아파트 실거래가 예측 프로젝트

목차

01 → 02 → 03

개요 및 분석목적

- 대회 설명
- 프롭테크 시장
- 구매자 보호

데이터 분석

- 활용데이터
- 파생변수 생성
- 기타전처리
- 최종전처리

최종결론

- 모델링 결과

개요 및 분석목적

개요 및 분석목적

1) 대회 설명

Train 데이터
서울/부산의
2008~2017년 거래데이터

↓

Test 데이터
서울/부산의
2018년 거래데이터

⇨

2018년의
아파트 실거래가 예측

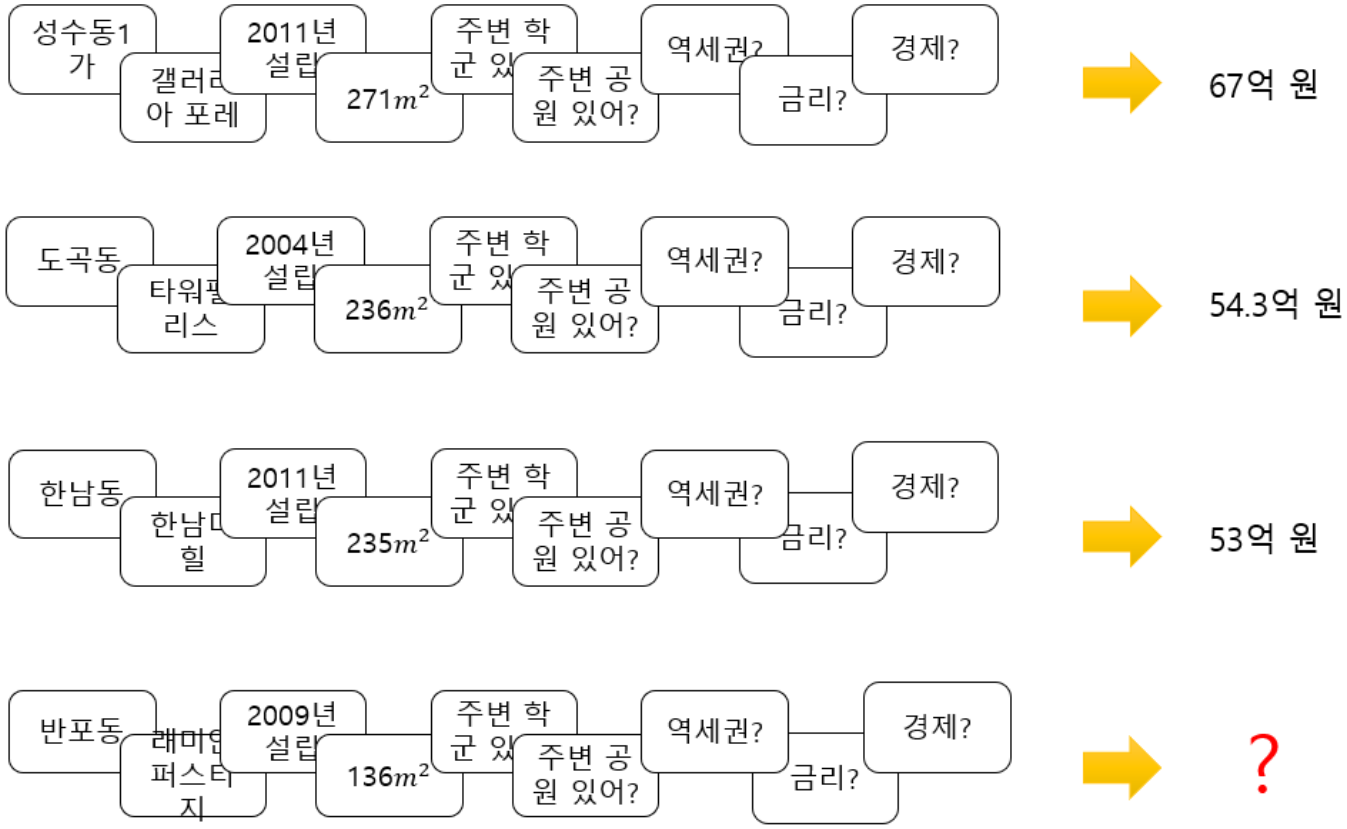
개요 및 분석목적

2) 프롭테크 시장

직방시세®

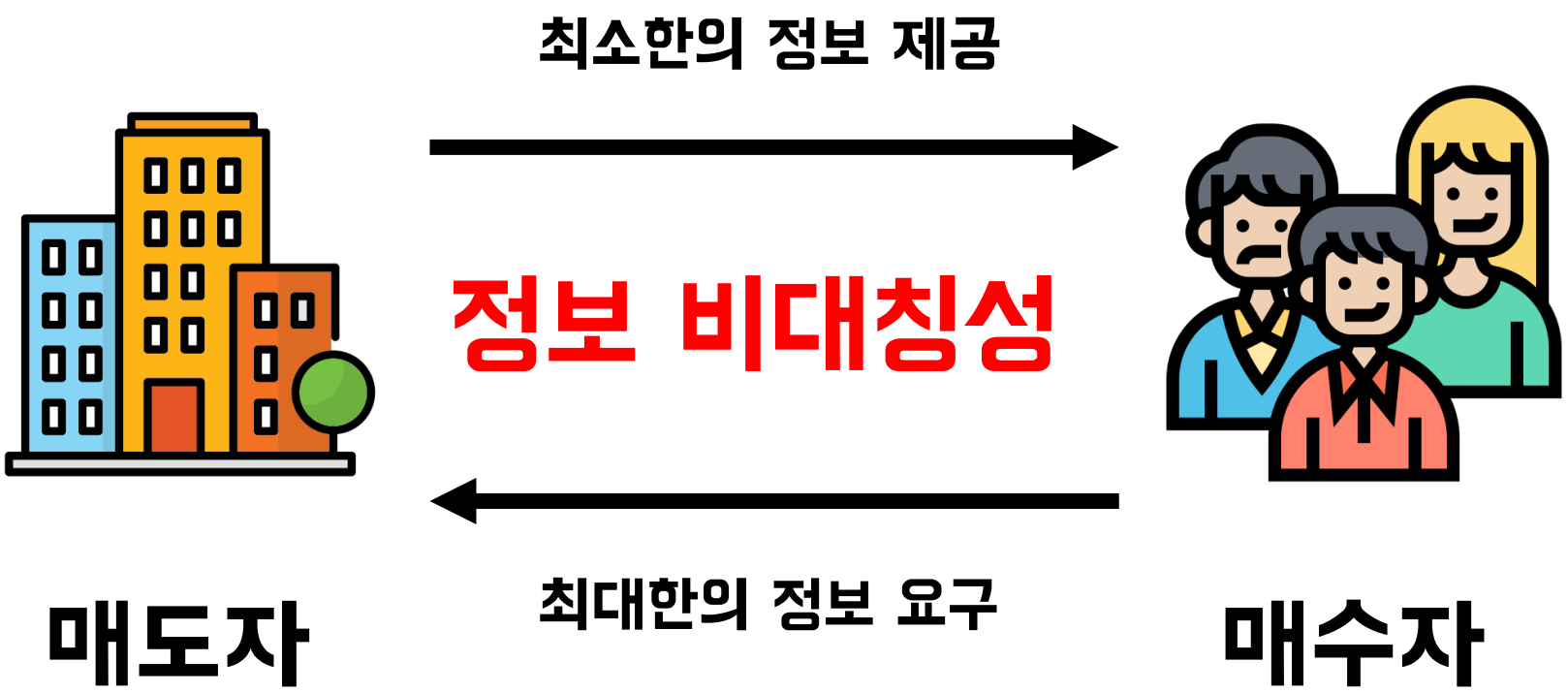
