2023 빅콘테스트

빅데이터플랫폼 활용 분야 자유주제

토지 가치 예측을 위한 AI 기반 모델 개발

김도환 (ehghks021203@gmail.com)
고건호 (rhrjsgh000@naver.com)

01.
분석 배경 및 목적

배경

- 토지 가치 예측은 부동산 투자, 도시 계획, 자산 관리 등 다양한 분야에서 중요한 역할을 함

- 부동산 시장은 투자 및 개발자, 정부 기관, 도시 계획자, 시민들에게 영향을 미치는 핵심적인 부분 중 하나로, 정확한 토지 가치 예측은 이들 모두에게 이익을 제공

- 인공지능 및 기계학습 기술의 발전은 대량의 데이터를 기반으로 한 토지 가치 예측의 정확도와 신뢰성을 향상시킬 수 있는 새로운 기회를 제공하고 있음

=> 토지 가치 예측을 위한 AI 모델의 개발은 현실적이고 필수적인 과제가 됨

2. 분석 목적 및 필요성

도시 계획

도시 계획에 관련된 결정에 과학적 근거를 제공하여 도시의 지속 가능한 발전 지원

부동산 투자

부동산 투자자들에게 정확한 투자 의사 결정 도움

부동산 시장 투명성

부동산 시장의 투명성을 높여 시장 건전성을 유지

3. 분석 절차

01

데이터 수집 및 전처리

- 토지 관련 데이터 수집
- 결측치 및 이상치 제거

02

AI 모델 개발

- XGBoost 모델 활용

03

모델 평가 및 성능 개선

- 평균 제곱근 오차(RMSE)
- 예측 오차

04

결과 해석 및 시각화

- 다양한 그래프 이용
- 예측 결과 지도 시각화

05

응용 및 확장

- 웹 서비스 구현

02.
문제 수행 내용

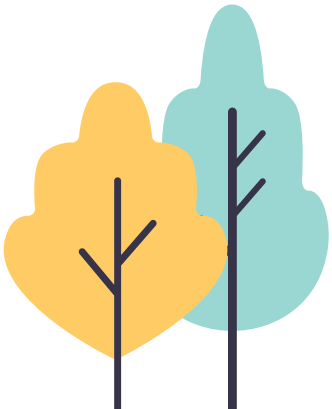
1. 데이터 선정

국토교통부
토지 매매 신고 조회 서비스

국토교통부
토지 특성 정보

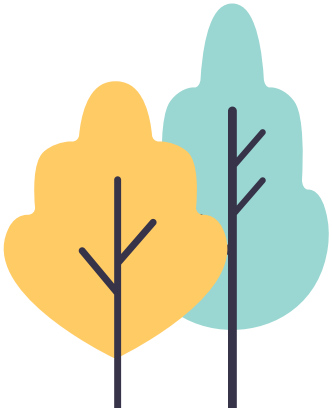
통계청 SGIS
통계지리정보

국토교통부
지가변동률정보



2. 데이터 조인 방식

PNU 코드 형식



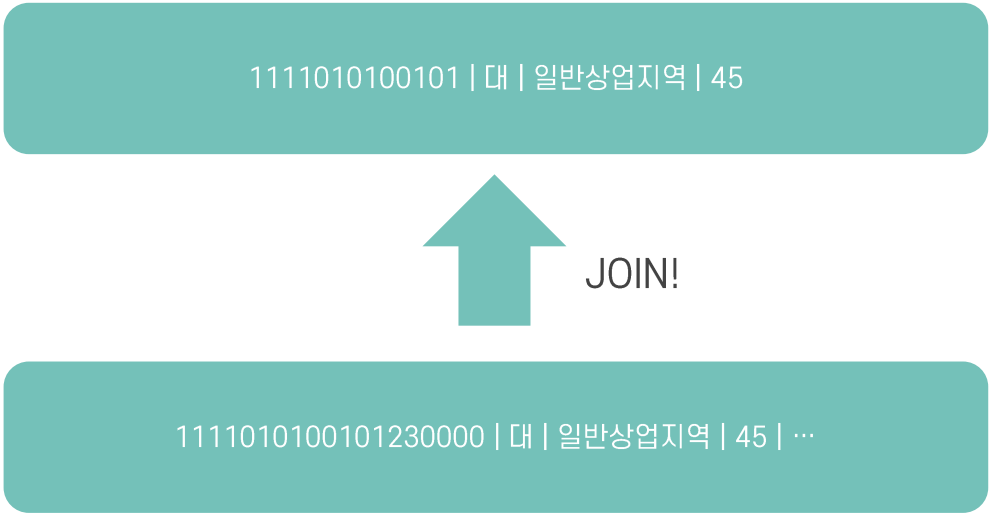
조인 방식

법정동 | 필지 | 지번 | 지목 | 용도지역 | 거래면적
서울특별시 종로구 청운동 (1111010100) | 일반 | 1** | 대 | 일반상업지역 | 45

1111010100101 | 대 | 일반상업지역 | 45

- 법정동 코드를 PNU 코드 10자리로 변환
- 10자리 PNU 코드 뒤에 필지가 일반이면 1, 산이면 2
- 마스킹 되기 전까지의 지번을 PNU 코드 뒤에 이어붙임

