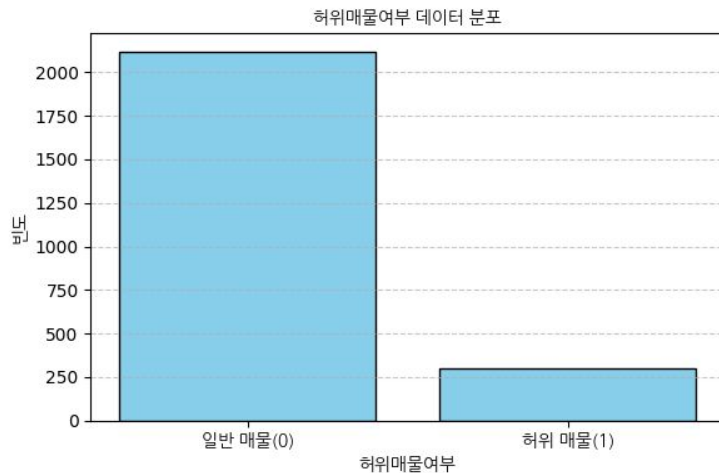
- 전용면적, 보증금, 월세의 관계 :
  - 사용하는 전용면적에 비해 보증금이나 월세가 터무니없이 높거나 낮을 경우 허위 매물일 가능성이 있음.  
따라서 보증금\_전용면적 비율 (보증금/전용면적) 과 월세\_전용면적 비율 (월세/전용면적) 라는 파생변수를 추가하여 비율에 따른 허위매물의 수를 비교해 보아야 함
- 해당층, 층 층 활용 :
  - 해당층에 따른 허위매물의 수를 살펴보고 몇 층에 허위 매물이 많은지 확인하고 상관관계를 살펴볼 것. 층층이 따른 허위매물의 수를 살펴보고 층수가 높은 건물에 포함되는 매물이 허위매물로 나오는지 층수가 낮은 건물에 포함되는 매물이 허위매물로 나오는지 비교해보고 상관관계를 살펴보아야 함
- 중개사무소 활용 :
  - 특정 중개사무소에서 허위매물들이 많이 나타날 수 있기 때문에 중개사무소 별로 허위매물 수를 분석
- 제공플랫폼 활용 :
  - 특정 제공플랫폼에서 허위매물들이 많이 나타날 수 있기 때문에 제공플랫폼 별로 허위매물 수를 분석
- 게재일 활용 :
  - 게재일을 활용해 허위 매물이 많이 나타나는 시기를 월별 & 분기별로 분석  
정부의 부동산 제제 및 정책이 허위 매물의 수에 영향을 줄 수 도 있으니 도메인 분석과 함께 진행할 예정

# 02 | EDA 및 데이터 전처리

## 타겟 변수 분석

- 허위 매물 여부

- 데이터 타입 : int (1 (허위매물 o) / 0 (허위매물 x))
- 허위 매물 각 298 (1), 2117 (0) 개의 데이터
- 허위 매물이 아닌 데이터가 허위 매물인 데이터보다 월등히 많은 것을 알 수 있음
  - 이는 모델 학습 과정에서 클래스 불균형에 의한 모델 분류 성능 (Recall & Precision)의 저하로 이어질 수 있음



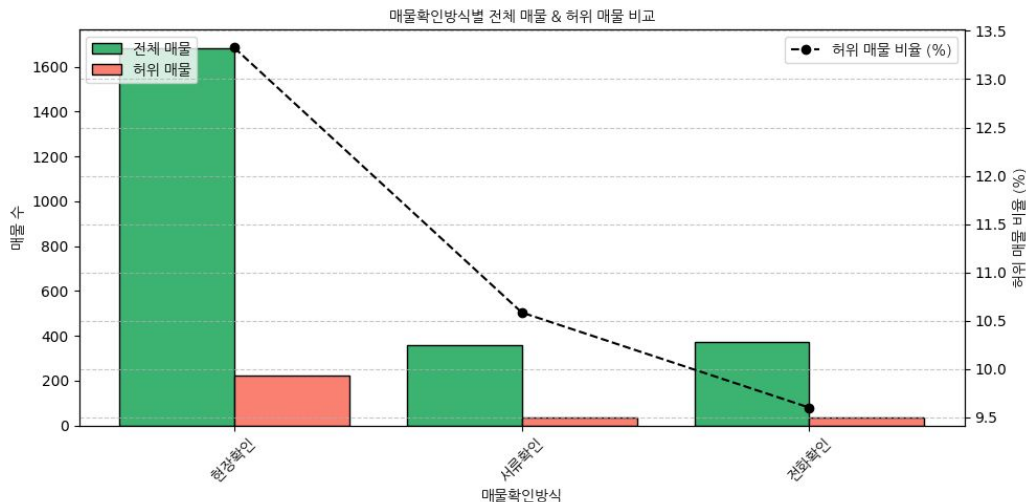


# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략\_매물확인방식

### ● 매물확인방식

- 데이터 타입 : object
- 현장방문, 전화확인, 서류확인 각 1705, 382, 365 개의 데이터
- 매물 확인 방식에 따라 허위 매물의 비율에 큰 차이가 없기 때문에 매물 확인 방식의 중요성을 못 느껴 변수 삭제



# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략\_보증금

- 보증금

- 데이터 타입 : float64
- 결측값 없음
- 모델의 부담을 줄이기 위해 (원) 단위를 (만원)으로 변경
  - 50000000 → 5000

보증금	
count	2.452000e+03
mean	1.574188e+08
std	1.212794e+08
min	5.000000e+06
25%	7.500000e+07
50%	1.325000e+08
75%	1.890000e+08
max	4.090000e+08

# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략\_보증금

### ● 보증금\_Train - 시각화

- 아래의 그래프에서 알 수 있다시피 **다봉 데이터 분포**를 가짐 (여러개의 봉우리를 가지는 형태)
- Train의 데이터를 Test 데이터와 유사하게 전처리 할 필요

