KUBIG
DATA SCIENCE & AI

중 탐구

부동산 허위 매물 탐지

Team | 집중탐구
19기 심서현
20기 김채원
21기 김지엽
21기 엄희문

KUBIG
DATA SCIENCE & AI

CONTENTS

01

Subject

- 주제 선정 배경
- 주제 설명

02

Processing

- EDA, Feature Engineering 진행
- 전처리 기법 및 모델 적용

03

Results

- ML 모델 결과
- DL 모델 결과
- TabPFN 결과

04

Conclusion

- 최종 성능 및 결론

KUBIG
DATA SCIENCE & AI

01. Subject

01. Subject - 주제 설명


```
0   ID          2452 non-null   object
1   매물확인방식    2452 non-null   object
2   보증금         2452 non-null   float64
3   월세          2452 non-null   int64
4   전용면적        1665 non-null   float64
5   해당층         2223 non-null   float64
6   총층          2436 non-null   float64
7   방향          2452 non-null   object
8   방수          2436 non-null   float64
9   욕실수         2434 non-null   float64
10  주차가능여부    2452 non-null   object
11  총주차대수      1756 non-null   float64
12  관리비         2452 non-null   int64
13  중개사무소      2452 non-null   object
14  제공플랫폼      2452 non-null   object
15  게재일         2452 non-null   object
16  허위매물여부    2452 non-null   int64
```

- 집을 직접 방문하거나 사진을 보지 않고

  수치적인 요소만으로 허위매물 예측이 가능한가?