조인 방식



- 토지 특성 정보 데이터에서 지목과 용도지역이 일치하는 데이터만 추출

- 추출된 데이터 중 토지 면적이 일치하는 데이터만 또 다시 추출

- 마지막으로 추출된 데이터가 단 하나라면, 해당 지역의 PNU 코드를 사용

3. 연속형 변수 및 범주형 변수 분석

변수 목록

변수명	설명	형식
PNU	토지의 PNU 코드 (19자리)	범주형
DealAmount	토지의 실거래 가격 (단위: 만원)	연속형
DealMonth	토지의 거래 월	연속형
LdCodeNm	토지의 법정동명	범주형
RegstrSeCodeNm	토지의 대장구분명	범주형
StdrYear	토지의 기준년도 (거래년도)	연속형
LndcgrCodeNm	토지의 지목명	범주형
LndpclAr	토지의 면적 (거래면적)	연속형
PrposAreaNm	토지의 용도지역	범주형
LadUseSittnNm	토지의 이용상황명	범주형
TpgrphHgCodeNm	토지의 지형 높이	범주형
TpgrphFrmCodeNm	토지의 지형 형상	범주형
RoadSideCodeNm	토지의 도로접면	범주형
PblntfPclnd	토지의 공시지가	연속형

변수명	설명	형식
Tot_ppltn	해당 법정동의 총 인구	연속형
Avg_age	해당 법정동의 평균 나이 (단위: 세)	연속형
Ppltn_dnsty	해당 법정동의 인구 밀도 (단위: 명/km²)	연속형
Aged_child_idx	해당 법정동의 노령화지수 (단위: 일백명당 명)	연속형
Oldage_suprt_per	해당 법정동의 노년부양비 (단위: 일백명당 명)	연속형
Juv_suprt_per	해당 법정동의 유년부양비 (단위: 일백명당 명)	연속형
Tot_family	해당 법정동의 총가구 수	연속형
Avg_fmember_cnt	해당 법정동의 평균 가구원 수	연속형
Tot_house	해당 법정동의 총 주택 수	연속형
Corp_cnt	해당 법정동의 종업원 수 (전체 사업체)	범주형
Tot_worker	해당 법정동의 사업체 수 (전체 사업체)	범주형
PrposAreaDstrcCodeNm	해당 법정동의	범주형
ChangePricePercent	해당 월 지가 변동률	범주형
ChangePricePercentTotal	해당 월부터 2022년 1월까지의 총 변동률	범주형
ChangePricePercentAveerage	해당 월부터 2022년 1월까지의 평균 지가 변동률	연속형

3. 연속형 변수 및 범주형 변수 분석

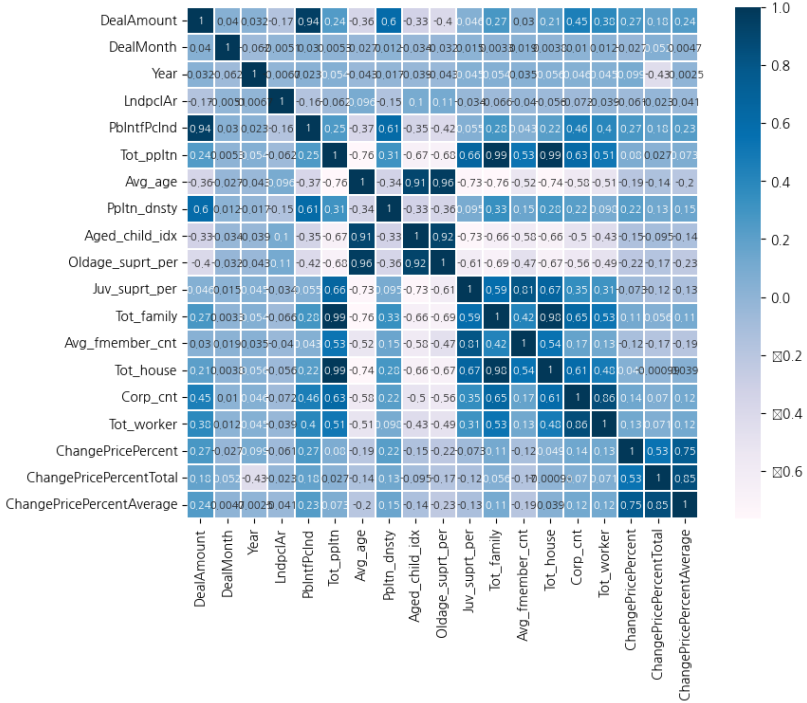
연속형 변수 목록

DealAmount	DealMonth	Year	LndpclAr	PblntfPclnd	Tot_ppltn	Avg_age	Ppltn_dnsty	Aged_child_idx	Oldage_suprt_per	Juv_suprt_per
6114.18	1	2021	21.40	4738.14	5929.00	48.60	2454.70	542.90	31.70	5.80
3097.25	1	2021	5.70	2537.72	4828.00	44.90	8010.40	532.60	25.00	4.70
5547.00	1	2021	5.20	3339.16	7453.00	42.00	9533.10	297.80	20.80	7.00
5363.12	1	2021	67.10	3339.16	7453.00	42.00	9533.10	297.80	20.80	7.00
4253.85	2	2021	2.10	3327.16	5929.00	48.60	2454.70	542.90	31.70	5.80

Tot_family	Avg_fmember_cnt	Tot_house	Corp_cnt	Tot_worker	ChangePricePercent	ChangePricePercentTotal	ChangePricePercentAverage
3636.00	1.60	1630.00	18586.00	118650.00	0.34	4.54	0.38
2692.00	1.70	1270.00	8343.00	25067.00	0.37	4.95	0.41
3883.00	1.80	2486.00	1693.00	17857.00	0.37	4.94	0.41
3883.00	1.80	2486.00	1693.00	17857.00	0.37	4.94	0.41
3636.00	1.60	1630.00	18586.00	118650.00	0.34	4.54	0.38

