ESC 2019 Spring

2조 Final Project

김주은, 유아현, 이건, 이솔희, 이재현, 차상훈

CONTENTS

01

Review

- Deriv1 : 구별 평당 평균 가격
- Deriv2 : 지하철역까지의 거리
- Deriv3 : 환승역까지의 거리
- Deriv4 : 골드라인까지의 거리
- Deriv5 : 쿼드러플 역세권

```
deriv1          : num  51684244 51684244 51684244 51684244 51684244 ...
deriv2          : num  0.435 0.435 0.313 0.605 0.597 ...
deriv3          : num  0.739 0.739 1.09 0.894 0.793 ...
deriv4          : num  1.32 1.32 1.55 1.54 1.31 ...
deriv5          : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
deriv6          : num  1.01 1.16 1.11 1.13 1.04 ...
deriv7          : num  0.504 0.504 0.727 0.569 0.515 ...
deriv8          : num  2.74 2.74 2.12 3.07 2.83 ...
```

- Deriv6 : 각 고등학교별 서울대 진학률
- Deriv7 : 자율형 사립고등학교까지의 거리
- Deriv8 : 한강까지의 거리

외부데이터를 사용하여 만든 파생변수

- Deriv6 : 각 고등학교별 서울대 진학률

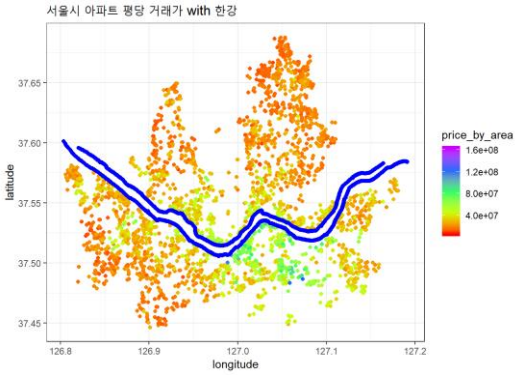
고등학교의 경우 학교의 거리보다는 학교의 수준이 더 중요하다는 점에 착안하여 특목고, 특성화고 제외 각 고교별 서울대 진학률을 파생변수로 생성하였다.

출처 : 학교알리미

- Deriv8 : 한강 프리미엄

한강과 인접한 곳은 평당 거래가격이 상대적으로 높다는 점에 착안하여 한강과의 거리를 파생변수로 만들었다.

