Predicting condominium price in Bangkok

Summarizing and updating the prior literature (Jirapon S. & Sarawut R., 2019, Predicting Condominium price in Bangkok using web mining techniques)

2023.5.17.

Yaowamal Luetrakulset

Introduction

- 방콕에서 콘도미니엄은 편리한 생활을 선호하는 많은 사람들에게 **첫 번째 집이 되고 있다**
- 아파트 가격은 구매자의 의사결정에 **가장 중요한 요소 중 하나입니다**
- 방콕 아파트 **평방미터당 매매가격** 예측모델을 만들었습니다

Attribute

- https://www.hipflat.co.th/에서 데이터를 수집힙니다
- 17 속성(x), 1 예측 값(y)
- **'Title', 'nearest_bts', 'nearest_mrt'** 변수가 범주형 변수인 것을 제외한 모든 속성이 숫자 변수입니다

Column	Description
Title	Name of condominium
Year_built	Year of condominium was built
No_floor	Total number of floors of condominium
Nearest_bts	The name of nearest BTS station
Dist_bts	Distance between the nearest BTS station and condominium
Nearest_mrt	The name of nearest MRT station
Dist_mrt	Distance between the nearest MRT station and condominium
Price_per_sqm	Condominium price per square
Price_chg_prev_quart	Condominium price change from last quarter
Price_chg_from_lastyr	Condominium price change from last year
Yield_amt	Condominium achievable gross rental yield
Rental_chg_from_lastyr	Condominium rent price change from last year
Price_for_sale	Condominium price for sale
Price_for_rent	Condominium price for rent
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms
Internal_area	Size of internal area
Tower	Number of towers

Data preprocessing

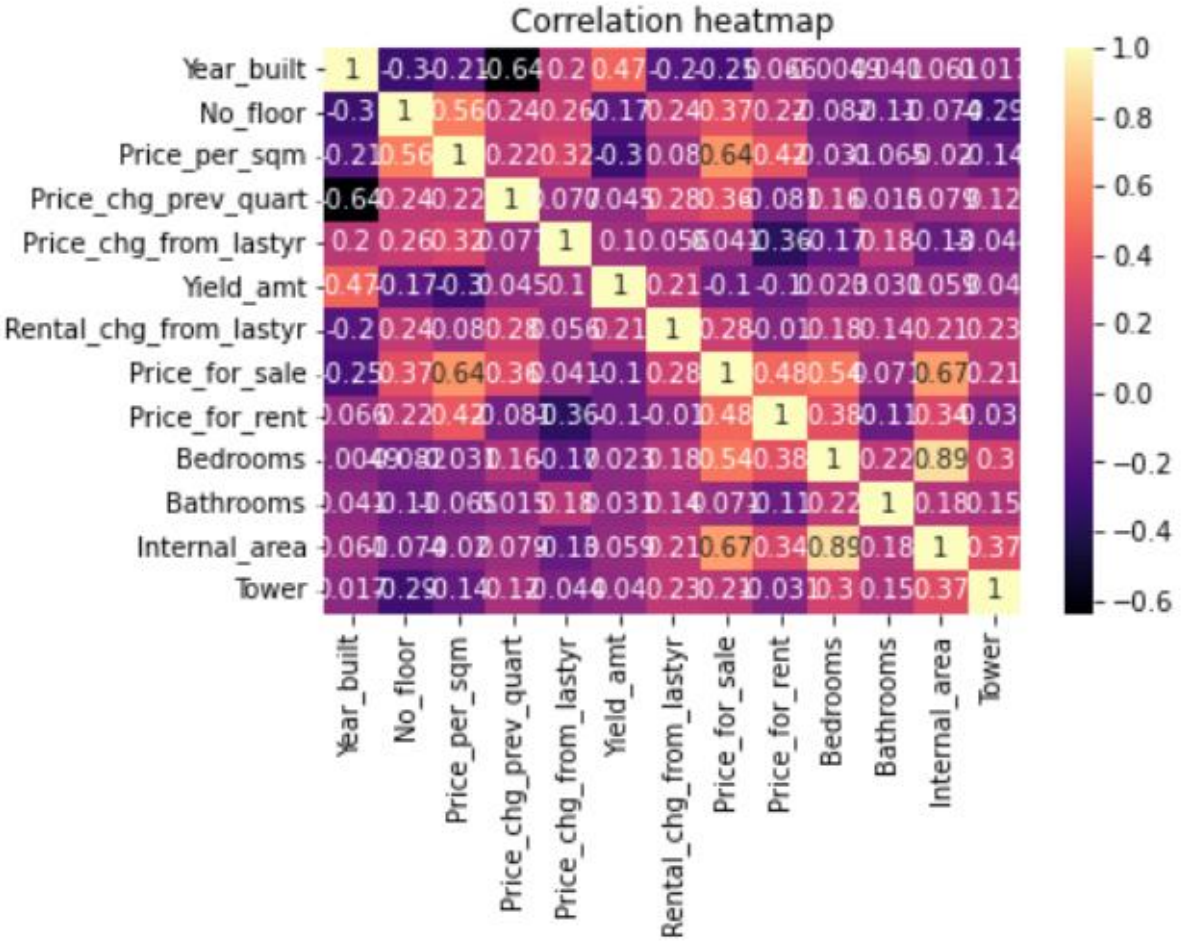
- 원시 데이터는 웹 크롤링 프로세스에서 일부 결측값이 있음을 보여줍니다
- **'Title'** 변수가 없습니다 -> **레코드를 제거합니다**
- **‘Price_per_sqm’** 변수가 없습니다 -> **레코드를 제거합니다**
- 결측값을 제거한 후 데이터 = **990 -> 713**

