Boston Housing Data 데이터 분석 리포트

1. 개발 환경

언어 : Python (anaconda)
개발툴 : Spyder
사용패키지 : numpy, pandas, seaborn, sklearn, xgboost, fuctools

소스 파일 : boston housing data.py, analysis.py

실행 방법 :
IDE Spyder 사용시 : boston housing data.py 불러들여 실행
python.exe 사용시 : python boston housing data.py <엔터>

2. 분석 개요

해당 데이터 셋에 대하여, 먼저 탐색적 데이터 분석을 수행하여 기초 통계 및 변수 간의 상관 관계 또는 비/선형성을 확인하고 수행할 후보 모델링 알고리즘 별로 성능테스트를 수행 후 가장 성능이 좋게 나오는 알고리즘을 선정 한 다음 해당 알고리즘에 대하여 최적화 작업을 수행

변수 리스트 : 'CRIM','ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE','DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT', 'MEDV'

타켓 변수 : 'MEDV'

1) 탐색적 데이터 분석
2) 모델링 알고리즘 테스트 (일반 선형 회귀, PLS(PCA+Regression), 결정트리 ,그라디언트 부스팅)
3) 모델링 최적화 선정 및 모델 선정 분석 결과 도출

3. 탐색적 데이터 분석

1) 기초통계

이미 해당 데이터셋은 목적을 두고 전처리 과정을 거치고 생성된 공개 데이터 이기에 결측치 나 특별히 분산이 0 에 가까운 특별한 변수는 없었으나 'CHAS' 와 같은 경우에는 0/1 의 데이터로 일반 회귀 모형보다는 결정트리 계열의 알고리즘에서 유효할 것으로 보임

- 기초통계 요약

```
Housing_data summary
row counts=506 col counts=14
             CRIM          ZN       INDUS        CHAS         NOX          RM  \
count  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000
mean     1.716290   11.363636   11.136779    0.069170    0.554695    6.284634
std      2.653510   23.322453    6.860353    0.253994    0.115878    0.702617
min      0.006320    0.000000    0.460000    0.000000    0.385000    3.561000
25%      0.081900    0.000000    5.190000    0.000000    0.449000    5.885500
50%      0.250895    0.000000    9.690000    0.000000    0.538000    6.208500
75%      2.326718   12.500000   18.100000    0.000000    0.624000    6.623500
max      9.966540  100.000000   27.740000    1.000000    0.871000    8.780000

              AGE         DIS         RAD         TAX     PTRATIO           B  \
count  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000
mean    68.574901    3.696228    4.332016  408.237154   18.455534  356.674032
std     28.148861    1.999689    1.417166  168.537116    2.164946   91.294864
min      2.900000    0.585700    1.000000  187.000000   12.600000    0.320000
25%     45.025000    2.073700    4.000000  279.000000   17.400000  375.377500
50%     77.500000    3.107300    4.000000  330.000000   19.050000  391.440000
75%     94.075000    5.112625    5.000000  666.000000   20.200000  396.225000
max    100.000000    9.222900    8.000000  711.000000   22.000000  396.900000

            LSTAT        MEDV
count  506.000000  506.000000
mean    12.653063   22.532806
std      7.141062    9.197104
min      1.730000    5.000000
25%      6.950000   17.025000
50%     11.360000   21.200000
75%     16.955000   25.000000
max     37.970000   50.000000
```

2) 상관관계분석

타겟변수인 'MEDV' 와 LSTAT, RM, PTRATIO, INDUS, TAX 가 5개의 변수가 최상위 5 위로 큰 관계성을 보였으나, 'MEDV' 제외한 나머지 독립변수들 간의 상관관계가 크게 존재하여 일반적인 회귀분석으로는 적정 모델을 도출하기 힘들 것으로 보임