- Tabular 데이터를 다루는 다양한 머신러닝 툴 복습
- 세션에서 배운 딥러닝 모델의 개념과 구조 활용 가능

KUBIG
DATA SCIENCE & AI

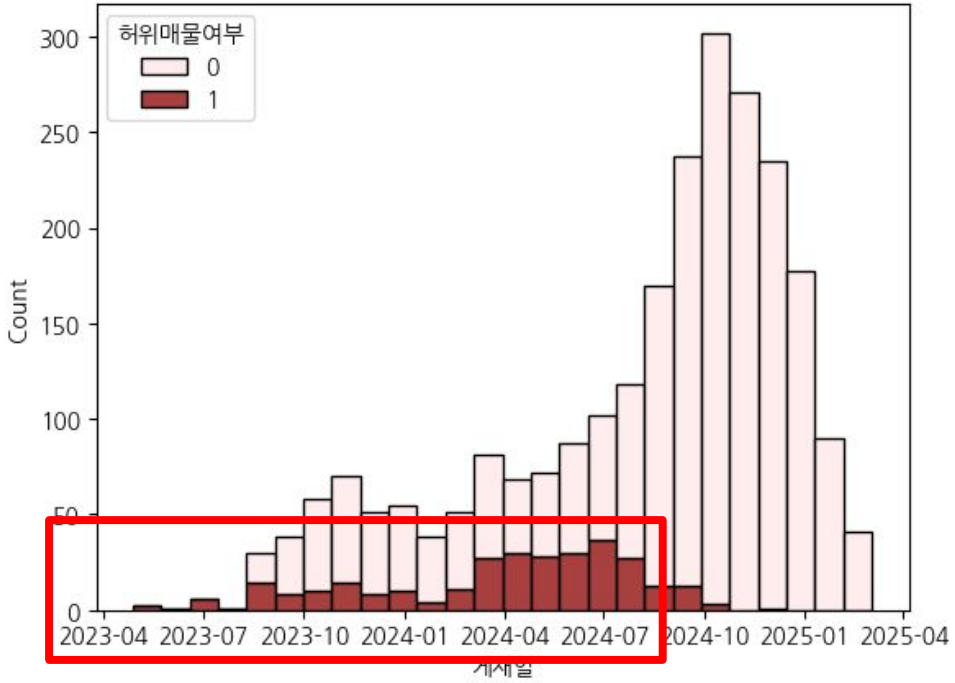
02. Processing

02. Processing - EDA, Feature Engineering

허위매물 걸러내는 6단계 체크리스트	
구분	내용
Step 1	시세보다 가격이 너무 낮다면 일단 의심해 보세요 : 고객들의 관심을 끌기 위한 미끼 상품용 허위매물일 가능성이 높음
Step 2	매물 등록일이 너무 오래 지난 건 아닌지 확인해 보세요 : 등록일이 너무 오래 지난 주택이라면 애초에 존재하지 않는 매물이 가능성이 높음
Step 3	올려진 사진이 너무 적지는 않은지, 주택 내부를 잘 보여주고 있는지도 살펴보세요 : 허위매물이라면 다양한 사진을 올리기가 힘듦. 내부 구조를 파악하기 힘든 사진들만 올려져 있다면 의심해 보아야 함
Step 4	기본 정보가 정확하게 적혀 있는지 꼼꼼히 읽어보세요 : 법에 따라 온라인 부동산 광고에는 소재지, 면적, 가격, 주택 유형, 거래 형태 등 12가지 기본 정보가 기재 되어야 함.
Step 5	특정한 문구가 등장하는지도 따져보세요 : '주택의 실제 모습은 사진과 다를 수 있다' 등의 문구가 적혀 있거나 '단기 임대', '저금리 대출이자' 등의 표현이 반복적으로 등장할 경우 허위매물일 가능성이 있음
Step 6	방문 전에 매물이 있는지 꼭 전화로 확인하세요. 방문 당일에도 다시 한번 확인하세요 : 마음에 드는 매물을 찾았다면 반드시 해당 중개업소에 연락해 계약 가능 여부를 확인한 뒤 일정을 잡고 방문해야 함