Title	Year_built	No_floor	Nearest_bts	Dist_bts	Nearest_mrt	Dist_mrt	Price_per_sqm	Price_chg_prev_quart	Price_chg_from_lastyr	Yield_amt	Rental_chg_from_lastyr	Price_for_sale	Price_for_rent	Bedrooms	Bathrooms	Internal_area	Tower
The Lofts Ekkamai	2016	28					192794	0	-0.23	4.43	1.86	12300000	35000	2	1	65	1
The Privacy Rama 9	2019	30					111577	0	-1.98	4.38	-2.4	2771000	0	1	1	27	1
RHYTHM Ekkamai	2018	32					217476	0	-0.09	4.51	-1.68	8500000	30000	1	1	38	1
Ceil by Sansiri	2013	17					132983	1.4	8.71	3.69	-17.04	6299999	0	1	1	47	3
M Jatujak	2018	34					147375	0	6.03	4.52	0.91	4590000	0	1	1	35	2
Juldis River Mansion	1996	16					49136	0	0	8.3	0	3100000	0	1	1	37	1
Sukhumvit House	1986	12					102404	0	7.06	4.23	2.27	9900000	0	2	2	109	1
Baan Chan	1988	8					88194	0	-10.96	4.2	-4.04	12700000	45000	4	2	160	3
Jewelry Trade Center	1996	56					77113	0	0.01	5.2	23.7	7490000	25000	2	1	91	1

Correlation Analysis

- Price_per_sqm and Price_for_sale (0.64)
- Price_per_sqm and No_floor (0.56)
- Price_per_sqm and Price_for_rent (0.42)
- Price_per_sqm and Price_chg_from_lastyr (0.32)
- Price_per_sqm and Price_chg_prev_quart (0.22)

- **‘Price_for_sale’, ‘No_floor’, ‘Price_for_rent’, ‘Price_chg_from_lastyr’, ‘Price_chg_prev_quart’**는 특성 중요도입니다



Data Modeling

- Linear Regression (선형 회귀)
- Regression Tree (회귀나무)
- Random Forest Regression (랜덤 포레스트)
- Gradient Boosting Regression (그래디언트 부스팅)