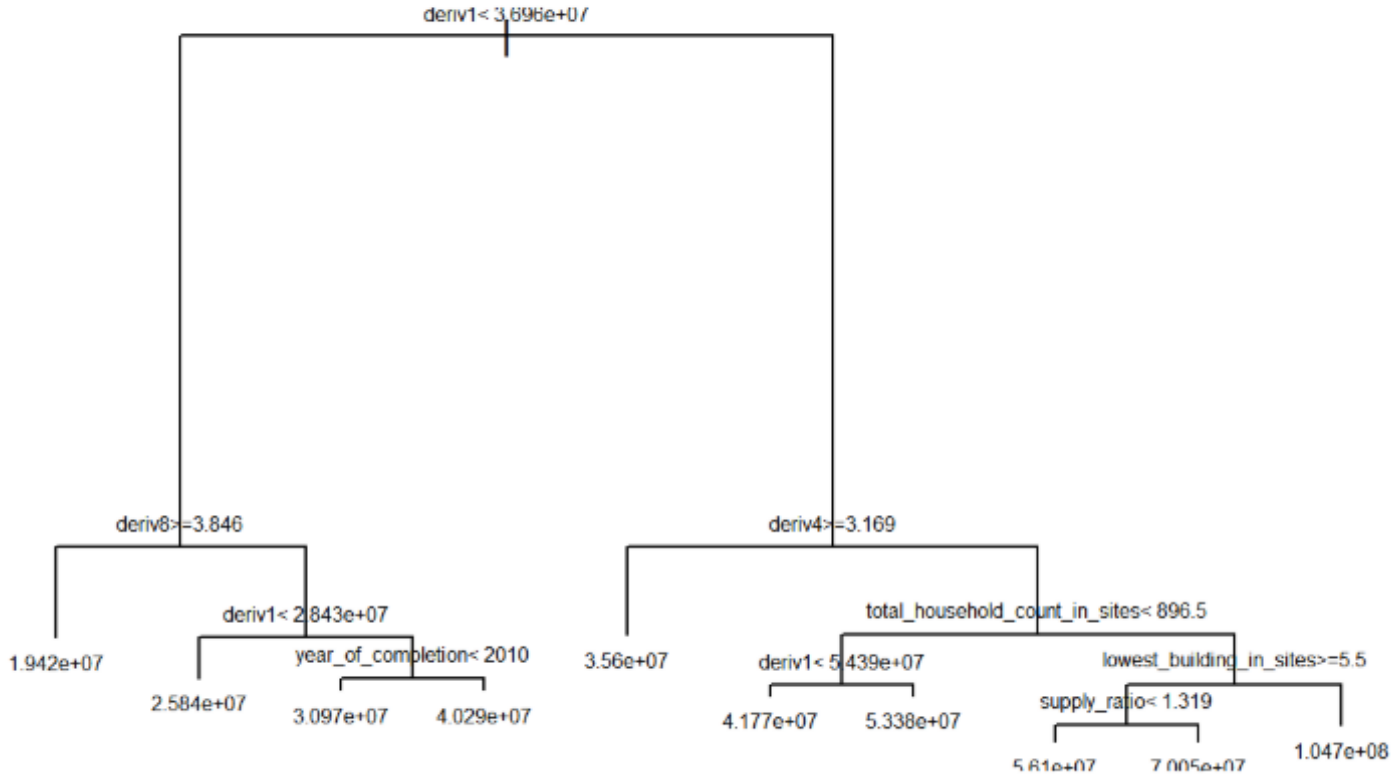
출처 : 이재현



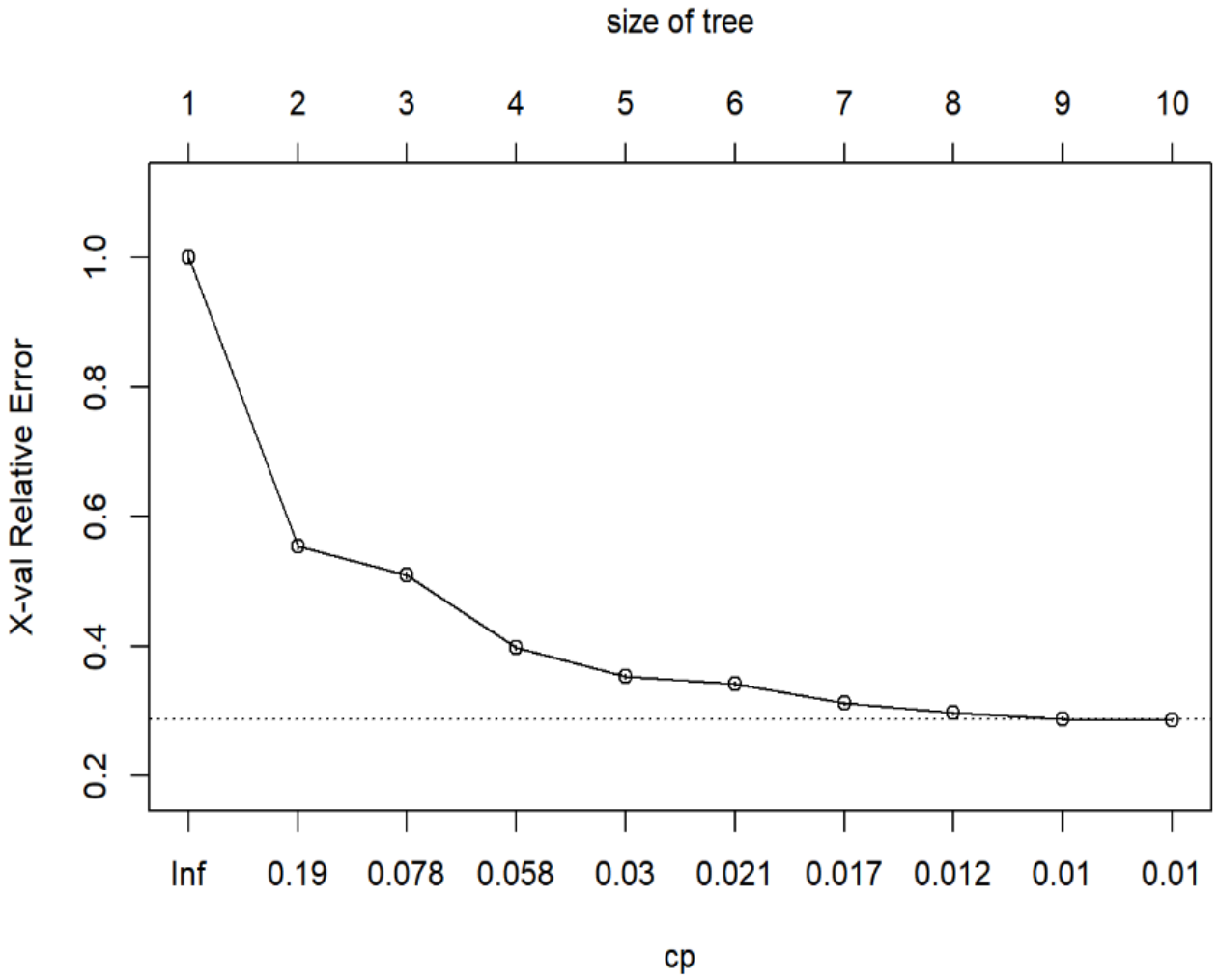
02

Decision Tree

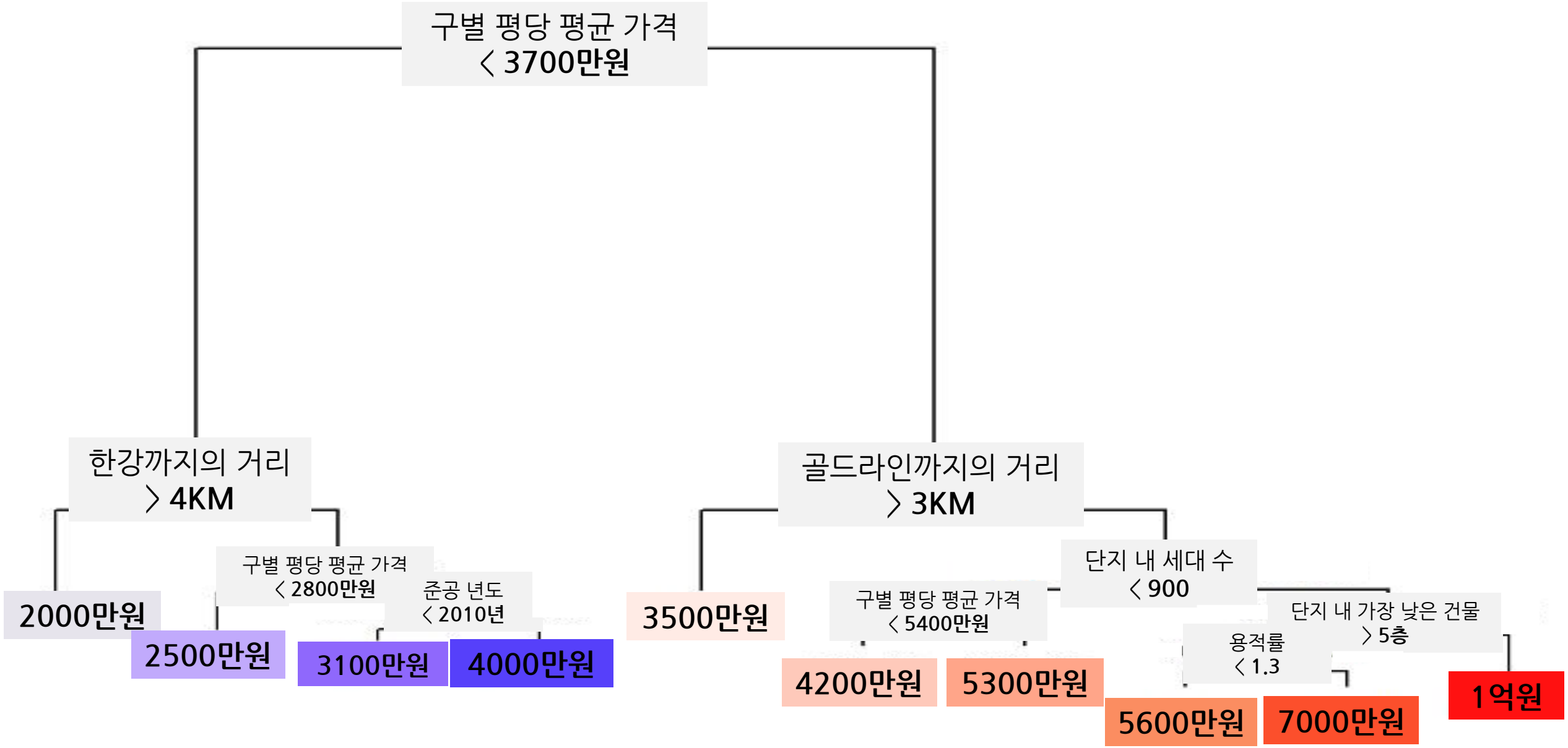
- **rpart 패키지**를 통해 decision tree 모델링
- 총 **10개의 설명변수**가 사용되었다.



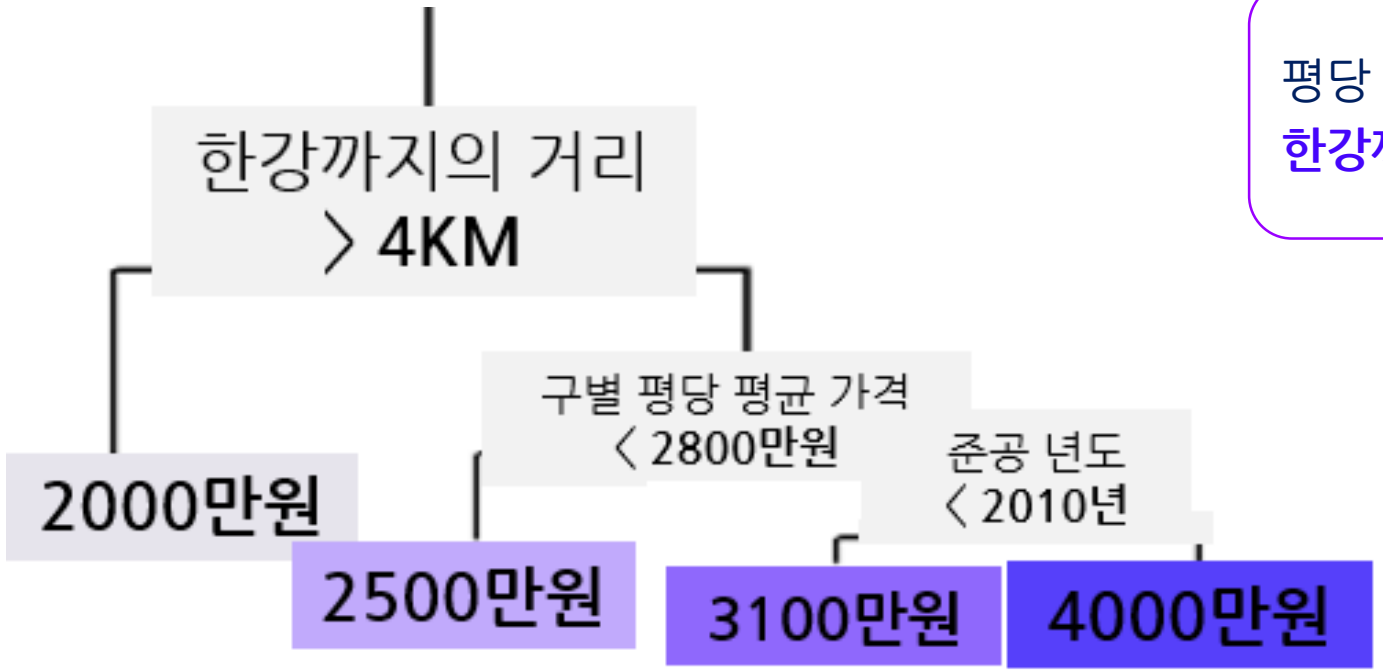

```
## Variables actually used in tree construction:
## [1] deriv1                          deriv4
## [3] deriv8                          lowest_building_in_sites
## [5] supply_ratio                    total_household_count_in_sites
## [7] year_of_completion
```



- Cp를 기준으로 가지치기를 진행한 결과 기존 모델과 똑같이 10개의 설명변수를 사용하는 것이 에러가 가장 작았다.
- 분류를 위한 데이터가 아니기 때문에 가지치기를 수행해도 별 효과가 없는 것으로 판단된다.
- 사용된 설명변수들 역시 동일하였기 때문에 모델 역시 동일하다.