3. 연속형 변수 및 범주형 변수 분석

연속형 변수의 상관계수 분석

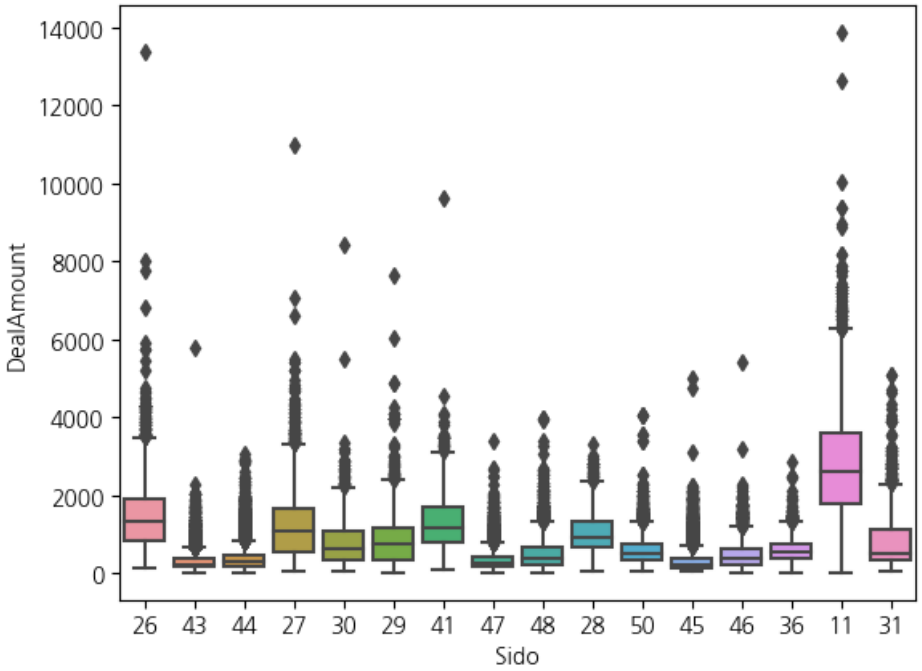


- DealAmount(실거래가)를 예측할 것이기 때문에 DealAmount에 해당하는 상관계수를 파악

- PblntfPclnd(공시지가)와 Ppltn_dnsty(인구밀도)가 가장 높은 상관계수를 나타냄

3. 연속형 변수 및 범주형 변수 분석

범주형 변수(시도) 분석

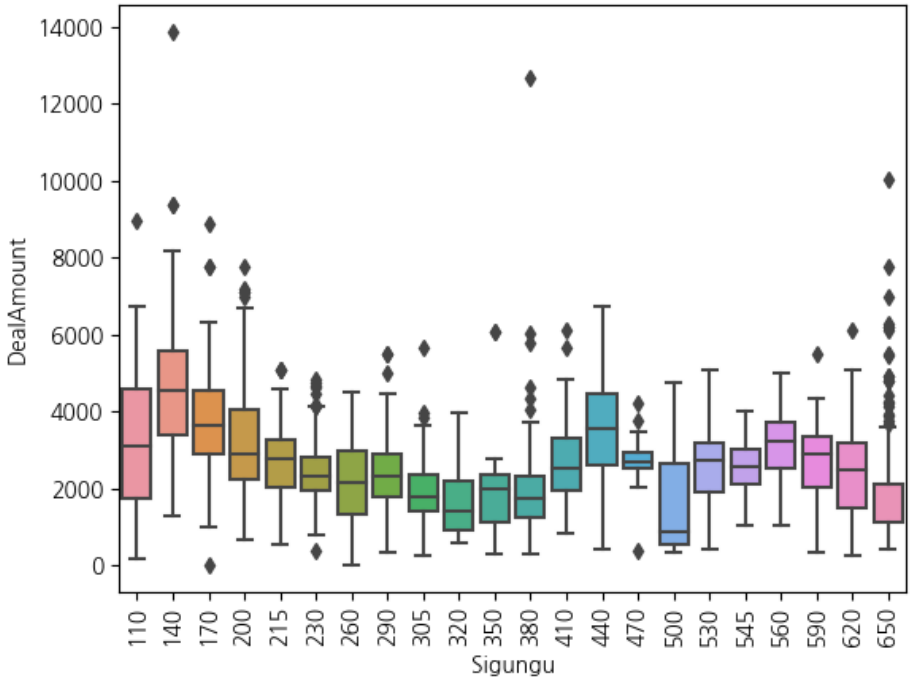


- 지역별 가격 편차가 크게 존재함

- 서울특별시(11)의 토지 가격이 가장 높게 나타남

3. 연속형 변수 및 범주형 변수 분석

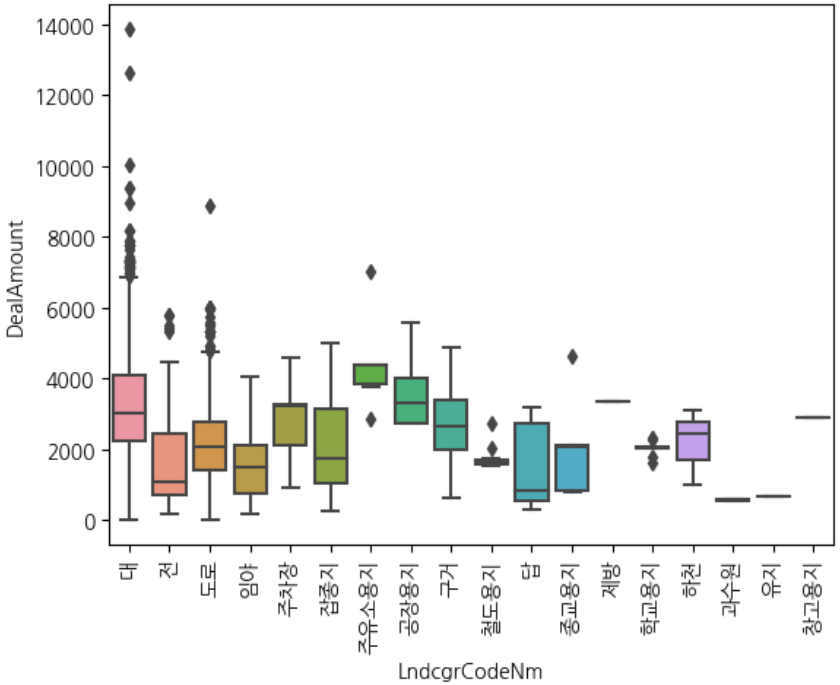
범주형 변수(시군구) 분석



- 서울특별시에 대한 시군구 박스 플롯

- 시군구별로도 실거래 가격이 상이함을 알 수 있음

범주형 변수(지목) 분석

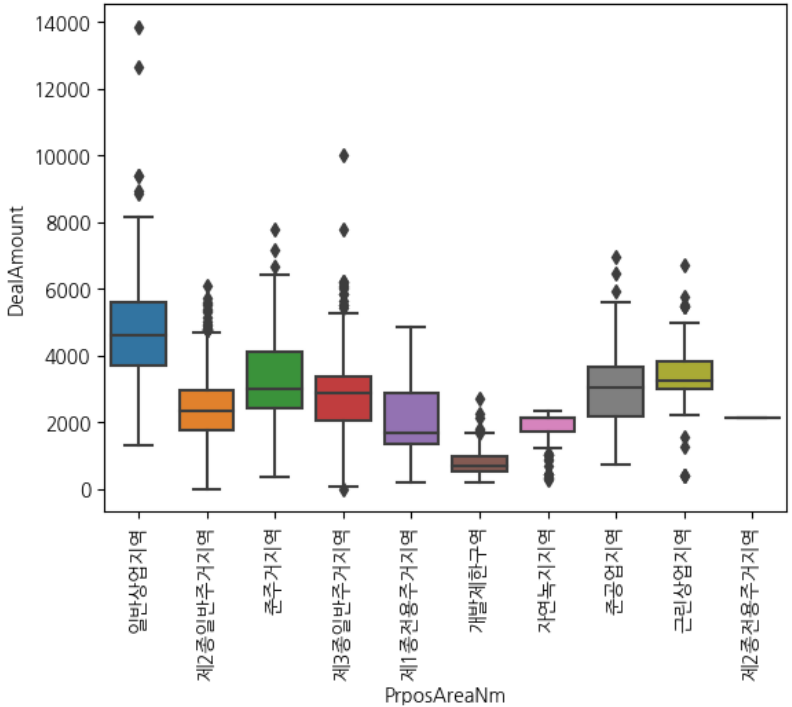


- 지목이 '대'인 토지 가격은 분산이 크지만 '대'를 제외한 대부분의 지목 가격 분산은 상대적으로 작음

- 평균 가격 차이가 유의미하게 있어 지목이 토지 가치 예측에 중요한 특징으로 작용할 수 있다고 봄

3. 연속형 변수 및 범주형 변수 분석

범주형 변수(용도지역) 분석



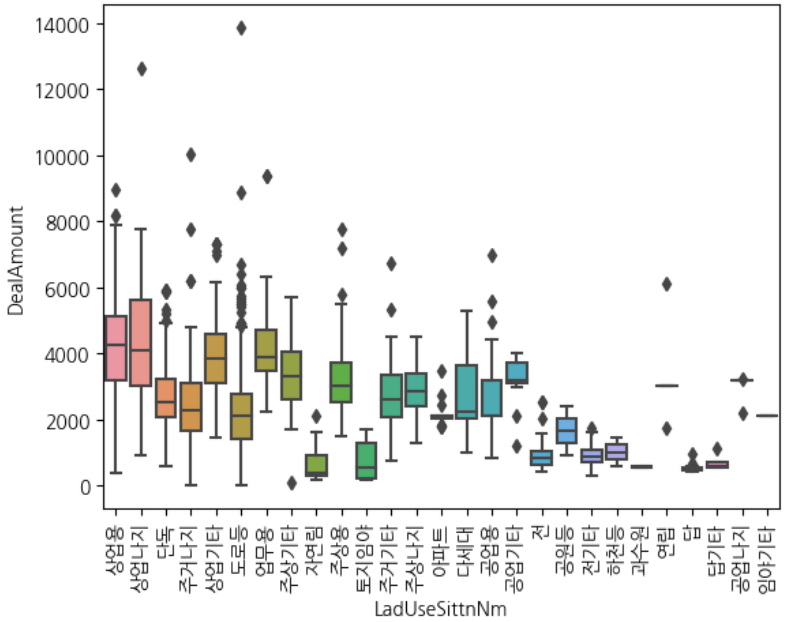
- 용도지역별 평균 매매 가격 차이가 남

- 일반상업지역의 가격 분포가 매우 큼