Data Modeling

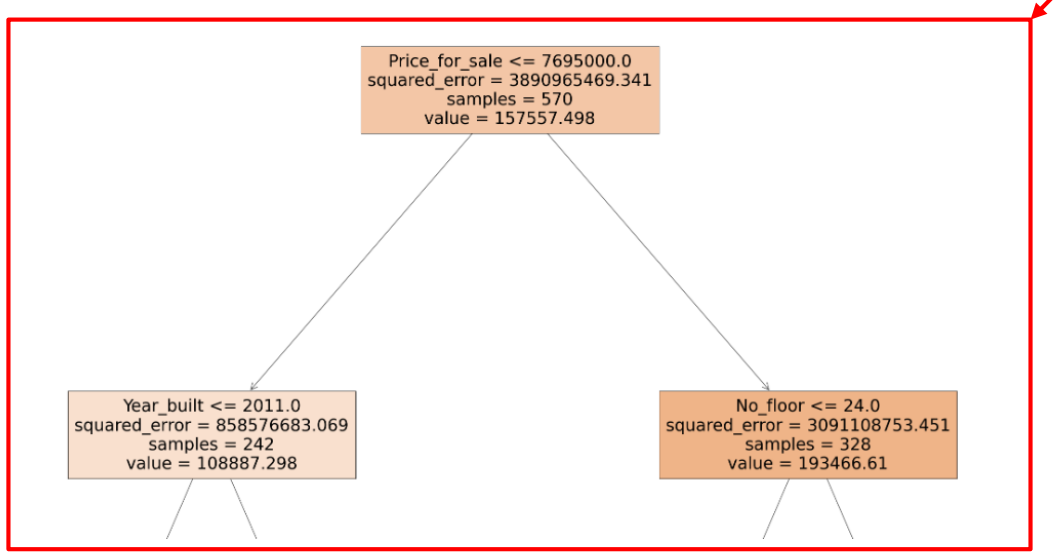
- **Linear Regression (선형 회귀)**
- Training : Testing = 80 : 20

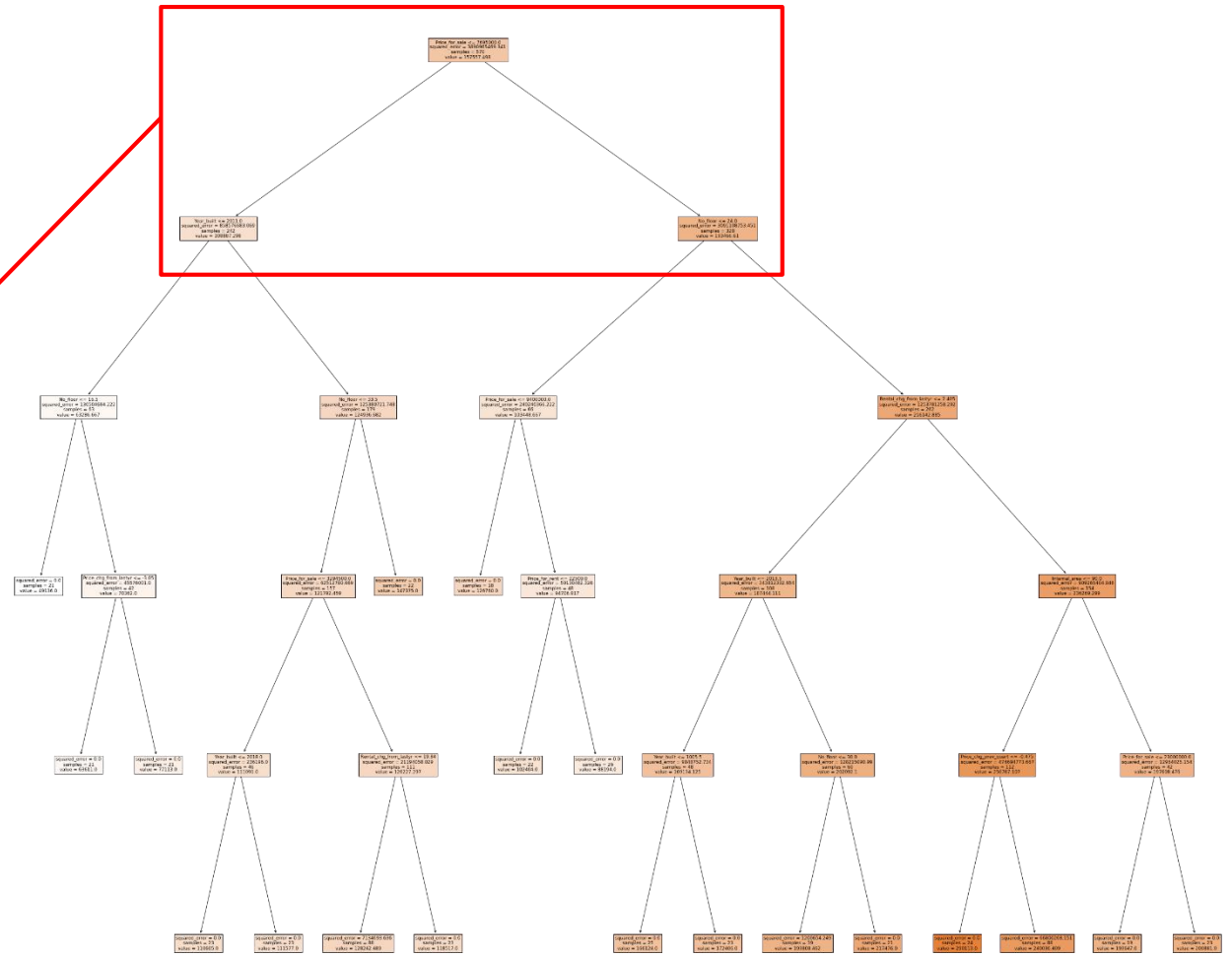
LR equation = (1.20140413e+01)**X1** + (-1.53316062e+02)**X2** + (-1.76258866e+03)**X3**
+ (2.18680357e+03)**X4** + (-9.81536639e+03)**X5** + (1.30766027e+02)**X6**
+ (1.09396752e-02)**X7** + (6.07946188e-01)**X8** + (2.05501104e+04)**X9**
+ (-7.72188246e+02)**X10** + (-1.67746384e+03)**X11** + (-3.76242248e+03)**X12**
+ 139360.18165476