매매 66억 9,000
1억 274 / 3.3㎡

전세 51억 3,000
7,878만 / 3.3㎡



개요 및 분석목적

3) 구매자 보호



데이터 분석

데이터 분석

1) 활용 데이터

제공 데이터

서울/부산 지역의 아파트 거래 데이터(train, test)
서울/부산 지역의 공원에 대한 정보
서울/부산 지역의 어린이집에 대한 정보

추가 데이터

공공데이터 포털 서울/부산 자치구별 법정동 데이터
카카오 오픈 API
아파트 브랜드 평판지수

데이터 분석

2) 파생변수 생성 - 평당가

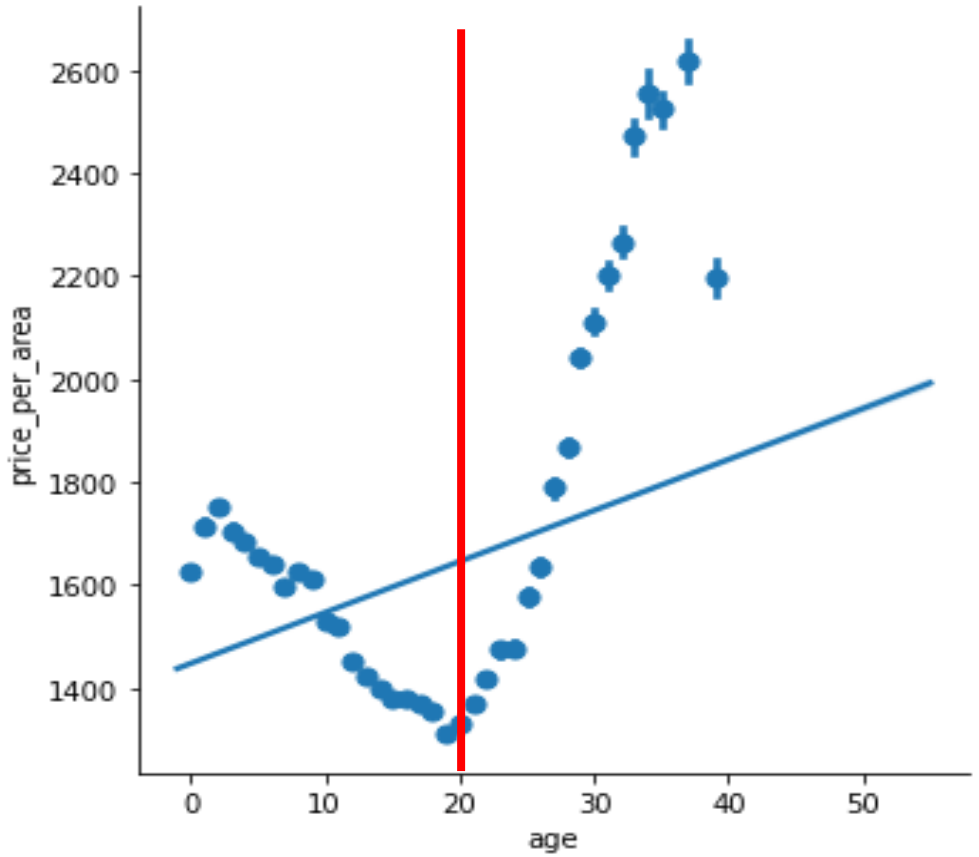
apartment_id	city	dong	jibun	apt	addr_kr	exclusive_use_area	year_of_completion	transaction_year_month	transaction_date	floor	transaction_real_price
4927	서울특별시	신당동	407-17	벨레어카운티	신당동 407-17 벨레어카운티	273.820	2004	200904	21~30	5	95000
10308	서울특별시	청담동	134-38	청담자이	청담동 134-38 청담자이	49.619	2012	201609	1~10	6	95000

같은 선상에 있다고 할 수 있을까?

```
train['price_per_area'] = train['transaction_real_price'] / train['exclusive_use_area']*3.3
```