- 개발제한구역의 토지에는 거래 가격이 매우 낮음

3. 연속형 변수 및 범주형 변수 분석

범주형 변수(이용상황) 분석



- 자연림이나 토지임야의 경우 가격이 매우 낮음

- 상업용, 상업나지의 경우 가격이 높음

연속형 및 범주형 변수 분석 결과

- 그 외에도 토지 형상, 지세, 도로접면 등의 데이터에도 약간의 가격 차이를 나타냄

- 전국 모델과 개별 모델을 모두 학습하여 어느 방식으로 모델을 구성하는 것이 가격 예측에 효과적일지 알아봄

4. 결측치 제거

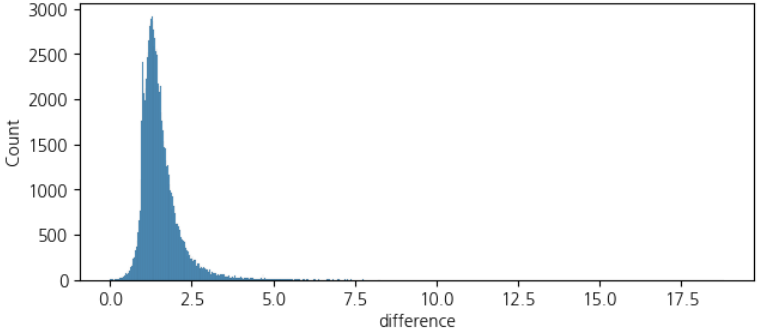
DealAmount	0
DealMonth	0
Year	0
LndpclAr	0
PblntfPclnd	0
Tot_ppltn	6315
Avg_age	6315
Ppltn_dnsty	6315
Aged_child_idx	6315
Oldage_suprt_per	6315
Juv_suprt_per	6315
Tot_family	6315
Avg_fmember_cnt	6315
Tot_house	6315
Corp_cnt	6315
Tot_worker	6315
ChangePricePercent	5
ChangePricePercentTotal	5
ChangePricePercentAverage	5
dtype: int64	

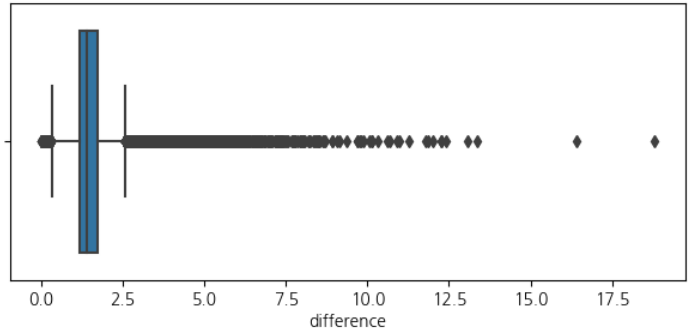
DealAmount	0
DealMonth	0
Year	0
LndpclAr	0
PblntfPclnd	0
Tot_ppltn	0
Avg_age	0