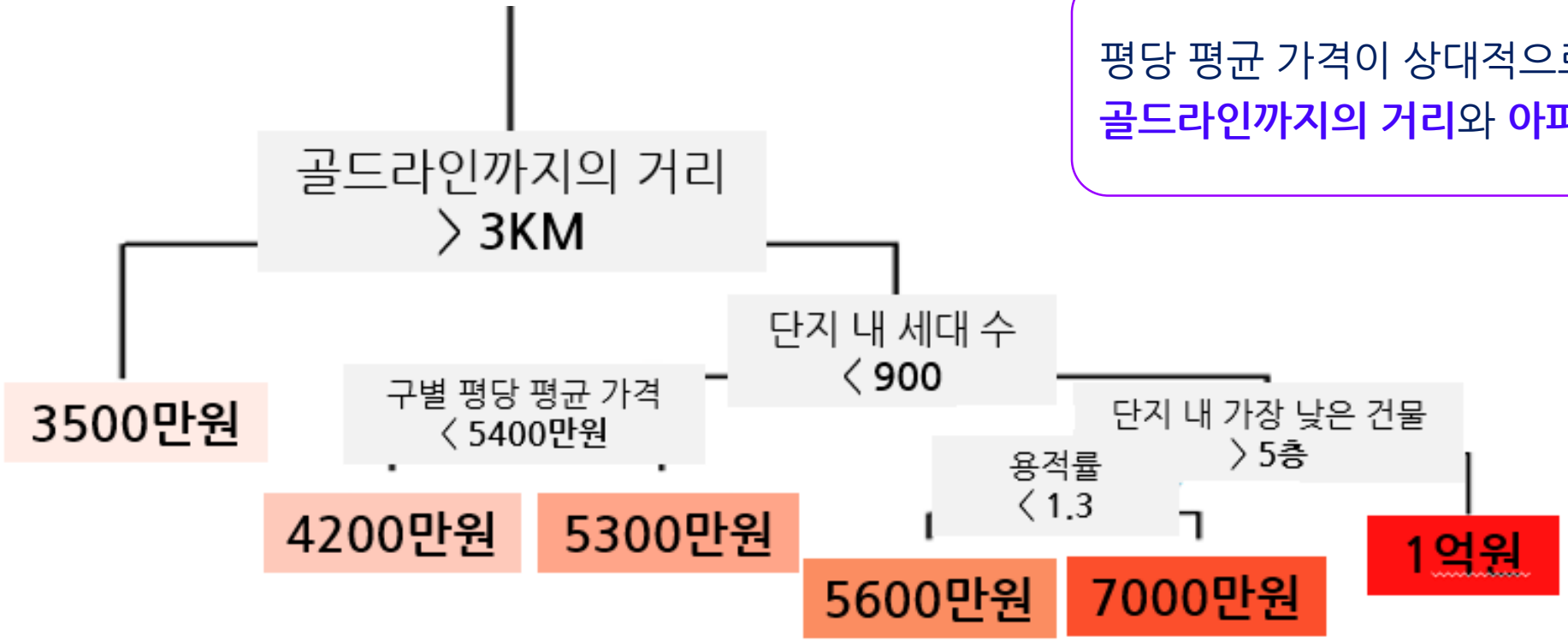

구별 평당 평균 가격
< 3700만원
한강까지의 거리
> 4KM
2000만원
구별 평당 평균 가격
< 2800만원
2500만원
준공 년도
< 2010년
3100만원
4000만원
골드라인까지의 거리
> 3KM
3500만원
구별 평당 평균 가격
< 5400만원
4200만원
5300만원
단지 내 세대 수
< 900
용적률
< 1.3
5600만원
단지 내 가장 낮은 건물
> 5층
7000만원
1억원

최근 1개월 간 평당 평균 가격 3700만원 미만인 구

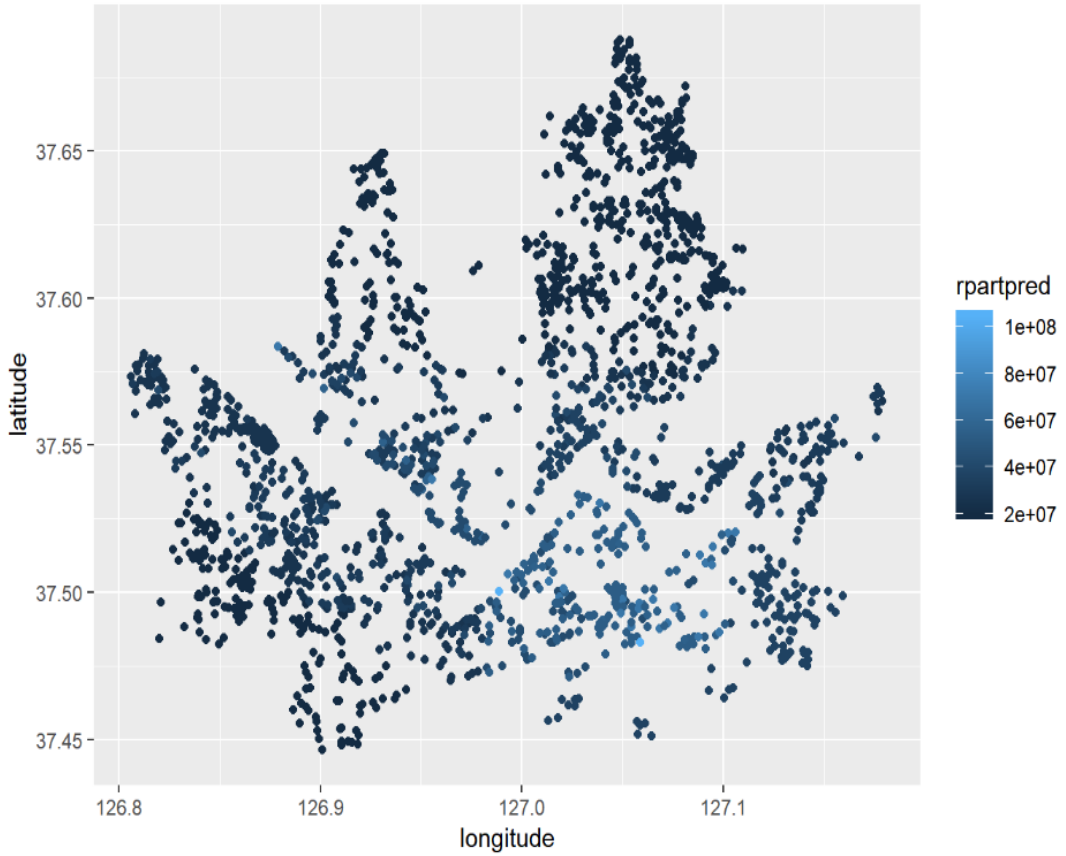


평당 평균 가격이 상대적으로 낮은 지역은
한강까지의 거리와 **준공 년도**가 중요

최근 1개월 간 평당 평균 가격 3700만원 이상인 구



평당 평균 가격이 상대적으로 높은 지역은
골드라인까지의 거리와 **아파트의 외관**이 중요



- 가지치기가 완료된 최종 모델로 예측을 수행한 결과 전체적으로 색깔의 차이가 크지 않다.
- Decision tree는 분류에 특화되어있는 만큼 예측에는 큰 효과가 없었다.
- 그러나 Decision tree를 통해 각 변수의 중요도를 파악할 수 있었다.