Train Score: 0.908
Test Score: 0.891

Data Modeling

- **Regression Tree (회귀나무)**
- Training : Testing = 80 : 20
- random_state = 42, max_depth=5





Data Modeling

- **Regression Tree**

```
|--- Price_for_sale <= 7695000.00
|   |--- Year_built <= 2011.00
|   |   |--- No_floor <= 16.50
|   |   |   |--- value: [49136.00]
|   |   |--- No_floor >  16.50
|   |   |   |--- Price_chg_from_lastyr <= -3.05
|   |   |   |   |--- value: [63611.00]
|   |   |   |--- Price_chg_from_lastyr >  -3.05
|   |   |   |   |--- value: [77113.00]
|   |--- Year_built >  2011.00
|   |   |--- No_floor <= 33.50
|   |   |   |--- Price_for_sale <= 3294500.00
|   |   |   |   |--- Year_built <= 2018.00
|   |   |   |   |   |--- value: [110605.00]
|   |   |   |   |--- Year_built >  2018.00
|   |   |   |   |   |--- value: [111577.00]
|   |   |   |--- Price_for_sale >  3294500.00
|   |   |   |   |--- Rental_chg_from_lastyr <= 19.84
|   |   |   |   |   |--- value: [128242.49]
|   |   |   |   |--- Rental_chg_from_lastyr >  19.84
|   |   |   |   |   |--- value: [118517.00]
|   |   |--- No_floor >  33.50
|   |   |   |--- value: [147375.00]
```

```
|--- Price_for_sale >  7695000.00
|   |--- No_floor <= 24.00
|   |   |--- Price_for_sale <= 9400000.00
|   |   |   |--- value: [126760.00]
|   |   |--- Price_for_sale >  9400000.00
|   |   |   |--- Price_for_rent <= 22500.00
|   |   |   |   |--- value: [102404.00]
|   |   |   |--- Price_for_rent >  22500.00
|   |   |   |   |--- value: [88194.00]
|   |--- No_floor >  24.00
|   |   |--- Rental_chg_from_lastyr <= 2.41
|   |   |   |--- Year_built <= 2013.50
|   |   |   |   |--- Year_built <= 1005.50
|   |   |   |   |   |--- value: [166124.00]
|   |   |   |   |--- Year_built >  1005.50
|   |   |   |   |   |--- value: [172406.00]
|   |   |   |--- Year_built >  2013.50
|   |   |   |   |--- No_floor <= 30.00
|   |   |   |   |   |--- value: [193808.46]
|   |   |   |   |--- No_floor >  30.00
|   |   |   |   |   |--- value: [217476.00]
|   |   |--- Rental_chg_from_lastyr >  2.41
|   |   |   |--- Internal_area <= 90.00
|   |   |   |   |--- Price_chg_prev_quart <= -0.47
|   |   |   |   |   |--- value: [290113.00]
|   |   |   |   |--- Price_chg_prev_quart >  -0.47
|   |   |   |   |   |--- value: [240036.41]
|   |   |   |--- Internal_area >  90.00
|   |   |   |   |--- Price_for_sale <= 23000000.00
|   |   |   |   |   |--- value: [193647.00]
|   |   |   |   |--- Price_for_sale >  23000000.00
|   |   |   |   |   |--- value: [200881.00]
```