- **저가 구간 (0 ~ 0.5억 구간)**

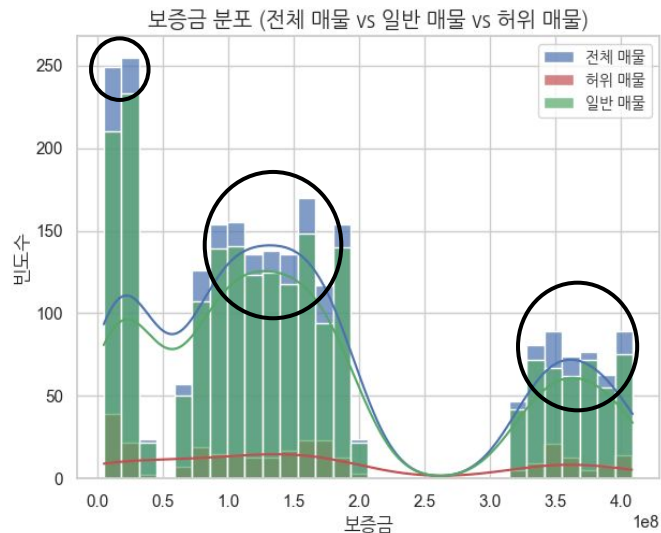
- 보증금의 수준으로 봤을 때, **원룸 & 오피스텔** 등의 거래가 활발
- 초저가 매물에서 허위 매물의 수가 가장 많은 것으로 미뤄 봤을 때, 저가 매물을 함정 마케팅의 용도로 활용할 가능성이 있음
- 저가 매물 중 고가 매물에서는 특이하게 대부분의 매물이 일반 매물임을 알 수 있음

- **중가 구간 (1 ~ 2억 구간)**

- 꽤 많은 매물을 가지고 있지만 비교적 허위 매물의 비율이 적음
- 저가 구간과 비슷한 의견

- **고가 구간 (3.5 ~ 4억 구간)**

- 해당 구간에서는 은행 규제 및 전세에 대한 선호도로 인한 매물 감소



## 02 | EDA 및 데이터 전처리

### 각 변수들의 활용 전략\_보증금

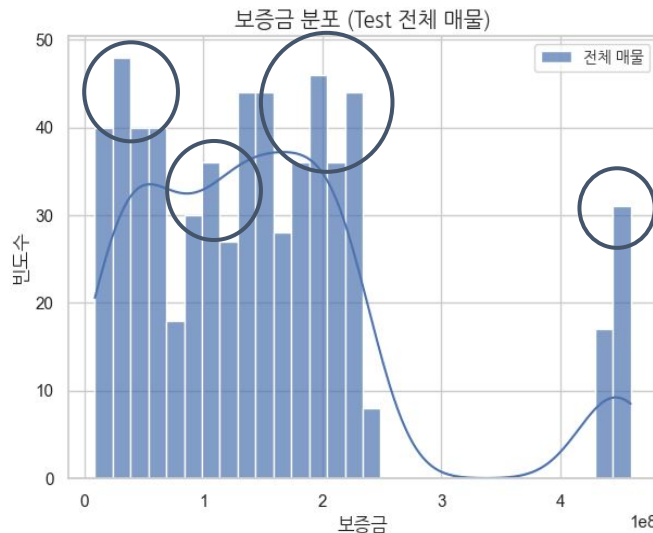
- 보증금\_Test - 시각화

- Train 데이터와 유사하게 다봉 데이터 구조이지만, 저가와 중가 매물의 구분이 모호하게 이어져 있음

- 저가 구간에서 **0.5억 이상**의 데이터가 **비교적 많이 존재**
- 이를 해결하기 위해 train 데이터에 저가와 중가 매물을 연결 할 수 있는 데이터 증강

- Train 데이터와 달리 고가 구간의 금액이 더 큼

- Train에서는 고가 구간의 상한값이 4억 이하였지만, Test에서는 고가 구간의 하한값이 4억 이상임
- 이를 통해  
 **$\text{Max}(\text{Test}) \geq \text{Max}(\text{Train})$  &  $\text{Min}(\text{Test}) \leq \text{Min}(\text{Train})$**   
이므로 Train에서 이상치 제거 불필요



# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략\_월세

- 월세

- 데이터 타입 : int64
- 결측값 없음
- 모델의 부담을 줄이기 위해 (원) 단위를 (만원)으로 변경
  - 200000 → 20
- 월세가 0인 경우는 전세임을 짐작할 수 있음

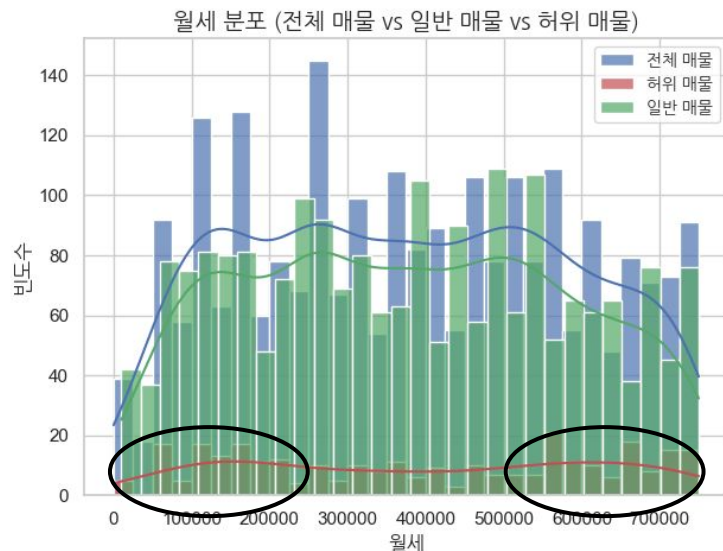
월세	
count	2415.000000
mean	379192.546584
std	206197.198776
min	0.000000
25%	200000.000000
50%	380000.000000
75%	550000.000000
max	750000.000000

## 02 | EDA 및 데이터 전처리

### 각 변수들의 활용 전략\_월세

- 월세\_Train - 시각화

- 보증금과 다르게 다봉 데이터 분포가 아닌 어느정도 균등한 분포를 가짐
  - 30만원 이하의 구간에서 상대적으로 높은 빈도
- 특정 구간에서 상대적으로 높은 허위 매물의 빈도를 보임
  - 10 ~ 20만원 구간에서 상대적으로 허위 매물의 높은 빈도
  - 50만 후반 ~ 고가의 구간에서 상대적으로 허위 매물의 높은 빈도
- Test에서 위의 구간이 균등하게 잘 분포하는지 관찰 필요