03

Regression

Before

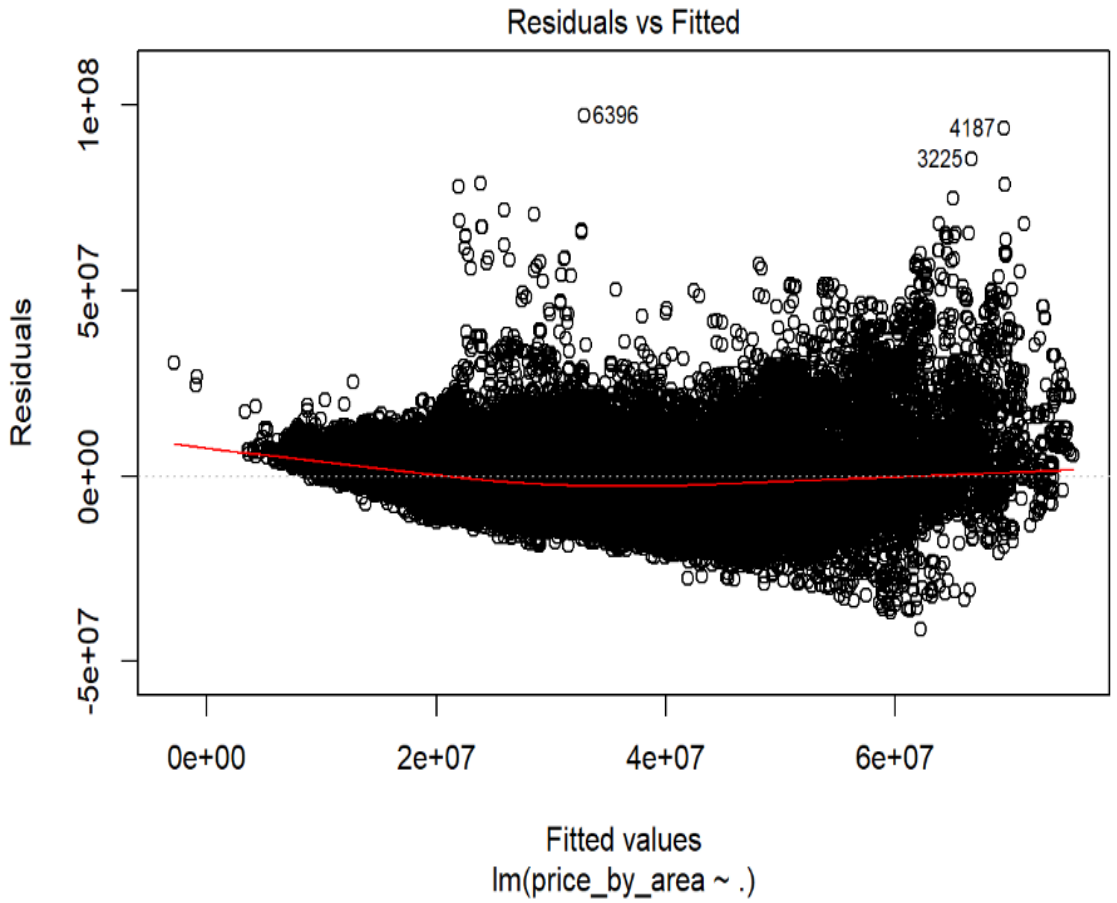
변수에 대한 전처리 과정 없이 모델에 적합한 결과, 대다수의 변수가 유의미하게 도출

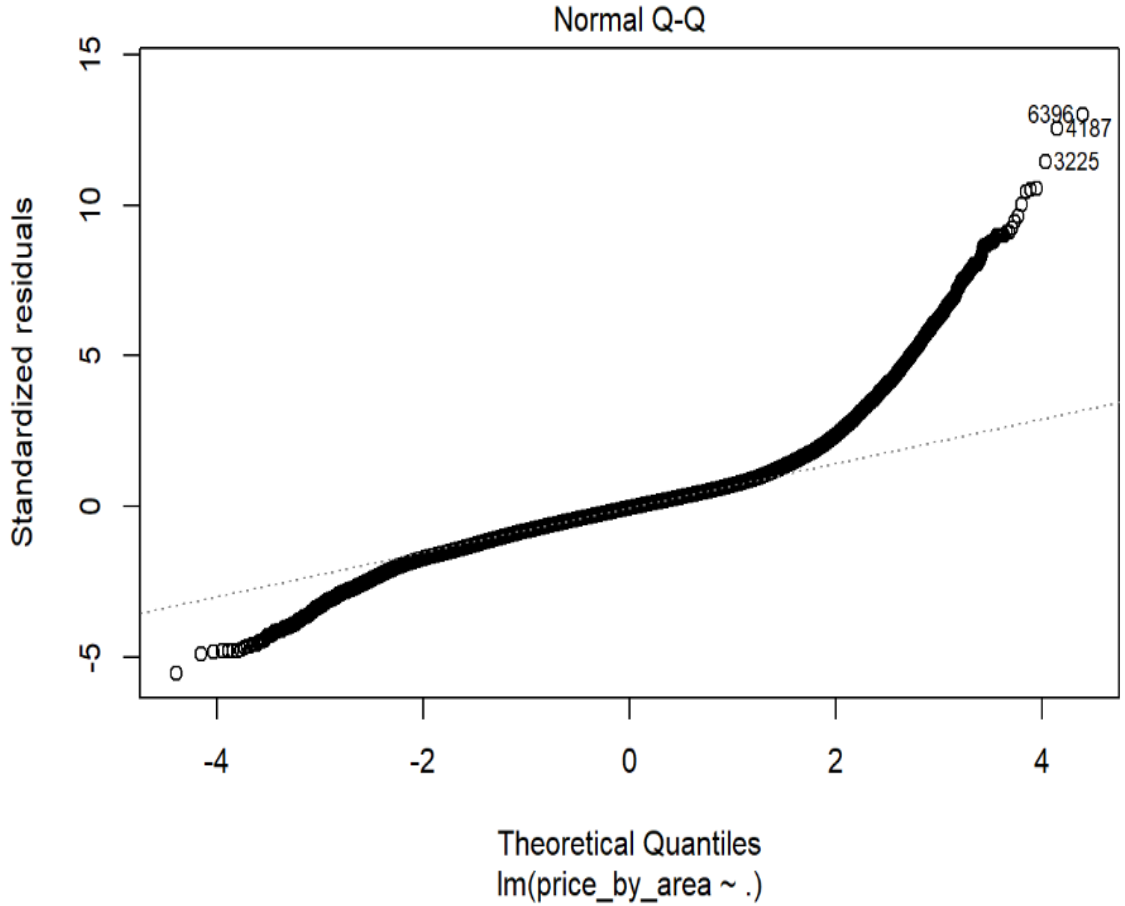
```
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -9.283e+07  8.709e+06 -10.660  < 2e-16 ***
## year_of_completion             5.402e+04  4.480e+03  12.057  < 2e-16 ***
## parking_ratio                  2.312e+06  7.069e+04  32.702  < 2e-16 ***
## total_household_count_in_sites 2.135e+03  2.458e+01  86.861  < 2e-16 ***
## type_ratio                    -3.588e+06  1.193e+05 -30.079  < 2e-16 ***
## tallest_building_in_sites      2.491e+04  5.936e+03   4.196 2.72e-05 ***
## lowest_building_in_sites      -2.843e+04  5.544e+03  -5.128 2.94e-07 ***
## front_door_structure_stairway  1.169e+06  2.106e+05   5.553 2.81e-08 ***
## front_door_structure_corridor -1.186e+06  2.178e+05  -5.447 5.14e-08 ***
## heat_type_district             8.501e+05  1.705e+05   4.986 6.17e-07 ***
## heat_type_individual          -1.593e+05  9.184e+04  -1.734   0.0829 .
## heat_fuel_cogeneration         1.327e+06  1.595e+05   8.315  < 2e-16 ***
## supply_ratio                   5.960e+05  4.289e+05   1.390   0.1646
## exclusive_use_area            -1.372e+05  2.066e+03 -66.406  < 2e-16 ***
## room_count                     9.006e+05  6.780e+04  13.283  < 2e-16 ***
## bathroom_count                 3.361e+04  9.271e+04   0.363   0.7169
## floor                          5.810e+04  4.447e+03  13.065  < 2e-16 ***
## deriv1                         8.358e-01  2.709e-03 308.490  < 2e-16 ***
## deriv2                        -2.564e+06  7.905e+04 -32.440  < 2e-16 ***
## deriv3                        -6.453e+05  3.568e+04 -18.084  < 2e-16 ***
## deriv4                        -7.013e+05  2.330e+04 -30.100  < 2e-16 ***
## deriv5                        -3.349e+06  1.423e+05 -23.537  < 2e-16 ***
## deriv6                        -2.564e+04  4.836e+04  -0.530   0.5960
## deriv7                        -8.983e+05  4.504e+04 -19.945  < 2e-16 ***
## deriv8                        -1.159e+06  2.930e+04 -39.569  < 2e-16 ***
```

- 24개의 설명변수를 모두 사용하여 **단순선형 회귀분석**을 시행한 결과 **20개의 설명변수**가 유의함.
- ADJ R^2도 **0.75**으로 높은 수준이다.