Ppltn_dnsty	0
Aged_child_idx	0
Oldage_suprt_per	0
Juv_suprt_per	0
Tot_family	0
Avg_fmember_cnt	0
Tot_house	0
Corp_cnt	0
Tot_worker	0
ChangePricePercent	0
ChangePricePercentTotal	0
ChangePricePercentAverage	0
dtype: int64	

- 수집한 데이터셋에서 누락된 데이터가 있는지 확인 후 누락된 행을 제거함

- 결측치를 제거함으로써 데이터셋의 완전성을 유지하고 모델의 안정성 확보

5. 이상치(Outlier) 제거

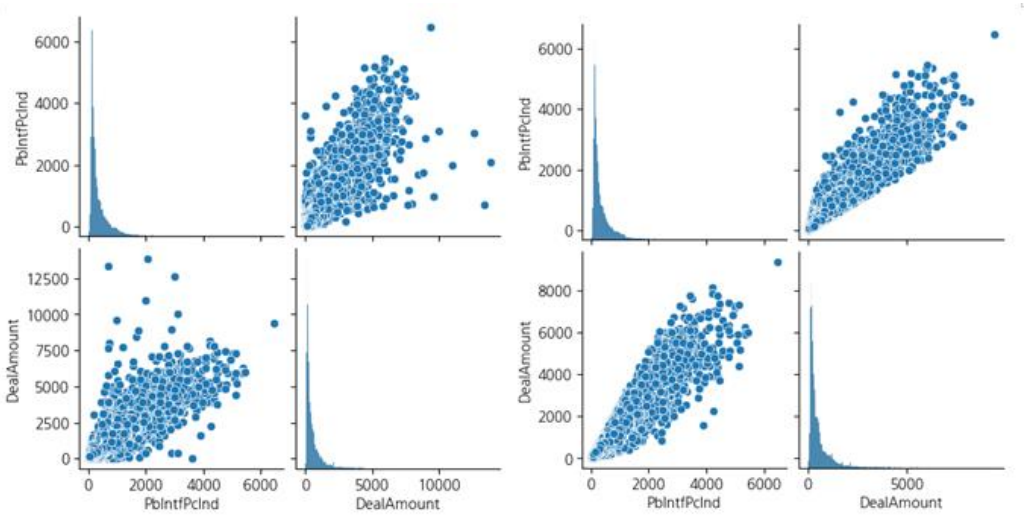




- 연속형 데이터들이 모두 정규분포를 따르지 않음을 확인했기 때문에, Z-Score 외에 다른 이상치 제거 방식을 고안함

- 실거래가를 공시지가로 나눈 값의 사분범위(InterQuartile Range)를 확인하여, 1.5*IQR 방식으로 이상치를 선정해 제거함

5. 이상치(Outlier) 제거



- 이상치를 검출해 제거해 준 후의 데이터 분포도를 산점도 그래프를 통해 나타냄

[왼쪽: 제거 전 / 오른쪽: 제거 후]

- 위 방식으로 이상치를 선정할 경우 과도하게 공시지가 값에 의존할 가능성이 생길 수 있음

- 현재 주어진 특성들을 가지고 최대한 좋은 성능의 모델을 생성해내고자 하였기에 이러한 점은 무시하였음

6. 원-핫 인코딩(One-Hot Encoding)

원-핫 인코딩이란?