데이터 분석

2) 파생변수 생성 - 나이 & 재건축여부



아파트 나이 = 거래년도 - 완공년도

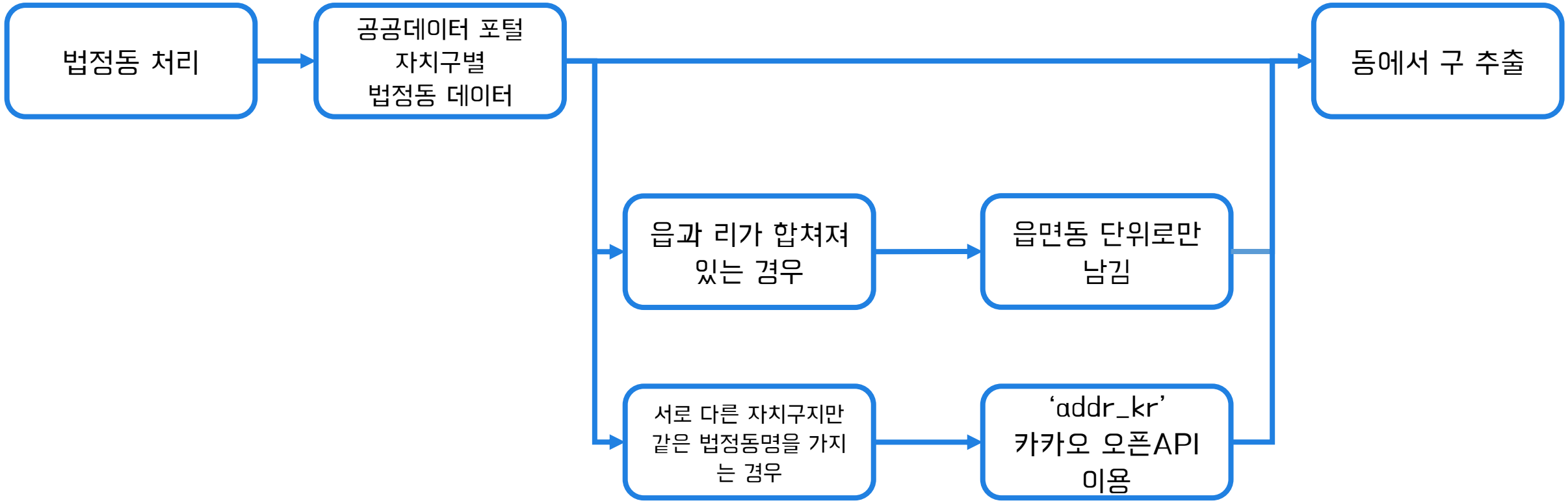
아파트 나이가 20년이 넘으면
재건축 대상 아파트이기 때문

```
train['year'] = train['transaction_year_month'].astype(str).str[:4].astype(int)
train['month'] = train['transaction_year_month'].astype(str).str[4:].astype(int)
train['age'] = train['year'] - train['year_of_completion']

train.loc[train['age'] < 20, 'reconstruction'] = 'NO'
train.loc[train['age'] >= 20, 'reconstruction'] = 'YES'
```

데이터 분석

2) 파생변수 생성 - 법정구



데이터 분석

2) 파생변수 생성 - 아파트 브랜드 평판

순위	아파트 브랜드	기업	참여지수	미디어지수	소통지수	커뮤니티지수	브랜드평판지수
1	힐스테이트	현대건설	956,750	779,369	1,188,274	1,101,454	4,025,847
2	푸르지오	대우건설	465,470	584,635	812,832	1,235,285	3,098,221
3	자이	GS건설	445,000	750,467	919,071	860,594	2,975,132
4	더샵	포스코건설	324,850	735,582	842,155	832,304	2,734,890
5	롯데캐슬	롯데건설	387,417	329,377	770,907	1,156,269	2,643,970
6	래미안	삼성물산	231,489	362,514	884,131	1,005,582	2,483,715
7	SK뷰	SK건설	103,329	169,352	868,120	1,333,163	2,473,964
8	e편한세상	DL이앤씨	190,015	319,351	425,039	1,074,693	2,009,098
9	더 플래티넘	쌍용건설	113,564	102,907	780,019	915,581	1,912,071
10	두산 위브	두산건설	140,798	160,851	652,687	657,367	1,611,703
11	우미린	우미건설	334,640	162,352	310,491	252,431	1,059,914
12	서희스타힐스	서희건설	273,853	134,556	200,668	207,129	816,206
13	한화포레나	한화건설	271,005	167,515	266,544	103,621	808,685
14	호반베르디움	호반건설	152,368	102,162	243,640	274,371	772,542
15	한라비발디	한라건설	177,822	99,497	224,975	213,121	715,415
16	하늘채	코오롱글로벌	61,766	89,698	269,639	282,295	703,398
17	아이파크	현대산업개발	36,107	133,024	159,526	365,256	693,913
18	코아루	한국토지신탁	59,452	54,306	272,457	239,598	625,813
19	센트레빌	동부건설	80,723	119,339	168,167	192,377	560,607
20	데시앙	태영건설	74,226	58,617	169,645	199,009	501,497
21	스위첸	KCC건설	36,045	68,248	141,200	153,818	399,310
22	리슈빌	계룡건설	44,767	36,373	145,899	164,201	391,240
23	벽산블루밍	벽산건설	89,890	46,538	103,314	103,748	343,491
24	동문굿모닝힐	동문건설	54,023	33,541	76,909	101,687	266,160