Before

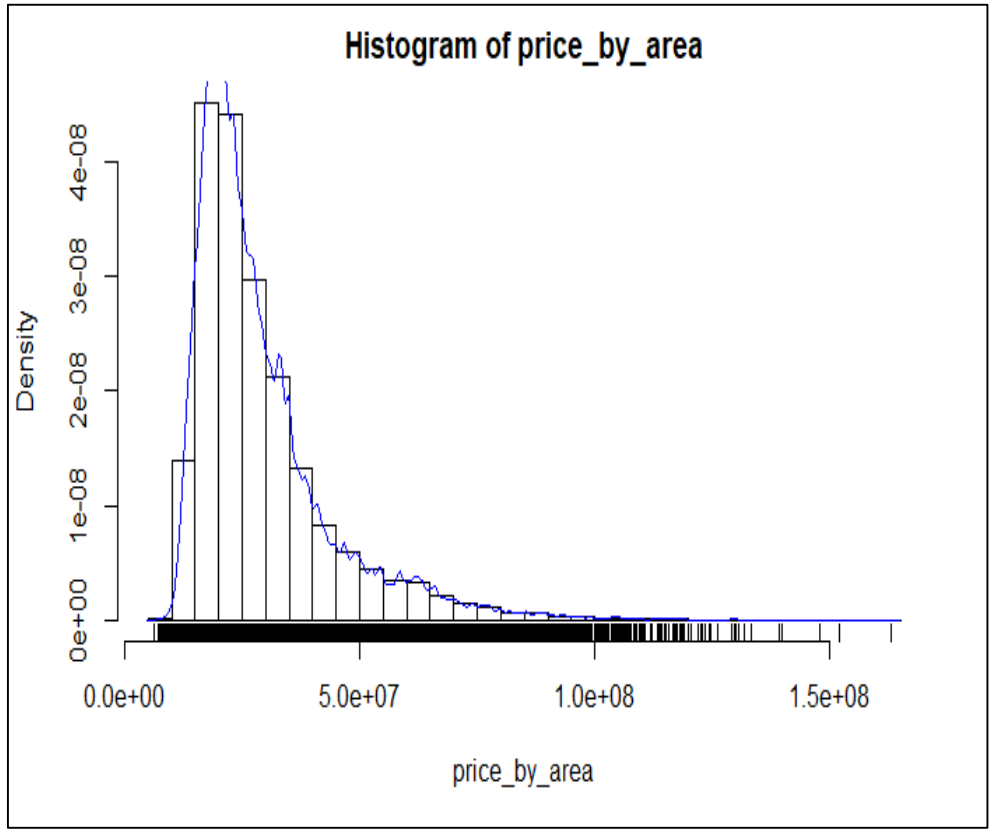
Note : 1) 그러나 **잔차의 분산이 일정하지 않고**, 2) **정규성 가정에 크게 위반**됨을 확인함