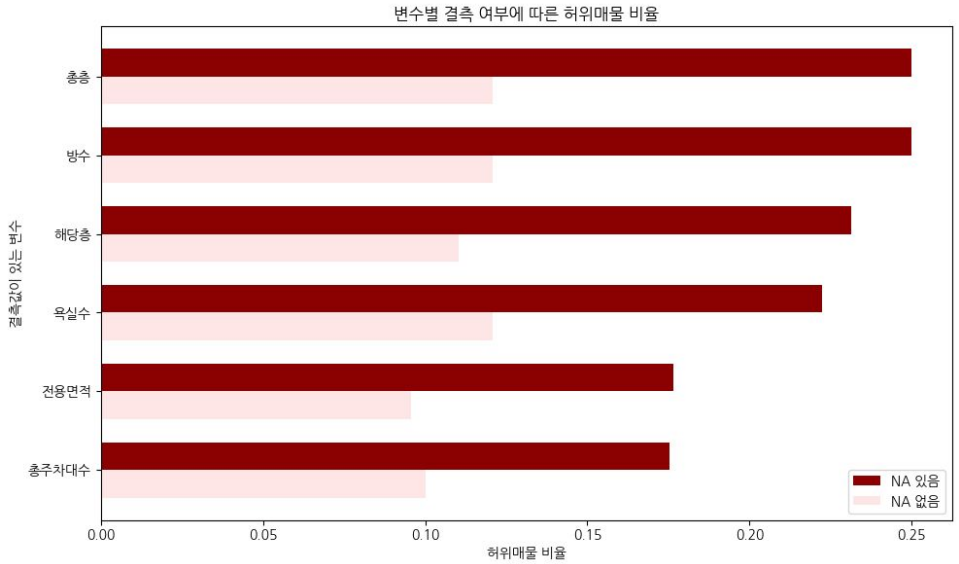
출처: https://brunch.co.kr/@dambee/12

02. Processing - EDA, Feature Engineering



수치형 변수 **게재일수 (최신 일자 - 등록 일자)** 추가

02. Processing - EDA, Feature Engineering



3: 월세 = 0

2: '총층', '방수', '해당층', 욕실' 이 NA

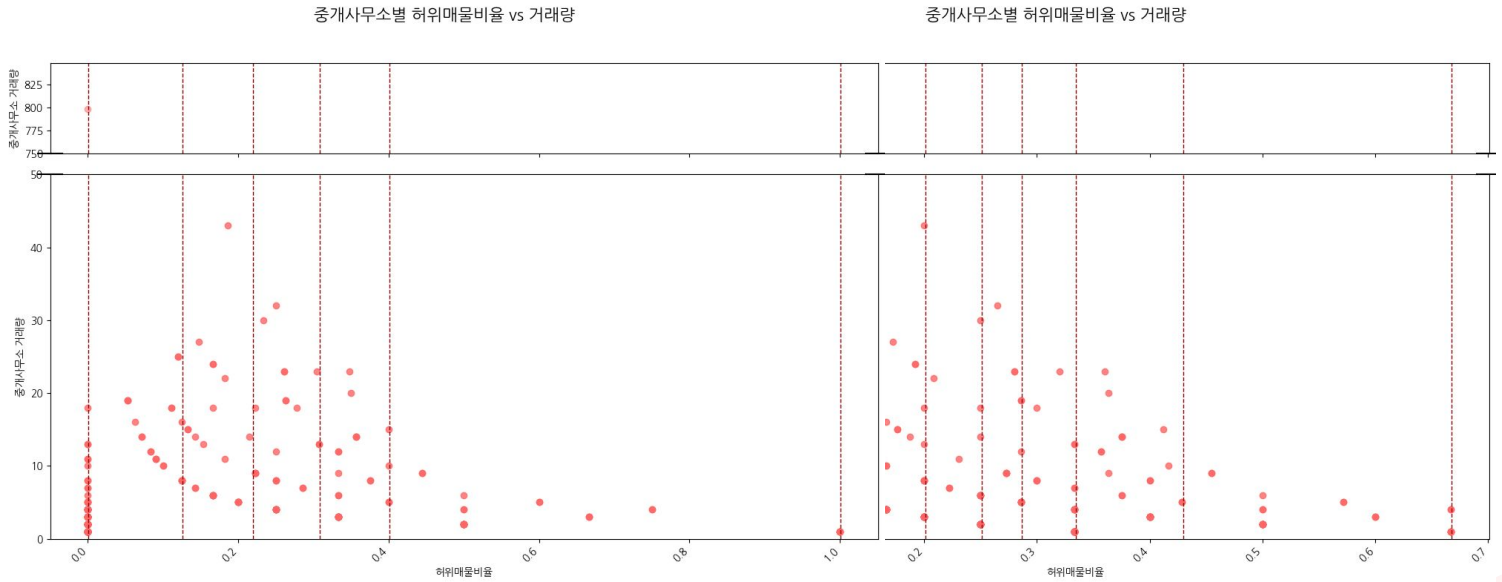
1: '전용면적', '총주차대수'이 NA

0: NA 없음

- 특정 변수의 NA 값과 허위 매물 여부 연관성 발견

  ➡ 해당 특징을 고려해 Label Encoding 진행

02. Processing - EDA, Feature Engineering



Best Result: 0.8871

In statistics, **additive smoothing**, also called **Laplace smoothing**[1] or **Lidstone smoothing**, is a technique used to smooth count data, eliminating issues caused by certain values having 0 occurrences. Given a set of observation counts $\mathbf{x} = \langle x_1, x_2, \ldots, x_d \rangle$ from a d-dimensional multinomial distribution with N trials, a "smoothed" version of the counts gives the estimator

$$\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ldots, d),$$

02. Processing - 모델 적용 및 선정

- **적용 모델**
  - **Random Forest**
    - 다수의 결정 트리를 결합하여 중요도를 해석하기 용이
  - **CatBoost**
    - 범주형 데이터를 자동 처리하는 gradient boosting 기반 모델
  - **LGBM(Light Gradient Boosting Model)**
    - 리프 중심 성장 방식으로 효율적인 학습이 가능한 부스팅 모델
  - **XGBoost**
    - 규제가 포함된 경사 부스팅 기반 모델로 병렬 학습이 가능
- **성능향상을 위한 DL 모델 접목 및 스태킹 시도**
  - **MLP -** 완전 연결층으로 구성된 신경망
  - **TabNet -** attention을 활용하여 중요한 특성을 집중적으로 학습하는 테이블 데이터 딥러닝 모델

KUBIG
DATA SCIENCE & AI

03. Results

03. Result - ML 모델

ML 모델별 최고성능

- Random Forest (Optuna)
- Catboost (SMOTE + Optuna)
- LGBM (SMOTE + roc-auc 기반 Bayesian Optimization)
- XGBoost (Optuna)

성능\모델	Random Forest	CatBoost	LGBM	XGBoost
marco F1 score	0.8564	0.8309	0.8710	0.8871

03. Result - DNN

- MLP
  - 전처리: KNNImputer, Ordinal Encoder 적용
  - Optuna + Smote + Adnam
- TabNet
  - Optuna + Smote