top_branded 컬럼 생성

```
brand_ranked = ['힐스테이트', '푸르지오', '자이', '더샵', '롯데캐슬', '래미안',
                '아이파크', 'e편한세상', '위브', '한화포레나', '우미린', 'SK뷰',
                '호반베르디움', '서희스타힐스', '한라비발디', '하늘채',
                '더플래티넘', '코아루', '센트레빌', '데시앙', '스위첸', '리슈빌',
                '벽산블루밍', '동문굿모닝힐']
train['top_branded'] = 0
for brand in brand_ranked:
    train['top_branded'].loc[train['apt'].str.contains(brand)] = 1
```

데이터 분석

3) 기타 전처리

거래연도별 데이터수(거래수)

```
[ ]  1 train.groupby('transaction_year').size()
```

```
transaction_year
2008    100066
2009    127869
2010    102823
2011     98586
2012     70065
2013    109738
2014    136649
2015    181195
2016    165664
2017    123898
dtype: int64
```

거래연도별 평당가 평균

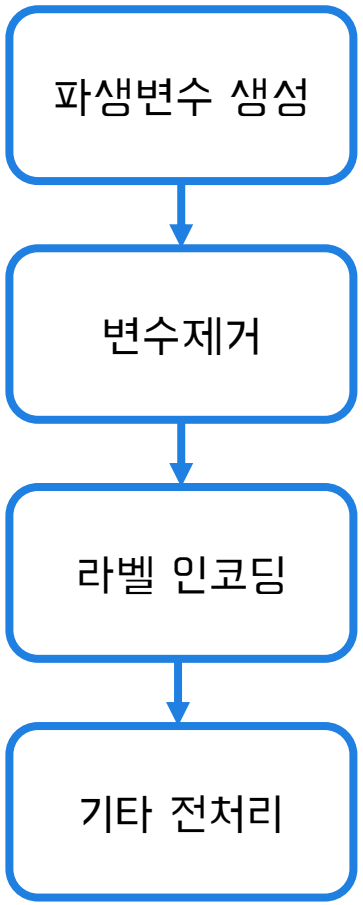
```
[ ]  1 train.groupby('transaction_year')['price_per_area'].mean()
```

```
transaction_year
2008    1227.554931
2009    1463.038274
2010    1277.597088
2011    1446.475808
2012    1462.791173
2013    1481.783184
2014    1538.405981
2015    1667.018121
2016    1847.773204
2017    2182.478714
Name: price_per_area, dtype: float64
```

2014년 이후의 데이터만 사용

데이터 분석

4) 최종 전처리 과정



법정구, 거래연도, 법정구별 평당 거래가 평균, 법정동별 평당 거래가 평균, 공원 수, 어린이집 수, 브랜드 평판 top 24 여부, 아파트 나이, 재건축 여부