- 각 범주형 변수를 이진 형태로 변환하여 모델이 범주 간의 관계를 이해할 수 있도록 함

- 이전에 분류한 범주형 데이터들을 원-핫 인코딩 하여 모델 학습에 적합한 형태로 변환함

```
Index(['DealAmount', 'DealMonth', 'Year', 'LndpclAr', 'PblntfPclnd',
       'Tot_ppltn', 'Ppltn_dnsty', 'Juv_suprt_per', 'Tot_family',
       'Avg_fmember_cnt',
       ...
       'RoadSideCodeNm_맹지', 'RoadSideCodeNm_세로각지(가)', 'RoadSideCodeNm_세로각지(불)',
       'RoadSideCodeNm_세로한면(가)', 'RoadSideCodeNm_세로한면(불)',
       'RoadSideCodeNm_소로각지', 'RoadSideCodeNm_소로한면', 'RoadSideCodeNm_중로각지',
       'RoadSideCodeNm_중로한면', 'RoadSideCodeNm_지정되지않음'],
      dtype='object', length=201)
```

- 범주형 데이터에 관한 One-Hot Encoding 작업 후에 생성된 변수

7. 데이터 분할

- 전처리 후 데이터셋의 70%를 학습에 사용하고, 나머지 30%를 모델의 성능 평가용으로 사용

- 데이터 분할은 무작위로 이루어지고, 학습데이터와 테스트 데이터 간 중복이 없음

- 결과는 재현될 수 있어야 함으로 시드값을 부여해 데이터셋을 분할함

8. XGBoost 모델 생성

- XGBoost(Extreme Gradient Boosting)는 트리 기반의 앙상블 학습 알고리즘

- 복잡한 데이터 패턴을 학습하고 예측하기에 효과적

- 토지 가격은 지역별로 상이하기에, 시도 변수을 사용하는 전국 모델과 시군구, 읍면동 변수를 사용하는 개별 지역 모델로 나누어 성능을 평가함

9. 하이퍼파라미터 튜닝

- 생성한 모델의 하이퍼파라미터 값을 변경하며 가장 좋은 예측 성능을 갖도록 함

변수	설명	값
learning_rate	학습 단계별 이전 결과 반영률	0.22
n_estimators	트리 모델의 개수	350
max_depth	트리의 최대 깊이	9
min_child_weight	child에서 필요한 모든 관측치에 대한 가중치의 최소합	1
gamma	트리에서 추가로 가지를 나눌지 말지 결정하는 최소 손실 감소값	0
subsample	각 트리별 데이터 샘플링 비율	0.91
colsample_bytree	각 트리별 feature 샘플링 비율	0.89
lambda	L2 규제	10
alpha	L1 규제	5
eval_metric	평가 척도	rmse
seed	시드값 고정	0

평균 제곱근 오차(RMSE)

```
26 ) RMSE score: 1491296.7160186877
44 ) RMSE score: 145192.1979993734
27 ) RMSE score: 1229925.8438525125
30 ) RMSE score: 413044.5705307837
29 ) RMSE score: 480700.43992318923
41 ) RMSE score: 1094569.7878177266
28 ) RMSE score: 575244.2040949115
45 ) RMSE score: 139795.41812162183
46 ) RMSE score: 192163.71745491488
36 ) RMSE score: 248377.6482810492
11 ) RMSE score: 3849554.880098103
31 ) RMSE score: 759229.5475266407
```

- RMSE는 예측값과 실제값 간의 차이를 평가하는 지표로, 모델의 예측 정확도를 확인할 수 있음

- RMSE 값이 작을수록 모델의 예측이 실제와 가까움을 나타냄

- 서울특별시(11)의 RMSE가 380만으로 가장 높았고, 전라북도(45)가 13만으로 가장 낮음

평균 제곱근 오차(RMSE) – 전국 모델과 개별 모델 비교

지역	전국 모델 RMSE	개별 모델 RMSE
서울	3,849,554	3,945,353
부산	1,491,296	1,690,463
대구	1,229,925	1,329,107
광주	480,700	608,289
대전	413,044	417,063
울산	759,229	715,394
경기	1,094,569	1,007,961
충남	145,192	133,489
전북	139,795	137,912
전남	192,163	225,712

- RMSE는 예측값과 실제값 간의 차이를 평가하는 지표로, 모델의 예측 정확도를 확인할 수 있음

- RMSE 값이 작을수록 모델의 예측이 실제와 가까움을 나타냄

- 서울특별시(11)의 RMSE가 380만으로 가장 높았고, 전라북도(45)가 13만으로 가장 낮음

10. 모델 평가

예측 오차 계산

	sido	pred	real	ratio(%)
2544	45	29927.91	29929	0.00
13919	29	2016566.75	2016400	0.01
11724	36	776234.88	776161	0.01
6109	36	99240.69	99225	0.02
5474	44	81209.54	81225	0.02
...	...	...	...	...
3	45	27859.10	1444	1829.30
58	36	122626.35	5776	2023.03
24	30	69864.97	3025	2209.59
2038	30	628556.88	25600	2355.30
7959	11	4779583.50	188356	2437.53