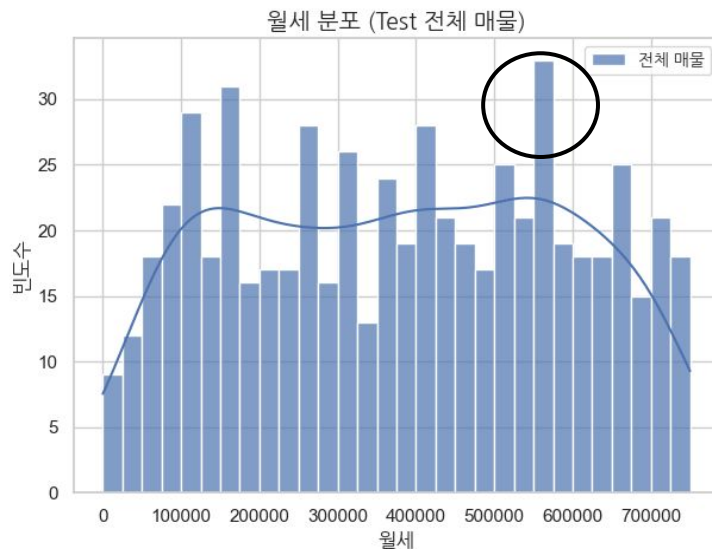


## 02 | EDA 및 데이터 전처리

### 각 변수들의 활용 전략\_월세

- 월세\_Test - 시각화

- Train 데이터와 유사하게 다봉 데이터 구조
- 50만 후반 ~ 60만 초반의 매물이 상대적으로 높은 빈도
  - Train에서 분석한 결과에 의하면 해당 구간에서 허위 매물의 빈도가 상대적으로 증가
- Train 데이터와 최저가 & 최고가가 유사함
  - 이를 통해  
 **$\text{Max}(\text{Test}) \geq \text{Max}(\text{Train})$  &  $\text{Min}(\text{Test}) \leq \text{Min}(\text{Train})$**   
이므로 Train에서 이상치 제거 불필요

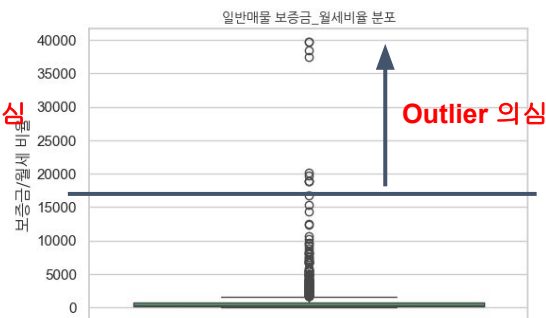
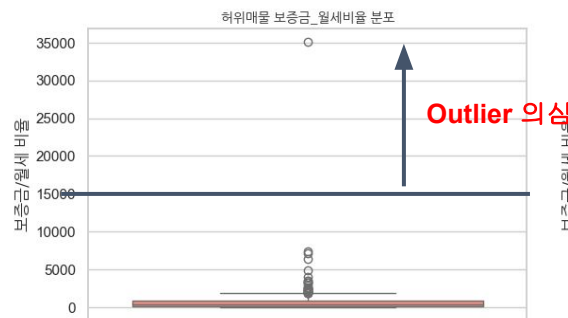
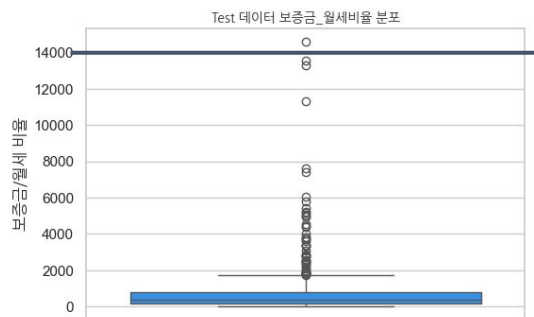


# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략\_보증금 & 월세

### ● 보증금-월세 비율 파생 변수

- 앞서 전처리 한 보증금과 월세 비율 파생 변수 생성
- 월세에 비해 보증금이 과도하게 높거나 낮은 경우 허위 매물 및 이상치 일 수 있음
- **Test** 데이터에서 보증금\_월세의 비율의 **최대 14000대**까지 가는 반면, **Train** 데이터에선 허위 매물 및 일반 매물에서 **14000을 훨씬 뛰어넘는 이상치** 데이터 존재
  - 이를 **Test** 데이터에 대한 이상치로 판단하여 제거



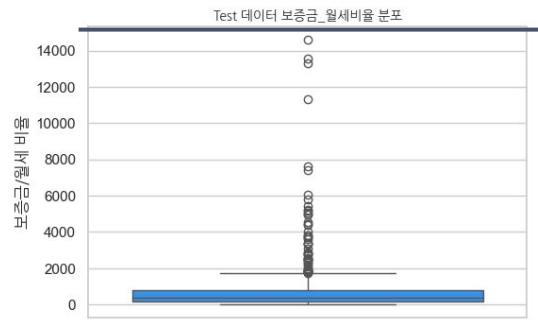
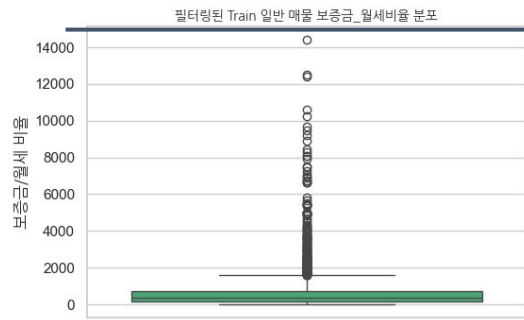
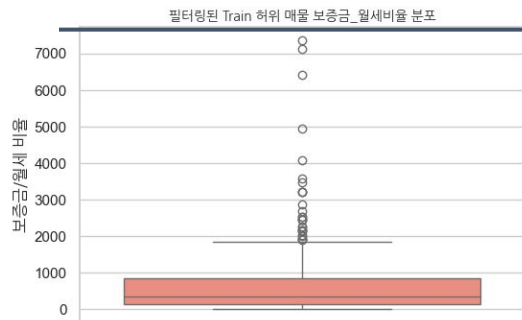


# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략\_보증금 & 월세

### ● 보증금-월세 비율 파생 변수

- Test의 보증금 / 월세 비율의 최대값으로 Train 데이터의 이상치 제거 후 boxplot
- 이상치 제거 후 Train 데이터의 boxplot은 Test와 유사한 데이터로 바뀜