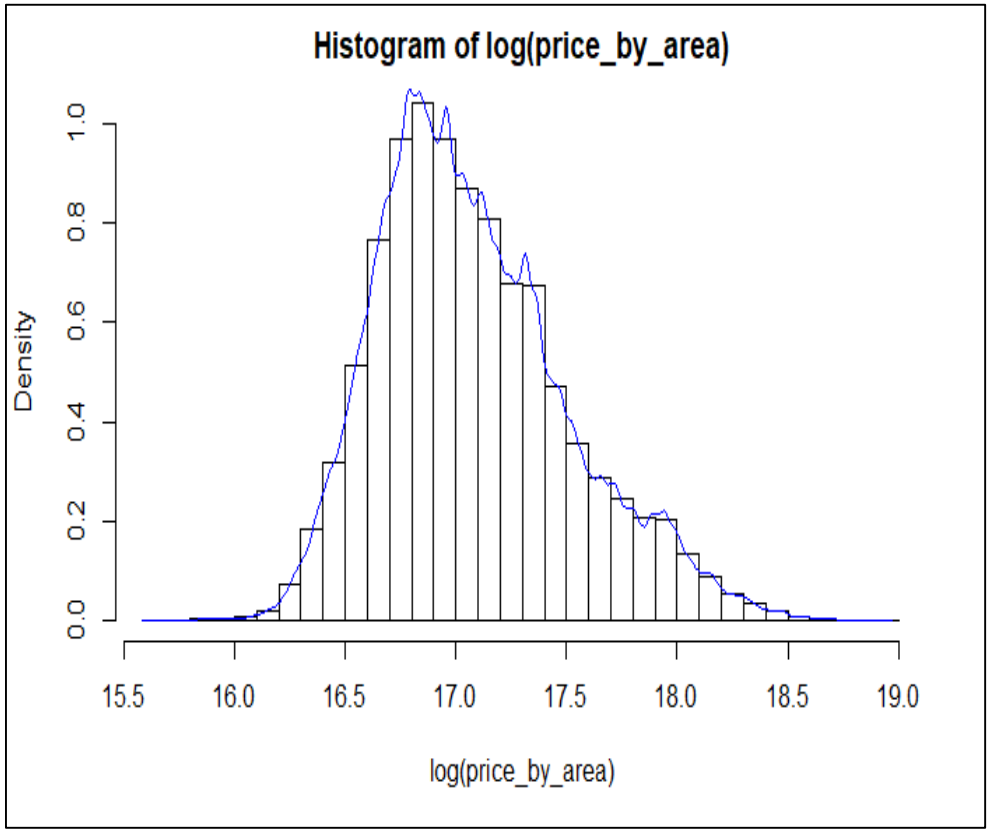


Solution

Skewed된 변수 8개에 대해 **log변환** 실시



Log



After

변수에 대한 **전처리 과정 이후** , **모든 변수**가 유의미하게 도출

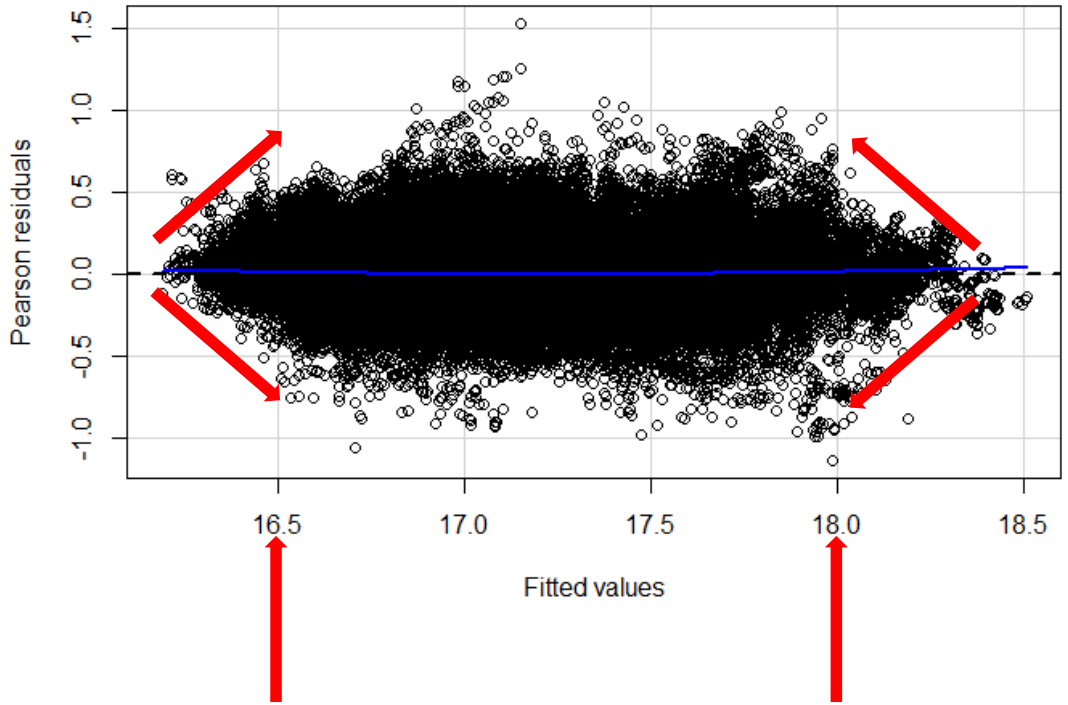
```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      3.781e+00  2.380e-01  15.885  < 2e-16 ***
year_of_completion               6.080e-03  1.225e-04  49.618  < 2e-16 ***
type_ratio                      -4.627e-02  3.316e-03 -13.954  < 2e-16 ***
tallest_building_in_sites       -3.014e-03  1.580e-04 -19.077  < 2e-16 ***
lowest_building_in_sites        -1.497e-03  1.514e-04  -9.892  < 2e-16 ***
front_door_structuremixed        7.108e-02  5.998e-03  11.851  < 2e-16 ***
front_door_structurestairway     7.306e-02  2.479e-03  29.468  < 2e-16 ***
heat_typedistrict                3.086e-02  4.693e-03   6.574 4.91e-11 ***
heat_typeindividual             -1.362e-02  2.545e-03  -5.350 8.80e-08 ***
heat_fuelgas                    -3.680e-02  4.387e-03  -8.389  < 2e-16 ***
supply_ratio                     3.875e-01  1.093e-02  35.461  < 2e-16 ***
room_count                      -7.343e-02  1.373e-03 -53.473  < 2e-16 ***
bathroom_count                  -5.769e-02  2.422e-03 -23.822  < 2e-16 ***
deriv1                           2.200e-08  7.803e-11 281.987  < 2e-16 ***
deriv3                          -3.135e-02  9.453e-04 -33.161  < 2e-16 ***
deriv5                          -6.876e-02  3.881e-03 -17.716  < 2e-16 ***
log(parking_ratio)               4.709e-02  2.059e-03  22.871  < 2e-16 ***
log(floor)                       1.794e-02  8.950e-04  20.046  < 2e-16 ***
log(total_household_count_in_sites)  9.544e-02  8.929e-04 106.888  < 2e-16 ***
log(deriv2)                     -5.984e-02  1.279e-03 -46.771  < 2e-16 ***
log(1 + deriv4)                 -2.451e-02  1.764e-03 -13.894  < 2e-16 ***
log(1 + deriv7)                 -4.907e-02  2.539e-03 -19.323  < 2e-16 ***
log(1 + deriv8)                 -1.660e-01  1.944e-03 -85.380  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2055 on 89977 degrees of freedom
Multiple R-squared:  0.7726,    Adjusted R-squared:  0.7725
F-statistic: 1.389e+04 on 22 and 89977 DF,  p-value: < 2.2e-16
```

- 22개의 설명변수를 모두 사용하여 **단순선형 회귀분석**을 시행한 결과 22개의 **모든 설명변수가 유의**함.

- ADJ R^2 도 **0.7725**으로 전처리 전에 비해 소폭 상승함

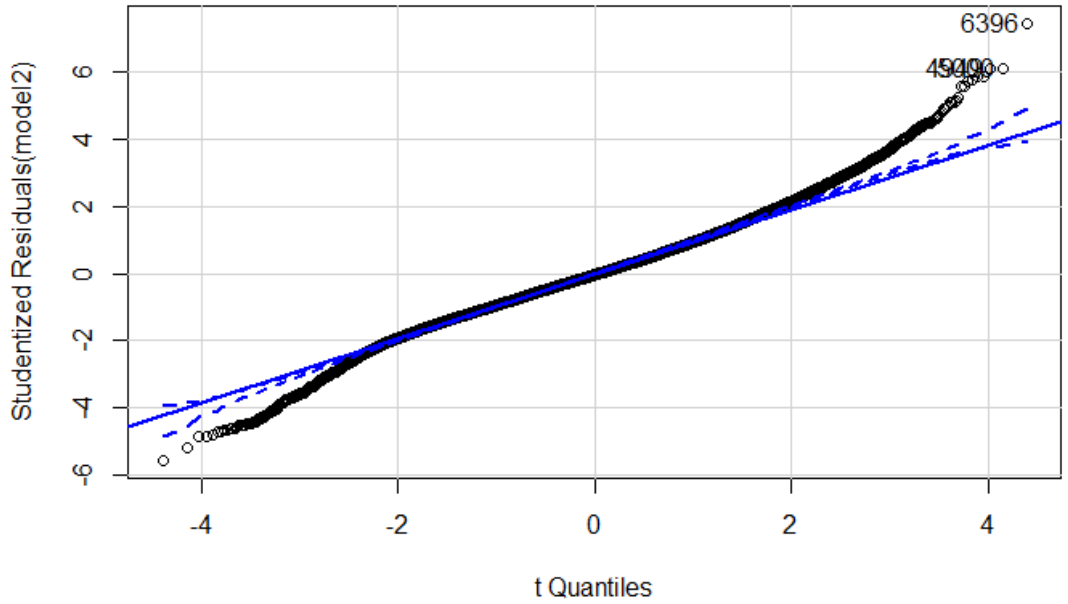
After

등분산 가정과 정규성 가정에 만족여부가 **개선**됨을 확인



평당 약 1500만원

평당 약 6600만원



그러나

평당 1500만원 미달 및 6600만원을 초과하는 구간은 등분산 가정이 잘 들어맞지 않음

→ 해당 모델은 **전 구간에 대한 일반화는 되지 못함**

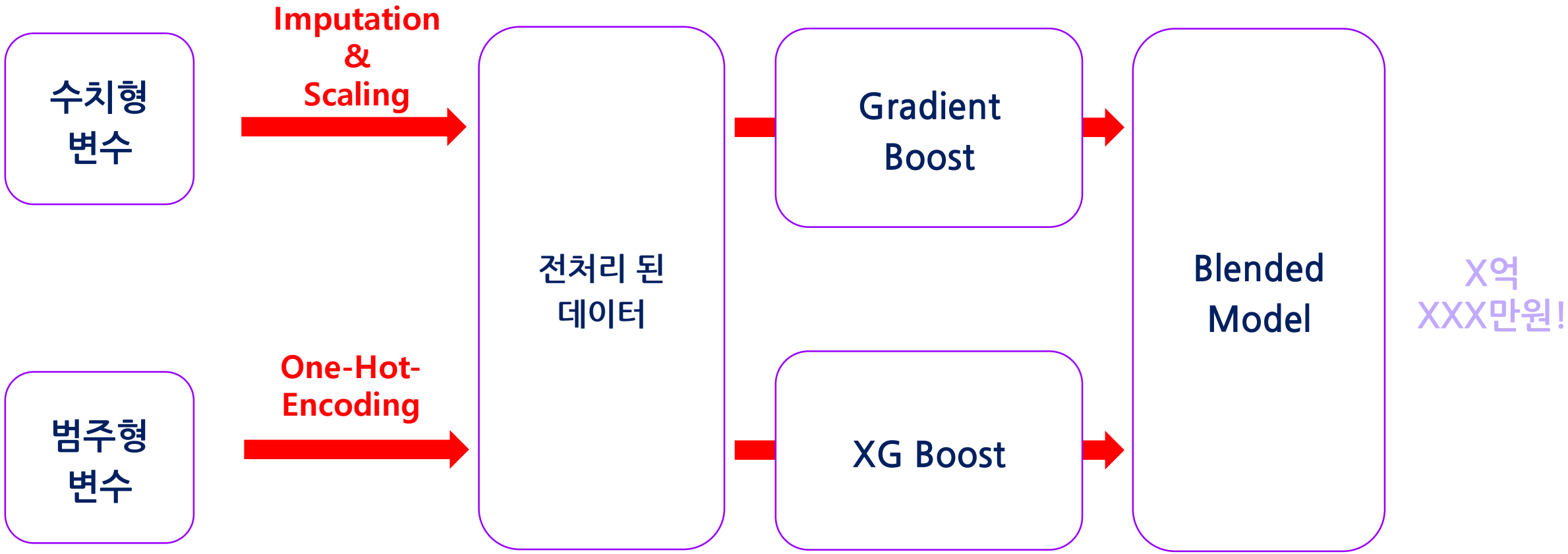
모델 해석

등분산 가정과 정규성 가정에 만족여부가 개선됨을 확인

변수명	t value	변수명	t value
year_of_completion	49.61	bathroom_count	-23.82
type_ratio	-13.95	deriv1	281.98
tallest_building_in_sites	-19.07	deriv3	-33.16
lowest_building_in_sites	-9.89	deriv5	-17.71
front_door_structuremixed	11.85	log(parking_ratio)	22.87
front_door_ structurestairway	29.46	log(floor)	20.04
heat_typedistrict	6.57	log(total_household _count_in_sites)	106.88
heat_typeindividual	-5.35	log(deriv2)	-46.77
heat_fuelgas	-8.38	log(1 + deriv4)	-13.89
supply_ratio	35.46	log(1 + deriv7)	-19.32
room_count	-53.47	log(1 + deriv8)	-85.38

04
Prediction Models

머신러닝 시스템 개요


수치형
변수
Imputation
&
Scaling
범주형
변수
One-Hot-
Encoding
전처리 된
데이터
Gradient
Boost
XG Boost
Blended
Model
X억
XXX만원!