[15885 rows x 4 columns]
ratio < 3: 1430 / 15885 (9.00%)
ratio < 5: 2291 / 15885 (14.42%)
ratio < 10: 4205 / 15885 (26.47%)

- 예측 오차는 모델이 예측한 값과 해당 토지의 실제 거래가를 비교하여 계산

- 예측값을 실제 거래가로 나눈 값에 절댓값을 취하고, 100을 곱해 백분위로 나타냄

- 예측 오차 10% 이내 값의 비중을 주요 모델 평가 지표로 삼았음

10. 모델 평가

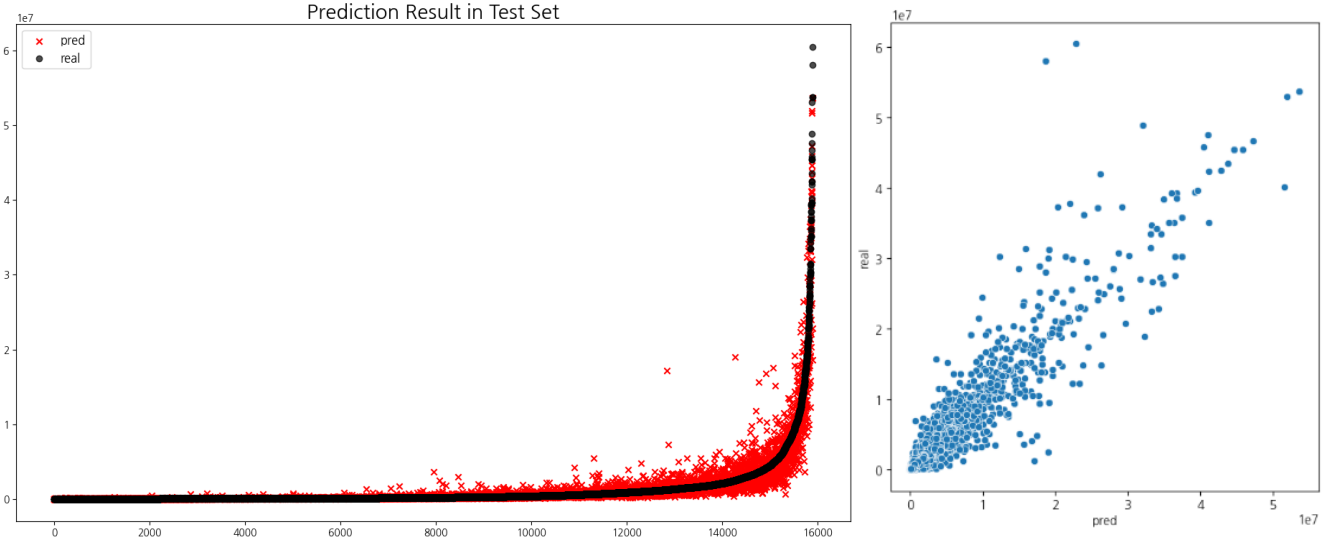
예측 오차 계산 – 전국 모델과 개별 모델 비교

지역	전국 모델 최대 오차율	전국 모델 예측 오차 10% 이내	개별 모델 최대 오차율	개별 모델 예측 오차 10% 이내
서울	1827.81%	41.50%	1648.19%	42.33%
부산	1208.82%	29.89%	10259.76%	33.85%
대구	1053.57%	30.61%	1156.48%	33.29%
광주	712.49%	28.56%	1067.13%	35.05%
대전	3074.86%	27.26%	533.21%	38.99%
울산	791.14%	31.28%	968.39%	37.40%
경기	673.83%	35.58%	855.33%	40.56%
충남	1973.33%	25.28%	2487.28%	32.36%
전북	1127.73%	23.21%	561.52%	33.23%
전남	1893.42%	21.97%	2704.59%	32.33%

- 전국 모델보다 개별 모델의 예측 오차가 더 낮은 모습을 확인할 수 있음

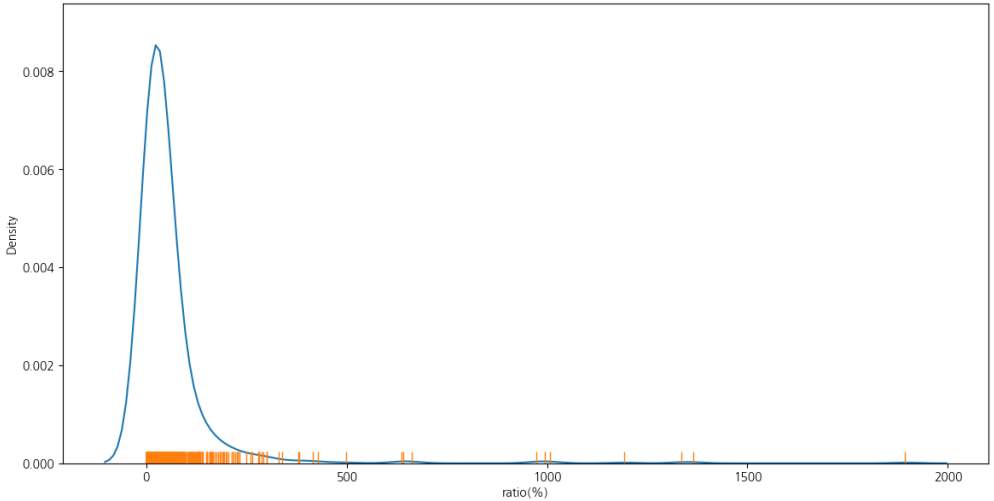
- 최대 오차율의 차이는 이상치를 완전히 제거해내지 못해 모델이 예측 과정에서 영향을 받았다고 예상