아파트아이디, 동, 지번, 주소, 거래년월, 거래날짜, 층

도시명, 아파트명, 법정구, 재건축여부

이상치 제거, 실거래가 로그변환, 2014년 이후 데이터

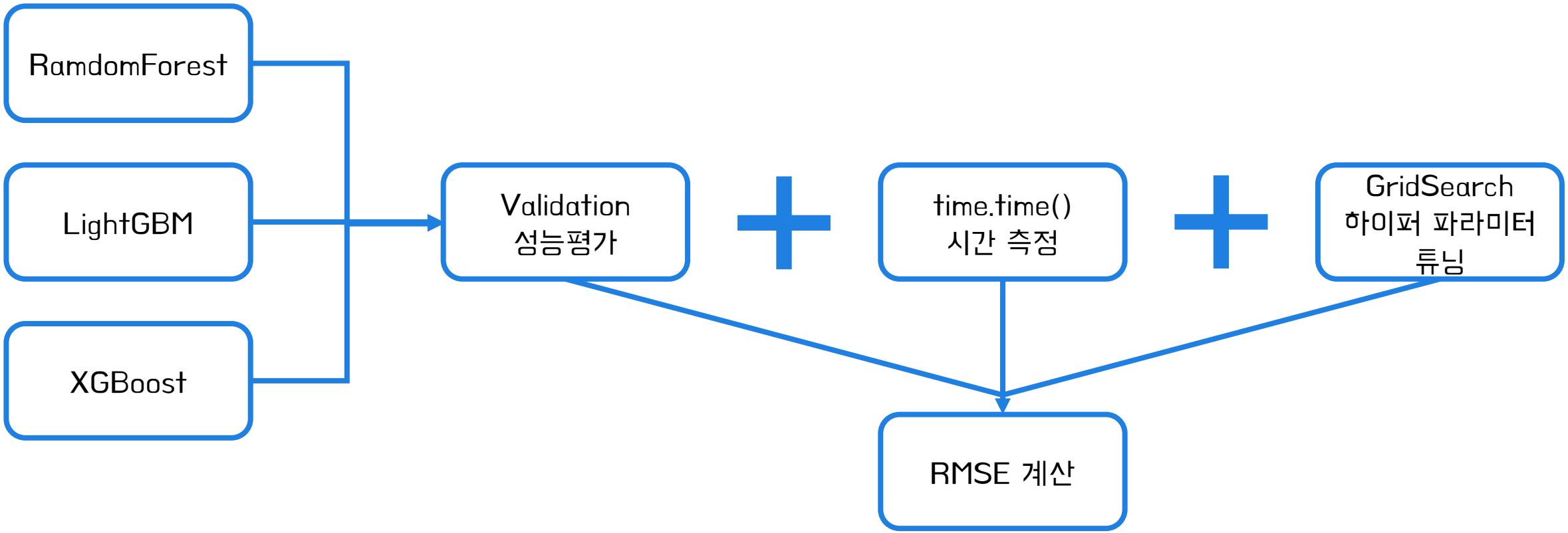
1216553 X 15

최종 COLUMNS

도시명(서울, 부산)
아파트명
전용면적
완공연도
법정구
평당가
건축연도
법정구별 평당 거래가 평균
법정동별 평당 거래가 평균
공원 수
어린이집 수
브랜드 평판 top 24 여부
아파트 나이
재건축 여부
실거래가

데이터 분석

5) 모델링


RamdomForest
LightGBM
XGBoost
Validation
성능평가
time.time()
시간 측정
GridSearch
하이퍼 파라미터
튜닝
RMSE 계산

최종 결론

최종 결론

1) 분석모델 선정

모델	데이터 추가 및 추가 전처리 전	데이터 추가 및 추가 전처리 후	2014년 이후 데이터로 모델링	하이퍼 파라미터 최적화 후
RamdomForest	0.2133	0.0979	0.0855	0.0800
LightGBM	0.3235	0.1650	0.1504	0.0794
XGBoost	0.3445	0.1936	0.1778	0.0839

모델	최종성능	시간(초)
RamdomForest	0.0800	90
LightGBM	0.0794	104
XGBoost	0.0839	652

LightGBM

RMSE

0.3235 → 0.0794


감사합니다