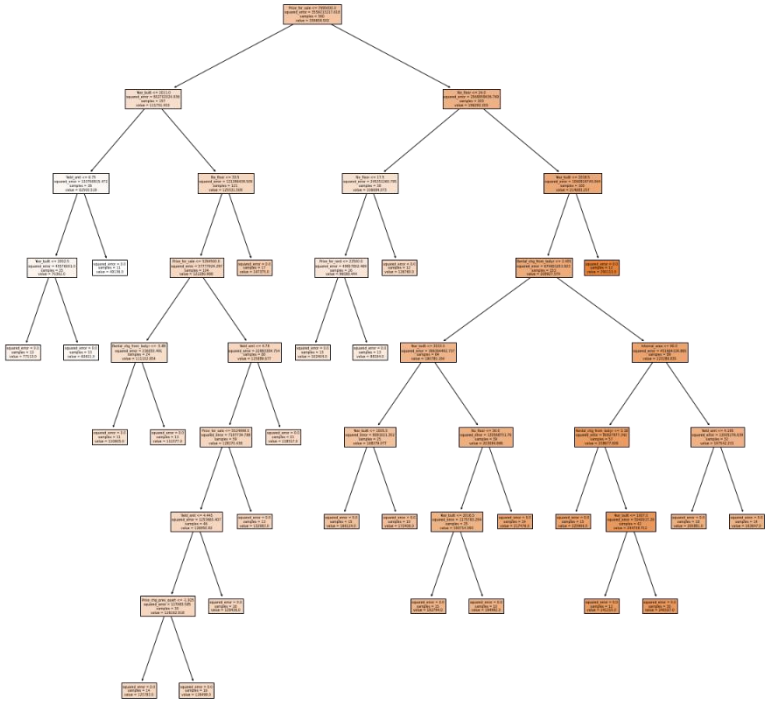
Train Score: 0.997

Test Score: 0.996

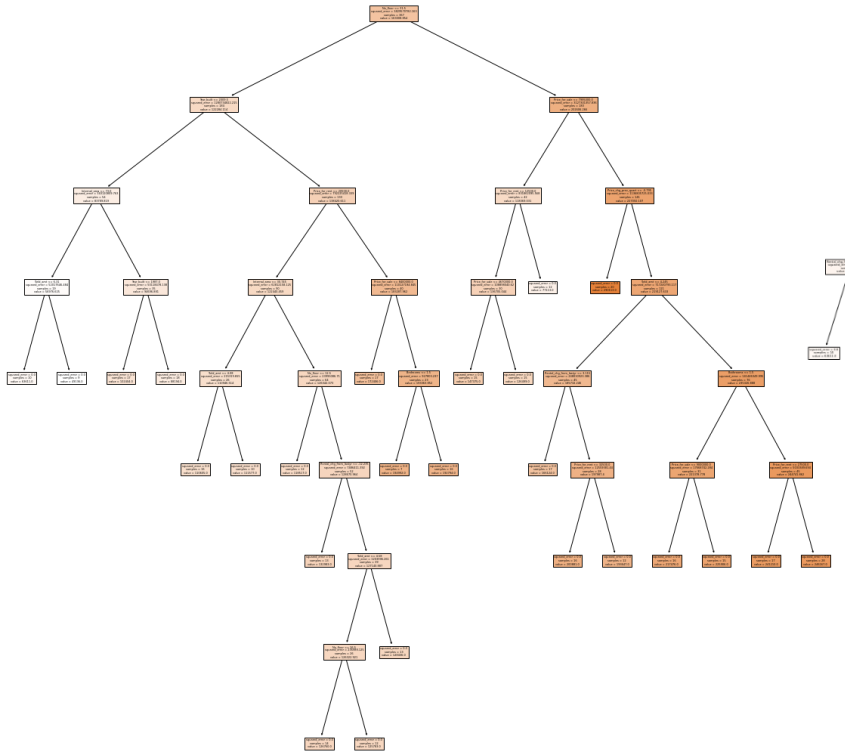
Data Modeling

- **Random Forest Regression (랜덤 포레스트)**
- Training : Testing = 80 : 20, n_estimators=100

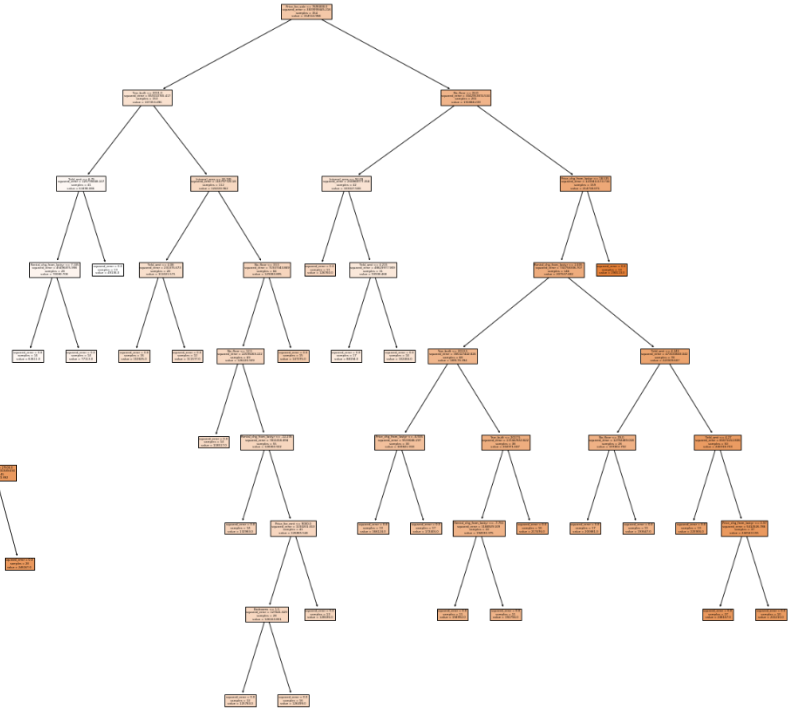
Train Score: 1.0
Test Score: 1.0



Tree1 (max_depth = 8)



Tree2 (max_depth = 8)



Tree3 (max_depth = 8)

Data Modeling

- **Gradient Boosting Regression (그래디언트 부스팅)**
- Training : Testing = 80 : 20
- n_estimators = 1000
- max_depth: 3
- min_samples_split: 5
- learning_rate: 0.01

Train Score: 0.9999984
Test Score: 0.9999982

Measurement

- RMSE (Root mean square error)

Model	RMSE 1 (Existing variable)	RMSE 2 (Existing + Adding variable)
Linear Regression	46933.428	19784.273
Decision Tree Regression	4974.157	3593.853
Random Forest Regression	0	0
Gradient Boosting	480.7	79.522

Conclusion

- 기존 변수에서 **‘No_floor’, ‘Price_chg_from_lastyr’, ‘Price_chg_prev_quart’**는 특성 중요도입니다
- 추가 변수에서 **'Price_for_sale', ‘Price_for_rent’**는 특성 중요도입니다
- **더 많은 특성**을 추가하면 **더 좋은 성능** 제공합니다
- **랜덤 포레스트 회귀 분석**이 최상의 예측 모델입니다