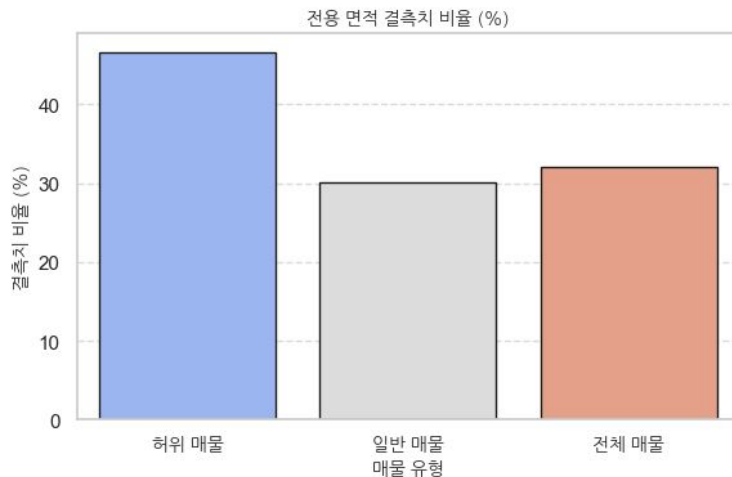
모두 보증금 / 월세 비율이 Test와 Train의 분포가 비슷해 진 것을 볼 수 있음

# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략\_전용면적

### ● 전용면적

- 데이터 타입 : float64
- 787개의 결측값
  - 전용 면적은 부동산 매물의 크기를 판단할 수 있는 중요한 지표로 결측치 전처리가 매우 중요
- 전용면적 결측치의 유무
  - 전체 허위 매물에서 **전용면적 결측치가 있는 매물**의 비율은 약 **46.64%**로 **전용면적의 결측치 유무**가 허위매물 여부에 많은 영향을 미칠 것 같음
- 전용면적의 중요성
  - 전용면적은 부동산 데이터에서 **매물의 크기**를 알 수 있는 가장 중요한 지표임
  - 전용면적 그 자체로도 활용하기 위해 결측치 전처리 필요

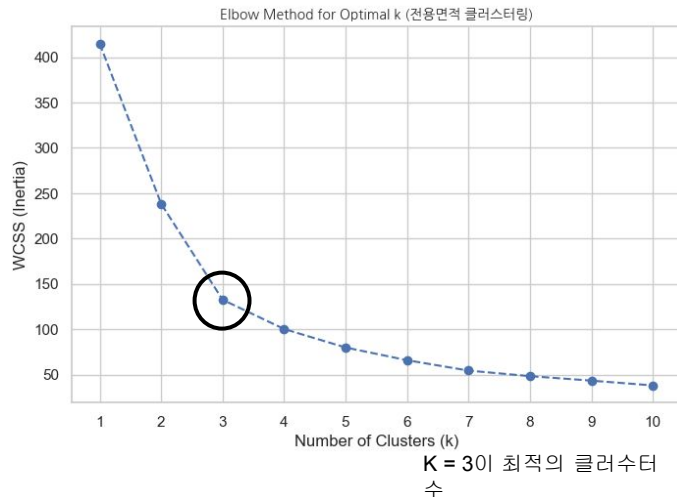


# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략\_전용면적

### ● 전용면적\_결측치 전처리

- 월세, 관리비, 보증금을 변수로 하는 클러스터링 진행
  - `max_clusters`를 하이퍼파라미터 튜닝
  - 군집내 밀집도를 측정하는 `WCSS`(Within-Cluster Sum of Squares, 군집 내 오차 제곱합)을 통해 `max_clusters`를 최적화
- Elbow Method로 최적의 클러스터수 산출
  - Elbow Method는 K를 여러번 테스트 후 `WCSS`가 급격히 줄어드는 지점 (팔꿈치와 닮아 Elbow)을 최적의 K로 지정하는 기법
- 각 결측치를 클러스터 별 중앙값으로 대체

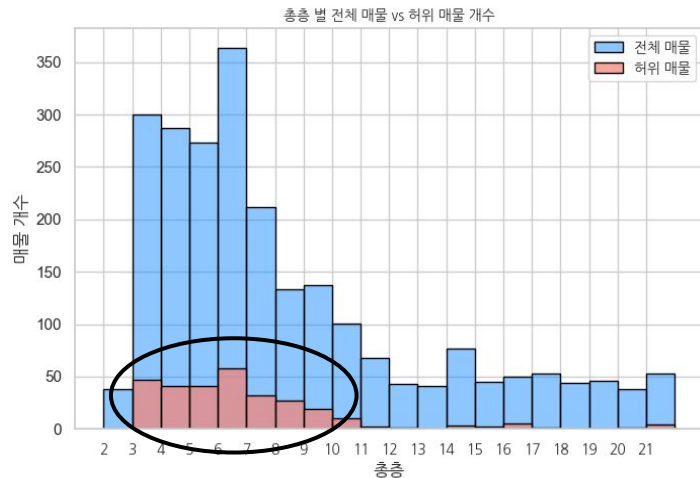


# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략

### ● 총층

- 데이터 타입 : float64
- 16개의 결측값
  - 총층이 결측치일 경우에는 해당층, 전용면적, 방수, 욕실수가 모두 결측치임
- 매물의 총 층수가 높을 수록 대형 건물 매물 → 대형 건물이면 허위 매물일 가능성이 낮다는 가설 설정
  - 오른쪽 사진처럼 총층이 낮은 저층 건물일 경우 허위 매물일 가능성이 비교적 높음을 확인
- 이러한 총층의 경향을 활용하기 위해 결측값 전처리 진행
  - K Means 클러스터링 후 해당 클러스터의 Median으로 결측치 대체

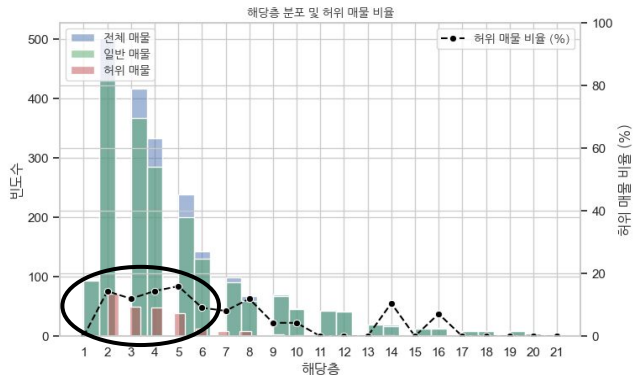
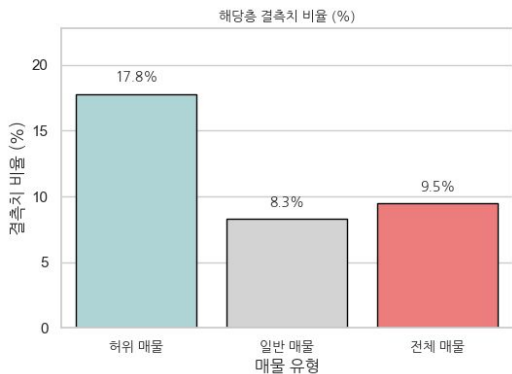


# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략

### ● 해당층

- 데이터 타입 : float64
- 229개의 결측값
  - 다른 변수들의 상관 관계를 통해 층수 예측은 불가능하여 결측치를 채우는 방식은 채택하지 않음
- 매물의 층수가 높을 수록 대형 건물 매물일 가능성이 높음
  - 저층 매물 (2 ~ 5층)에 허위 매물이 상대적으로 높은 빈도로 존재
  - 비율로 비교하였을 때, 큰 유의미성을 가지지 않음
- 해당층의 결측치 유무가 허위 매물 분류에 더 큰 영향을 끼침

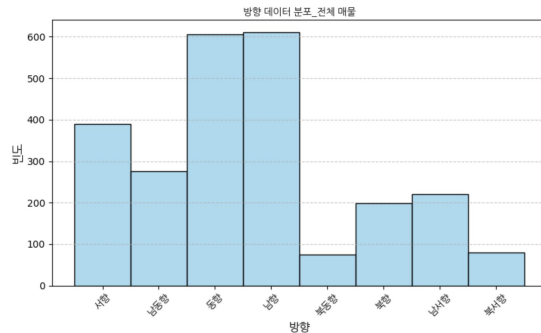


# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략

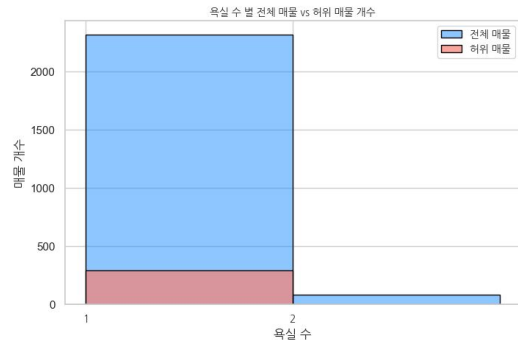
### ● 방향

- 데이터 타입 : **object** (동, 서, 남, 북, 남동, 남서, 북동, 북서향)
- 도메인 조사 시 및 상관관계 분석 시 허위매물 여부에 전혀 영향을 주지 않아 변수 제거



### ● 욕실수

- 데이터 타입 : **float64**
- 16개의 결측값
- 욕실수는 허위 매물 분류에 영향을 주지 않아 변수 제거
  - 대부분의 부동산과 같이 대부분 1,2개의 욕실 수를 가짐

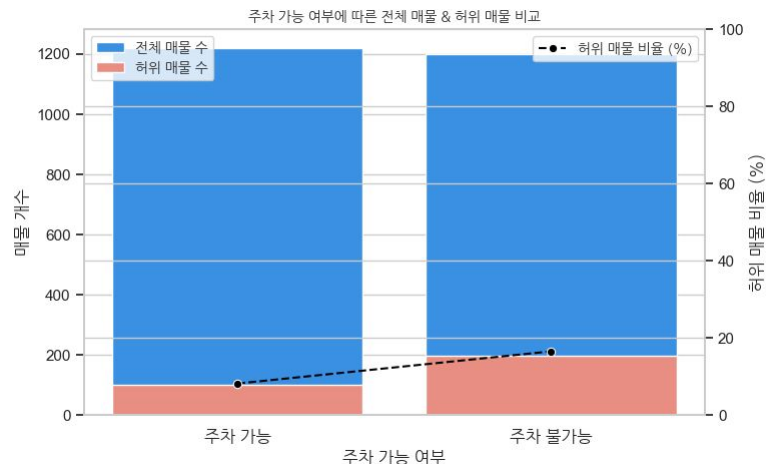


# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략

### ● 주차가능여부

- 데이터 타입 : object
- 가능, 불가능 총 두 개의 값
- 결측값 없음
- 가능 / 불가능의 모집단은 약 1200 : 1200으로 반반
  - 허위 매물과 일반 매물의 비율은 약 1 : 2로 주차가능여부는 유의미한 변수
  - 주차가 가능하다는 내용을 허위로 제공하는 허위 매물일 가능성도 있음

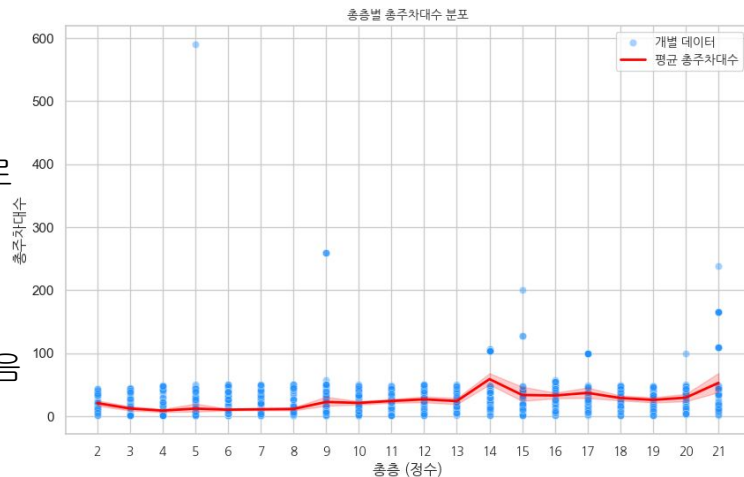


# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략

### ● 총주차대수

- 데이터 타입 : float64
- 696개의 결측값
- 총주차대수는 부동산 매물의 크기에 따라 차이가 존재
- 총주차대수로 사용자를 유도하기 보단 주차 가능 여부로 유도할 것으로 예측
  - 총주차대수를 활용하여 주차 가능 여부 업데이트  
총주차대수 = 0, 주차 불가능  
총주차대수 > 0, 주차 가능
- 총주차대수가 너무 많은 이상치 존재 및 유의미하지 않음
  - 앞서 언급한 주차 가능 여부 업데이트로만 활용 후 변수 제거





## 02 | EDA 및 데이터 전처리

### 각 변수들의 활용 전략

- 관리비

- 데이터 타입 : int64
- 결측값 없음
- 보증금과 월세를 미루어 보아 단위가 (천원)임을 짐작할 수 있음
  - 앞선 보증금, 월세의 단위를 (만원)으로 맞췄으므로 10으로 나누어 단위를 맞춤  
천원 단위를 만원 단위로 대체  
7(천원) → 0.7(만원)