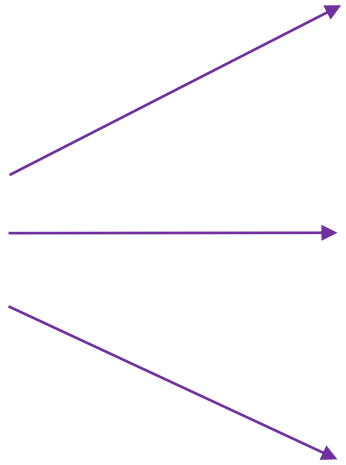
주요 설정 하이퍼 파라미터 – GBRegressor

HyperParameter	설정 값	설명	구분	용도
max_depth	15	트리 깊이	Base function	Feature Space 분할!
max_feature	11	선택 피쳐 최대 수		
min_sample_leaf	60	노드에 할당된 샘플 숫자		
min_sample_split	200	분할 시 필요한 샘플 수		
subsample	0.9	트레이닝 셋 사용율	Boosting 설정	모델 분산 줄이기!
learning_rate	0.01	학습율	최적화	최적해 찾기!
n_estimators	15000	부스팅 횟수		

하이퍼 파라미터 튜닝 과정 – GBRegressor

하이퍼 파라미터 튜닝 과정

1. Base function 튜닝
2. 부스팅 설정 튜닝
3. 최적화 방법 튜닝



max_depth
min_sample_leaf
max_feature
min_sample_split

Subsample

learning_rate
n_estimators

[튜닝 방법]

Cross Validation

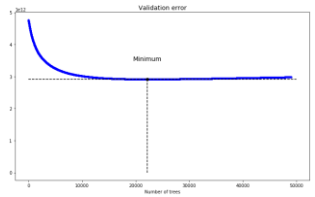
staged_predict
early_stopping

하이퍼 파라미터 튜닝 과정 – GBRegressor

Cross Validation
GridSearchCV를 통해 최적의 하이퍼파라미터 선택

하이퍼파라미터
max_depth
min_sample_leaf
max_feature
min_sample_split
subsample
learning_rate