모델 학습 결과



- 전반적으로 비슷하게 가격 예측을 하지만, 특정 토지에 대해서 오차가 크게 발생하는 것을 확인할 수 있음

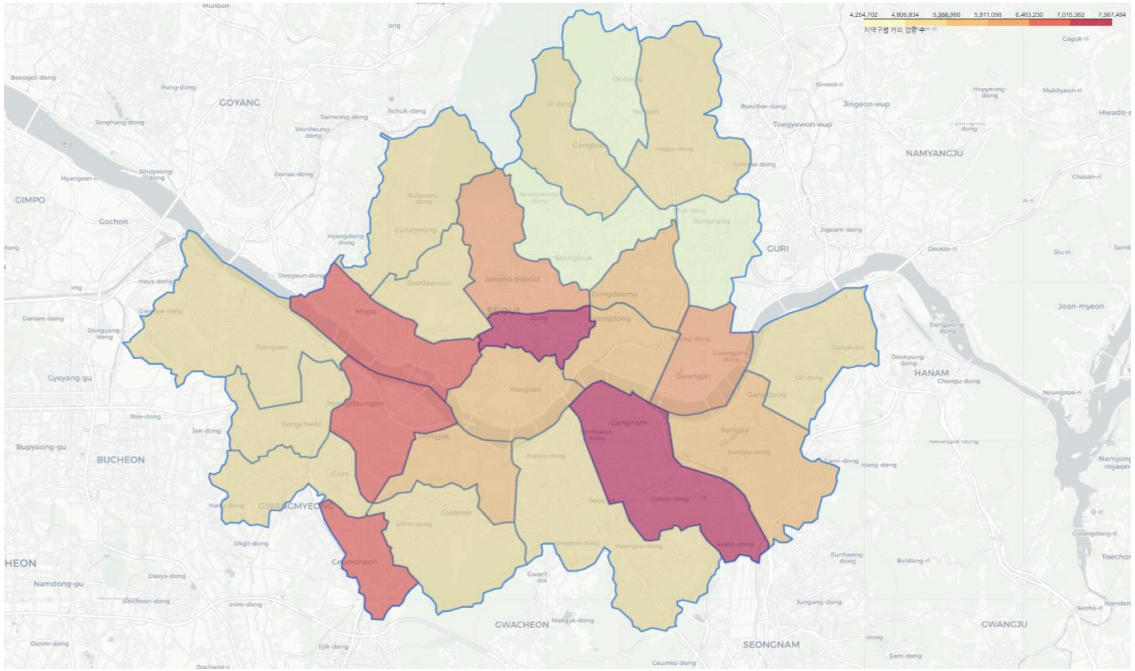
모델의 예측 오차



- 커널 밀도 추정 그래프 및 1차원 실수 분포 플롯을 사용하여 시각화 함

- 대개 200% 이내로 예측을 하지만, 1,000%를 넘어가는 데이터도 몇몇 존재함

Folium을 활용한 토지 가치 시각화



- 서울의 중심부로 갈수록 가격이 높아지고, 강남구, 마포구 등의 가격이 높게 책정됨

03.
결과 활용 및 시사점

1. 결과 활용

1. 부동산 투자

개발한 모델을 활용하여 부동산 투자자들은 토지 가치를 더 정확하게 예측하고 투자 리스크를 최소화할 수 있음

2. 도시 계획

도시 계획자들은 모델을 활용하여 도시의 발전 방향을 결정하고 지속 가능한 도시 발전을 지원할 수 있음

3. 부동산 시장 투명성

정확한 토지 가치 예측은 부동산 시장의 투명성을 증가시키고 시장의 건전성을 유지하는데 도움이 됨

4. 기타 분야 활용

개발한 모델은 다른 분야에서도 활용 가능하며, 예를 들어 자산 관리, 도시 인프라 개선, 부동산 투자 컨설팅 등에 활용 가능함

모델 개선과 확장

- 향후 모델의 정확도를 높이기 위해 데이터 수집과 전처리 과정을 보완하고, 더 많은 변수와 데이터를 활용하여 모델을 확장하는 것이 중요함

- 현재의 모델은 국가에서 지정한 공시지가 값에 너무 많이 의존하는 경향이 크기 때문에, 위와 같은 문제를 해결할 필요성이 있음

- 통계청 SGIS 오픈플랫폼에서 제공하는 가장 최신 데이터가 2021년도이고, 인구 밀도와 사업체 통계의 상관계수가 나름 높았다는 점을 생각해보면, 인구 통계 및 사업체 통계 데이터의 최신화도 중요하게 작용할 것으로 예상됨

시스템 구축

- 모델을 실제 응용에 적용하기 위해 웹 기반 시스템을 구축하여 관련 이해 관계자들이 쉽게 접근하고 활용할 수 있도록 함

- 사용자로 하여금 토지 예측 가격을 직관적으로 알 수 있고, 토지 거래 및 투자에 도움이 될 수 있음

감사합니다.