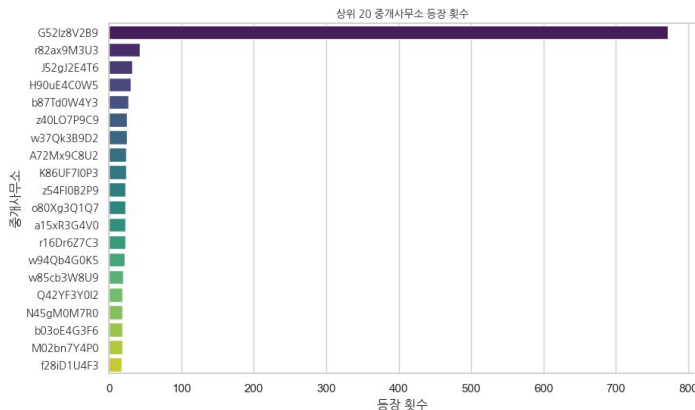
관리비	
count	2419.000000
mean	5.486151
std	5.647353
min	0.000000
25%	0.000000
50%	5.000000
75%	9.000000
max	96.000000

# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략

### ● 중개사무소

- 데이터 타입 : **object** (영어&숫자 조합 문자열)
- 총 279개의 고유값
- 결측값 없음
- **G52Iz8VB9** 중개사무소의 매물이 가장 많으며, 이 중개사무소의 799개 매물 모두 허위매물이 아님
  - 중개사무소 종류가 매우 많아 인코딩 시 학습에 어려움을 줄 수 있음
  - **G52Iz8VB9** 중개사무소인지 아닌지 여부만 나타내는 칼럼으로 대체



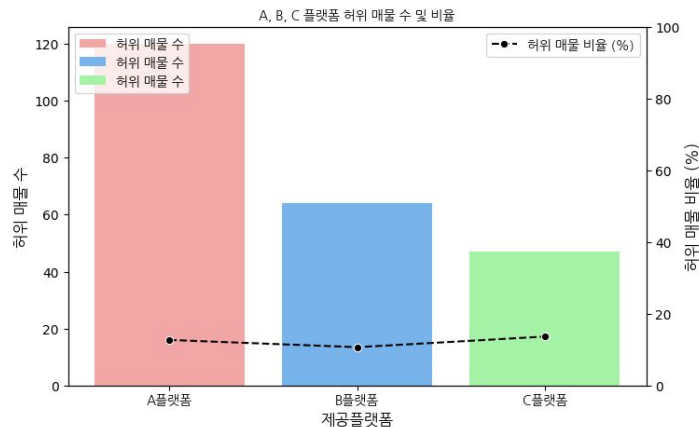
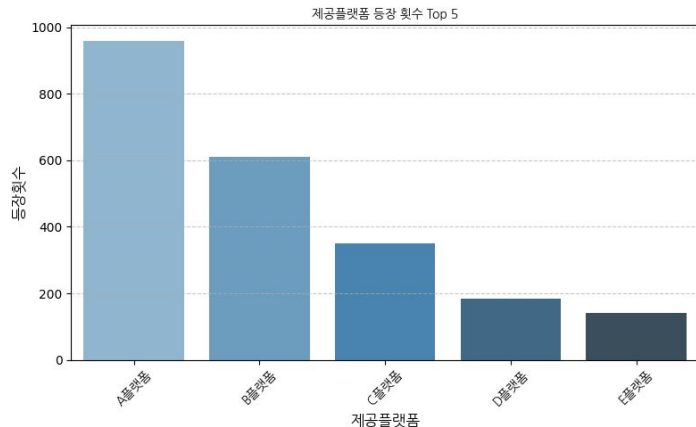
중개사무소	
G52Iz8VB9	799
r82ax9M3U3	43
J52gJ2E4T6	32
H90uE4C0W5	30
b87Td0W4Y3	27
...	
m75Dz8P6I7	1
A21Yr4B1U8	1
g11ci7P5V1	1
D26uW0Q2N3	1
L27J03N6S2	1

# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략

### ● 제공플랫폼

- 데이터 타입 : object ( 총 13개의 A~M플랫폼 )
- 결측값 없음
- 특정 플랫폼에서 높은 빈도를 보임
  - A, B, C 플랫폼에서 상대적으로 높은 빈도
  - 절대적인 매출의 양이 많은 A에서 허위 매출의 수가 더 많았지만, 허위 매출의 비율은 비슷함을 확인



## 02 | EDA 및 데이터 전처리

### 각 변수들의 활용 전략

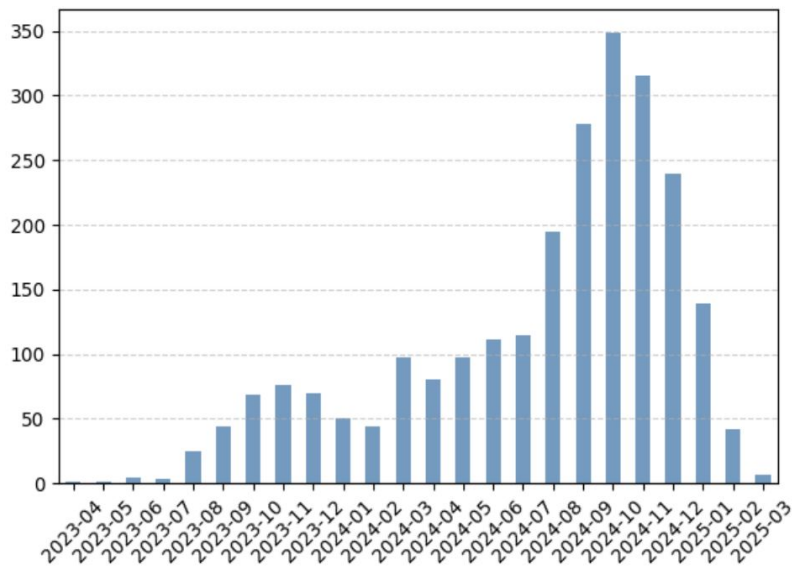
- **게재일**
  - 데이터 타입 : `object` ( 형식 : 2025-02-14 )
  - 결측값 없음
- 값이 매우 제각각이므로 인코딩 시 어려움  
→ 게재 연-월로 분리하여 게재연 & 게재월로 활용