Staged_predict() : early Stopping
Validation error를 통해 최적의 하이퍼파라미터 선택

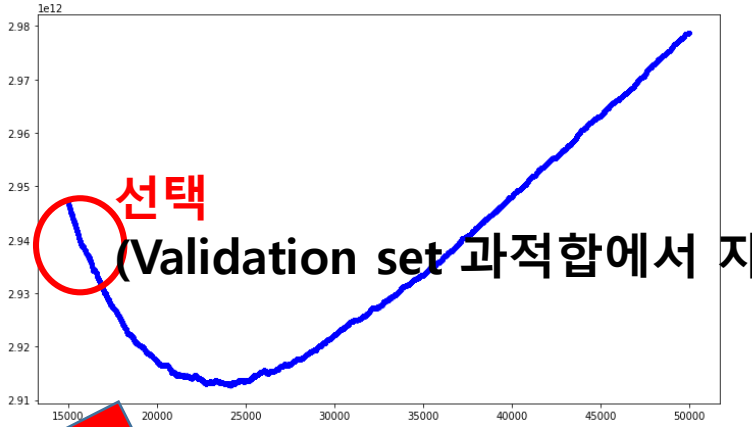


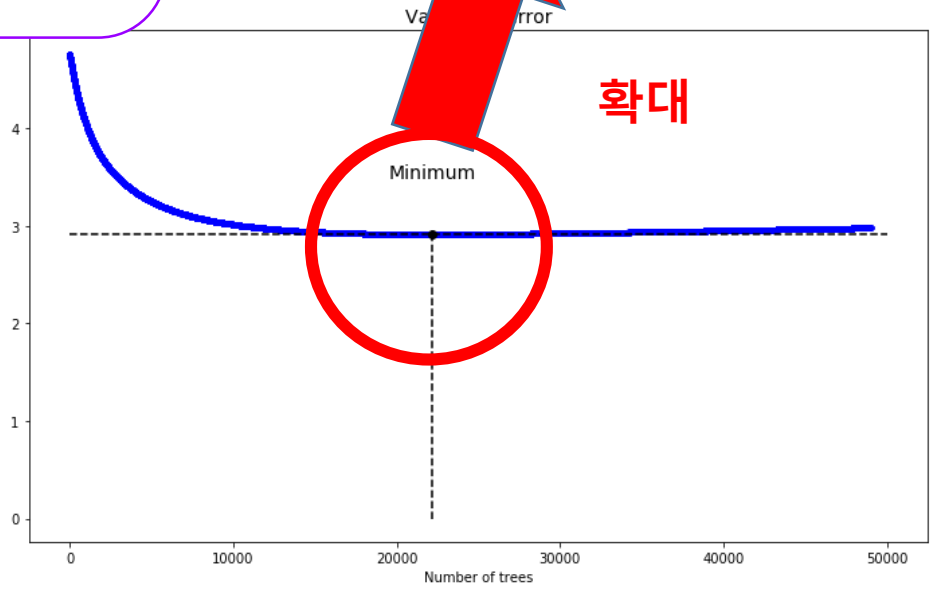
하이퍼파라미터
n_estimators

staged_predict()
early_stopping

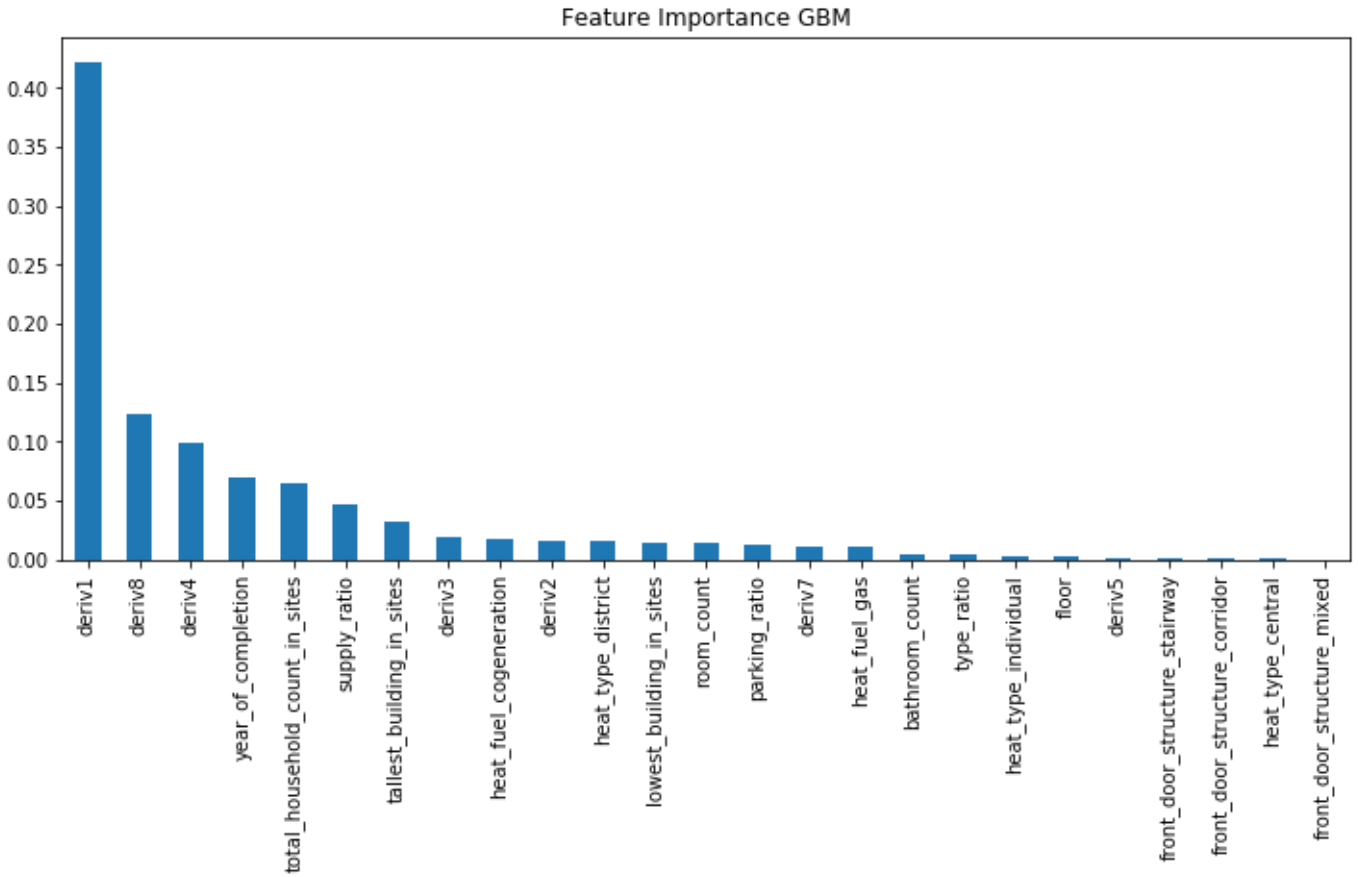
Validation error 경향의 관찰을 통해

학습률, 부스팅 횟수 선택





Importance– GBRegressor



[변수의 중요도]

- 무슨 구인가?
- 한강과의 거리는 어떤가?
- 골드라인과의 거리는?
- 준공년도는 언제인가?

Importance– XGBRegressor

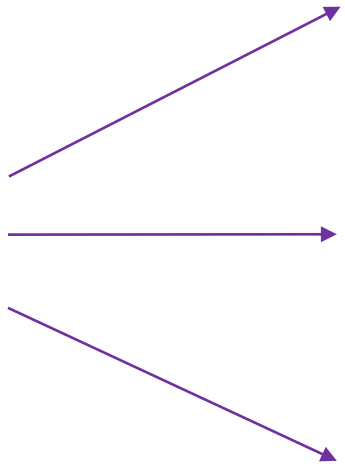
XGBoost

HyperParameter	설정 값	설명	구분	용도
max_depth	15	트리 깊이	Base function	Feature Space 분할!
min_child_weight	3	분할에 보수적인 성향		
colsample_bytree	0.7	피쳐 선택 비율		
reg_alpha	100	L1 regularization		
reg_lambda	1	L2 regularization		
Subsample	0.9	트레이닝 셋 사용율	부스팅 설정	모델 분산 줄이기!
learning_rate	0.01	학습율	최적화	최적해 찾기!
n_estimators	2175	부스팅 횟수		

하이퍼 파라미터 튜닝 과정 – XGBRegressor

하이퍼 파라미터 튜닝 과정

1. Base function 튜닝
2. 부스팅 설정 튜닝
3. 최적화 방법 튜닝



max_depth
min_child_weight
colsample_bytree
reg_alpha
reg_lambda

Subsample

learning_rate
n_estimators

[튜닝 방법]

Cross Validation

staged_predict
early_stopping

하이퍼 파라미터 튜닝 과정 – XGBRegressor

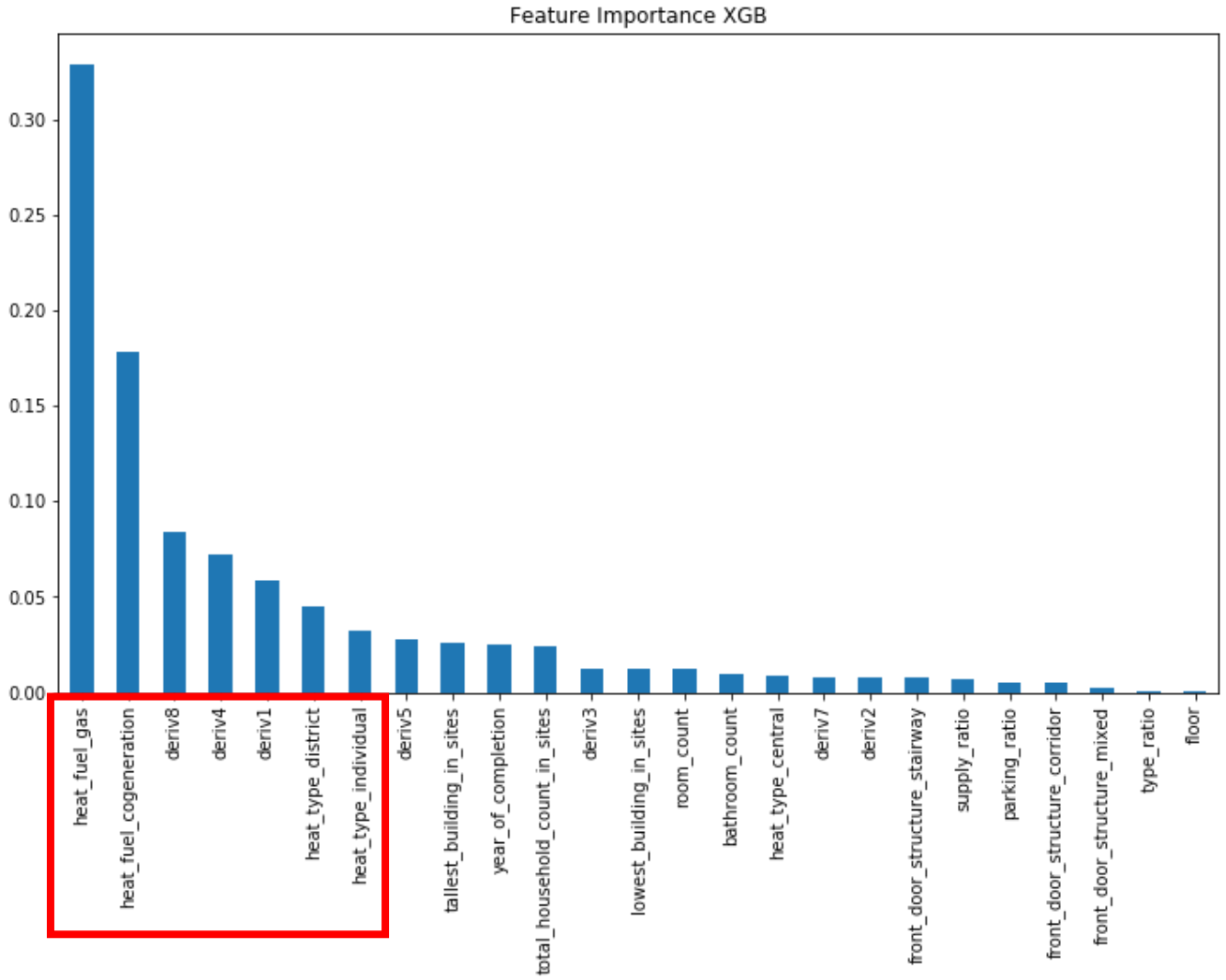
Cross Validation
GridSearchCV를 통해 최적의 하이퍼파라미터 선택

하이퍼파라미터
max_depth
Min_child_weight
Colsample_bytree
Reg_alpha
reg_lambda
learning_rate

Staged_predict() : early Stopping
CrossValidation 25번 연속 Error개선이 이루어지지 않으면 부스팅 중단

하이퍼파라미터
n_estimators

●Feature Importance– **XGBRegressor**



변수의 중요도

1. 열 난방 방식
2. 한강과의 거리
3. 골드라인과의 거리
4. 무슨 구인가

구의 종류는 25개이기 때문에 분할 횟수가 정해져있다.

깊게 파고들 경우엔 아파트 자체가 중요해진다는 결론

HyperParameter	설정 값	설명	TestError(RMSE)
GBRegressor			5033만원
XGRegressor			5081만원
Blended Model(Mean Model)			4966만원

THANK YOU