성능\모델	MLP	TabNet
marco F1 score	0.84	0.77

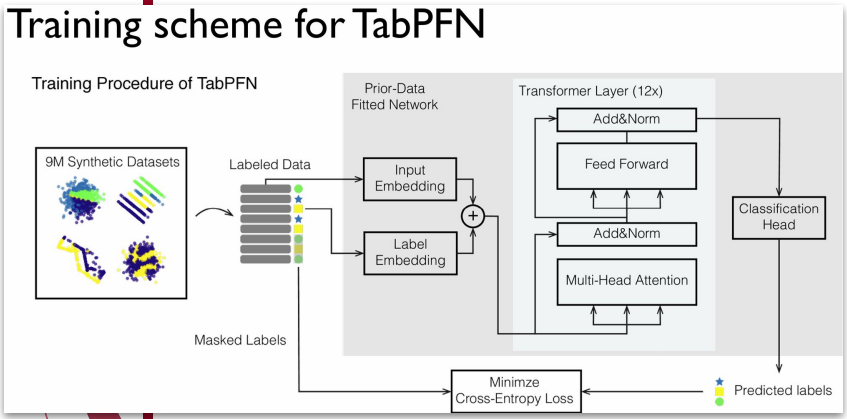
03. Extra Trial - TabPFN

Accurate predictions on small data with a tabular foundation model

https://doi.org/10.1038/s41586-024-08328-6

Received: 17 May 2024

Noah Hollmann[1,2,3,7], Samuel Müller[1,7], Lennart Purucker[1], Arjun Krishnakumar[1], Max Körfer[1], Shi Bin Hoo[1], Robin Tibor Schirrmeister[4,5] & Frank Hutter[1,3,6]



- **TabPFN**
  - Meta-learning을 활용한 사전 학습된 Transformer 모델
    - meta-learning: 다양한 문제를 학습하여 새로운 문제에서도 빠르게 적응할 수 있도록 학습(learning algorithm을 학습)
  - 사전 학습 과정 시 여러 합성 데이터 활용
  - 추가 학습 없이 주어진 context만으로 예측을 수행하는 In-Context Learning 방식

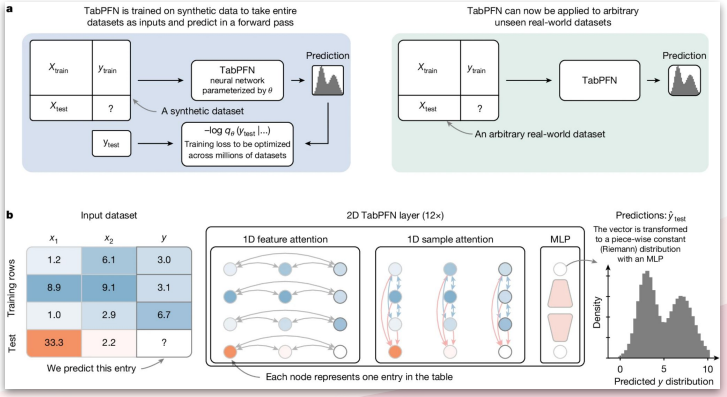
03. Extra Trial - TabPFN

- marco f1 score : 0.88811
- 전처리 기법과 smote 적용 후 가장 높은 F1 score 기록
- 다른 기법들과 비교하였을 때 가장 좋은 성능을 보임

→ 기존 머신러닝 모델보다 우수한 성능

→ 별도 학습 과정 없이 즉시 추론 가능

→ 더 빠른 시간 내 계산



KUBIG
DATA SCIENCE & AI

04. Conclusion

04. Conclusion

- 사용한 데이터셋과 실제 부동산 데이터 특징 간 괴리 존재
  - ex) 월세 - 보증금 간 상관관계 지수 : 0.01
  - 이상치로 보이는 데이터 분포 확인
- 결측치 다수
  - 데이터 간 상관관계 파악 어려움
- 정상 매물과 허위 매물 데이터 간 불균형
  - 과적합 문제
- 소규모의 데이터셋

→ 허위 매물 데이터의 특징 파악 어려움

→ EDA 및 전처리 필요성 높음

04. Conclusion

- 초기 예상

- Tree 기반 ML 모델(Xgboost, LGBM)이 우세

- Deep Learning 적용 결과

- 일반적으로 정형 데이터에서는 DL 모델이 성능이 낮았으며, 데이터 크기의 한계로 인해
  학습이 어려운 측면이 존재

- 반면, TabPFN은 예외적으로 높은 성능을 보였으며, 특히 데이터가 적거나 불균형할 때에도
  높은 성능을 보임

KUBIG
DATA SCIENCE & AI

Thank You

Team 집중탐구