# 02 | EDA 및 데이터 전처리

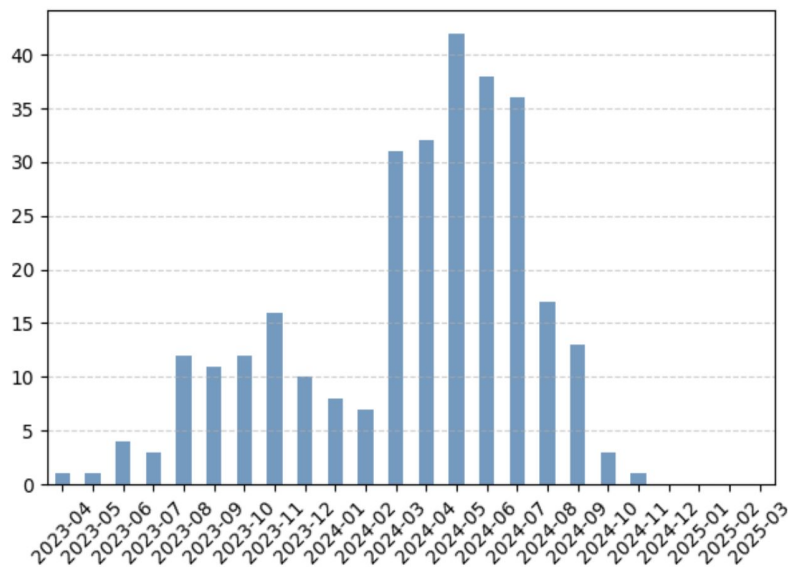
## 각 변수들의 활용 전략

- **계재일**

- 연-월 별 매출 수 히스토그램



- 연-월 별 허위매출 수 히스토그램



# 02 | EDA 및 데이터 전처리

## 각 변수들의 활용 전략

### ● 게재일

- 2024년 10월, 11월 두 번에 걸친 한국은행 기준금리 인하  
→ 부동산 공급 감소

- 2024.07 정부는 부동산관계장관회의를 열어  
부동산시장 거래질서 확립 의지 밝힘
- 2024.08 '제2차 부동산 시장 및 공급상황 점검 TF'  
회의 및 현장 점검 진행

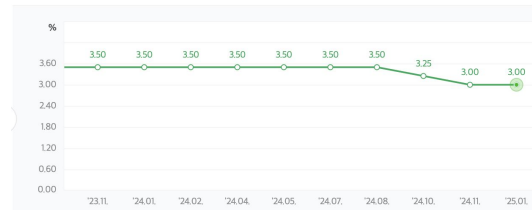
한국 중앙은행 기준금리 ①

통화 및 금리

+ 알림

3.00 %

시장영향력 매우 높음



국제신문 PICK · 2024.08.01 · 네이버뉴스

정부 "수도권 전 지역 대상 부동산 허위매물 등 현장 점검"

최근 서울 등 일부 지역을 중심으로 아파트 가격 상승세가 이어지자 정부가 수도권 전 지역을 대상으로 부동산 허위 매물·신고와 편법 증여·대출 등 위법 행위 발생 여부를 현장 점검하기로 했다. 정부는 1일 '제2차 부동산 시장 및 공급상황 점검 TF' 회의를...



정부 "수도권 전 지역 부동산 현장점검... 허위매물·편법증여 조사" 뉴스1 · 2024.08.01 · 네이버뉴스

서울 아파트 가격상승에... 정부 "허위매물·편법증여 등 현장점... 부산일보 PICK · 2024.08.01 · 네이버뉴스

한국경제 PICK · A5면 1단 · 2024.07.18 · 네이버뉴스

투기 수요에 칼 댄 정부... 편법증여·허위매물 살살이 찾는다

정부는 18일 부동산관계장관회의를 열어 부동산시장 거래질서 확립 의지를 밝혔다. 국토교통부는 올해 수도권 주택 거래분을 대상으로 대출 규제 회피, 편법 증여, 허위 매물 등을 전수조사하고 있다. 불법행위가 확인되면 사안에 따라 행정조치 등을 할 계획이다. 최상목 부총리 겸 기획재정부 장관은 "시장...

# 02 | EDA 및 데이터 전처리

## 최종 전처리

- 최종 전처리 정리

- 보증금 :
  - (원) 단위를 (만원) 으로 변경
- 월세 :
  - (원) 단위를 (만원)으로 변경
- 관리비 :
  - (원) 단위를 (천원)으로 변경
- 전용면적 :
  - 결측치를 클러스터링을 이용해 대체
  - 전용면적 결측치 여부 컬럼 추가
- 총층 :
  - 결측치를 클러스터링을 이용해 대체
- 총주차대수 :
  - 주차 가능여부 컬럼 추가
- 해당층 :
  - 결측치 여부 컬럼 추가
- 중개사무소 :
  - G52Iz8VB9 중개사무소 여부
- 게재일 :
  - 게재년월 로 변경하여 추가

# 02 | EDA 및 데이터 전처리

## 최종 전처리

- **train data**

- 전처리를 마친 데이터

	매물확인방식	보증금	월세	전용면적	해당층	총층	주차가능여부	관리비	제공플랫폼	허위매물여부	G52중개사무소여부	게재연	게재월
0	현장확인	40250.0	47.0	26.892268	4.721569	15.0	1	9.6	B플랫폼	0	0	2024	10
1	현장확인	17050.0	20.0	27.144463	3.000000	4.0	0	0.0	D플랫폼	0	0	2024	12
2	전화확인	11400.0	38.0	27.257531	2.000000	3.0	0	0.0	A플랫폼	0	0	2024	11
3	현장확인	16350.0	3.0	36.300000	3.000000	9.0	1	1.0	A플랫폼	0	1	2024	11
4	현장확인	34600.0	53.0	26.892268	3.000000	3.0	0	0.0	B플랫폼	1	0	2024	6



# 03 | 모델링 및 모델 학습

## Macro F1 Score

- **Macro F1 Score**

- Macro F1 Score를 성능 평가 지표로 선택한 이유는 타겟 클래스 1 (허위 매물), 0 (정상 매물)의 분류에서 클래스 불균형을 평가하기 위하여 선택함
  - Macro F1 Score는 아래처럼 precision과 recall의 조화 평균을 성능 지표로 활용
  - 이를 통해 분류 모델이 클래스 불균형을 잘 해결했는지 평가 가능

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

# 03 | 모델링 및 모델 학습

## Base 분류 모델

- 트리 기반 모델

- 부동산 데이터는 범주형 데이터 + 수치형 변수가 혼합된 구조의 데이터이며 변수들이 선형적인 관계를 갖지 않음 (비선형)
- 허위 매물 여부는 특정 조건에서 분류할 수 있는 것이 아닌, 다양한 조건이 조합되어 구분할 수 있으므로 이를 잘 구분할 수 있는 트리 기반 모델이 분류에 유리함
- 아래처럼 **Boosting(부스팅)** & **Ensemble(앙상블)** 기능을 가진 모델을 활용
  - XGBoost
  - LightGBM
  - CatBoost
  - Random Forest
  - ...

- 최종 모델 - XGBoost

- XGBoost 선정 이유 - 특징 및 장단점
  - L1, L2 규제를 활용한 일반화 성능 향상
  - min\_child\_weight, subsample, colsample\_bytree, gamma 등의 하이퍼 파라미터 튜닝으로 과적합 방지 가능
  - XGBoost 예측 과정에서의 Feature Importance 분석 가능

# 03 | 모델링 및 모델 학습

## 모델 하이퍼파라미터 튜닝

### ● 타겟 클래스 불균형 완화

- 불균형 데이터에서는 모델이 대부분 다수 클래스(일반 매물)에 맞춰 학습되어, 소수 클래스(허위 매물)를 제대로 예측하지 못하는 문제가 발생
- 이를 해결하기 위해 타겟 클래스의 분포 비율을 학습 가중치로 제어하여 소수 클래스에 더 집중하여 학습 진행
  - `scale_pos_weight = class_counts[0] / class_counts[1]` (다수 클래스) / (소수 클래스)  
ex) 다수 : 900, 소수 : 100  $\rightarrow 900/100 = 9$  즉, 소수 클래스의 손실 (loss)를 9배 더 크게 학습하여 이전보다 집중하여 학습 가능
- But, 성능이 좋지 않아 최종적으로 적용하지 않았음  
오히려 일반 매물을 허위 매물이라 오판단하여 Precision이 낮아져 Macro F1 Score가 낮아지는 현상 발생

# 03 | 모델링 및 모델 학습

## 모델 하이퍼파라미터 튜닝

- **Optuna 하이퍼파라미터 튜닝**

- 이전 슬라이드에서 소개한 모델들은 다양한 하이퍼파라미터로 모델 성능 향상을 할 수 있음
- 하이퍼파라미터 튜닝에서 가장 많이 사용되는 **Optuna** 알고리즘을 활용하여 하이퍼 파라미터 튜닝 진행

- **사용된 파라미터**

**max\_depth** : 사용되는 트리의 최대 깊이

**learning\_rate** : 업데이트되는 가중치의 크기

**subsample** : 트리를 학습할 때 사용되는 샘플의 비율

**colsample\_bytree** : 트리를 학습할 때 사용되는 특징의 비율

**alpha** : L1 정규화의 항 개수

**lambda** : L2 정규화의 항 개수

**min\_child\_weight** : 리프노드에 필요한 최소 가중치의 합

```
Best trial:
Value: 0.19298907835955548
Params:
  max_depth: 9
  learning_rate: 0.08778580254757205
  subsample: 0.6036768645184584
  colsample_bytree: 0.6272593267504225
  alpha: 0.3029189693494863
  lambda: 0.4289455433238625
  min_child_weight: 10
```

최종 하이퍼파라미터 값

# 04 | 모델 평가

## 모델 성능

- 모델 평가


**부동산 허위매물 분류 해커톤: 가짜를 색출하라!**

데이콘 해커톤 | 알고리즘 | 정형 | 분류 | 허위매물 | Macro F1 Score

₩ 상금 : 데이스쿨 프로 구독권

🕒 2025.01.06 ~ 2025.02.28 09:59 [+ Google Calendar](#)

👤 866명 📅 D-17



1132017

submission.csv

edit

2025-02-11 15:46:08

0.8742982336

24

0.87429

51

6시간 전

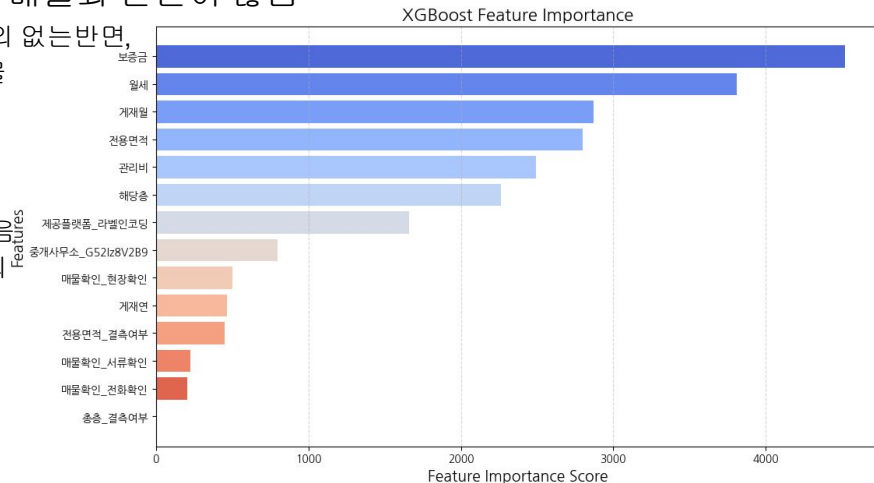
Dacon 리더보드 25등(02/12/9PM 기준) Top 10%  
이내!!

# 05 | 분석 결론

수정

## ● XGBoost의 Feature Importance를 통한 결론 및 인사이트

- 보증금 & 월세가 가장 중요한 변수
  - 허위 매물은 주로 보증금 & 월세를 활용하여 사용자를 유인하는 경우가 많음
- 부동산 시장의 분위기 (부동산 매물의 게재 타이밍)이 중요한 변수
  - 특정 시기, 전체적인 부동산 시장의 분위기 (정책 및 은행 규제 등)에 따라 허위 매물의 수가 달라짐
- EDA 분석 결과 특정 중개사무소 & 플랫폼이 허위 매물과 관련이 많음
  - 몇몇 중개사무소 및 플랫폼은 허위 매물이 아예 또는 거의 없는 반면, 몇몇 중개사무소 및 플랫폼은 대다수의 매물이 허위 매물
- 아쉬운 점
  - 부동산 매물이라는 데이터 특성 상 지역적 영향을 받을 지역을 특정 및 예측할 수 있는 데이터가 없어서 아쉬웠음
  - 분석 과정의 오류인지 분석 결과를 통해 진행한 전처리의 결과가 성능 저하로 이어져 아쉬웠음





Q & A

---

감사합니다