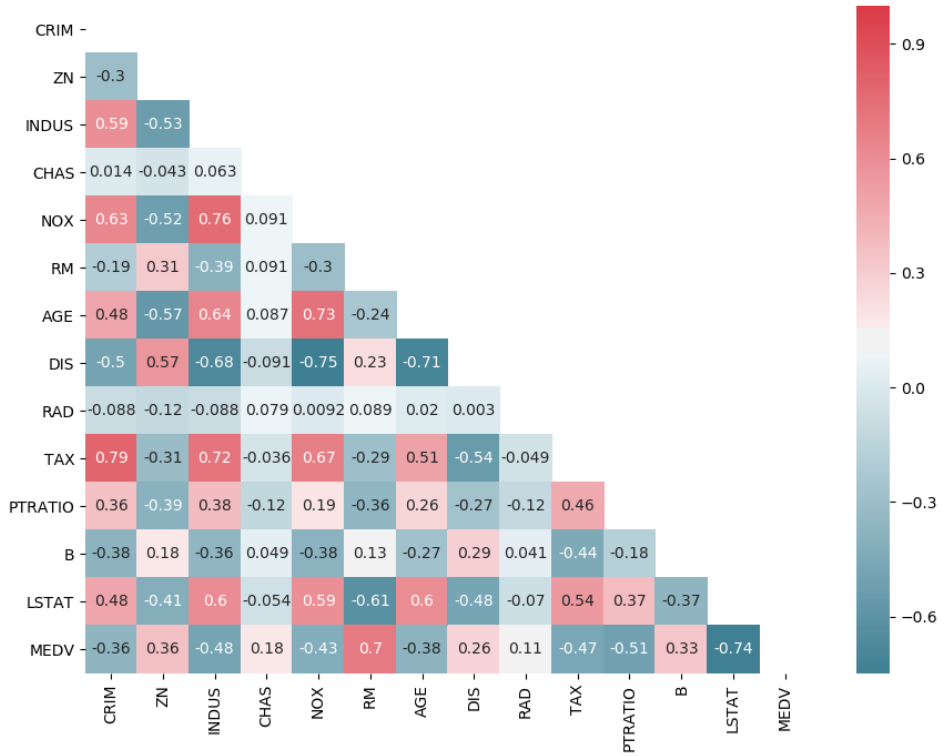
- 전체 변수 간의 상관관계 표

Housing_data Correlations

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
CRIM	1.000000	-0.300774	0.590822	0.013922	0.634679	-0.190197	0.482013
ZN	-0.300774	1.000000	-0.533828	-0.042697	-0.516604	0.311991	-0.569537
INDUS	0.590822	-0.533828	1.000000	0.062938	0.763651	-0.391676	0.644779
CHAS	0.013922	-0.042697	0.062938	1.000000	0.091203	0.091251	0.086518
NOX	0.634679	-0.516604	0.763651	0.091203	1.000000	-0.302188	0.731470
RM	-0.190197	0.311991	-0.391676	0.091251	-0.302188	1.000000	-0.240265
AGE	0.482013	-0.569537	0.644779	0.086518	0.731470	-0.240265	1.000000
DIS	-0.495148	0.566660	-0.678498	-0.090950	-0.748872	0.225052	-0.713313
RAD	-0.088451	-0.119290	-0.087615	0.079105	0.009217	0.088753	0.019658
TAX	0.793392	-0.314563	0.720760	-0.035587	0.668023	-0.292048	0.506456
PTRATIO	0.362615	-0.391679	0.383248	-0.121515	0.188933	-0.355501	0.261515
B	-0.377013	0.175520	-0.356977	0.048788	-0.380051	0.128069	-0.273534
LSTAT	0.481907	-0.412995	0.603800	-0.053929	0.590879	-0.613808	0.602339
MEDV	-0.362077	0.360445	-0.483725	0.175260	-0.427321	0.695360	-0.376955

	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	-0.495148	-0.088451	0.793392	0.362615	-0.377013	0.481907	-0.362077
ZN	0.566660	-0.119290	-0.314563	-0.391679	0.175520	-0.412995	0.360445
INDUS	-0.678498	-0.087615	0.720760	0.383248	-0.356977	0.603800	-0.483725
CHAS	-0.090950	0.079105	-0.035587	-0.121515	0.048788	-0.053929	0.175260
NOX	-0.748872	0.009217	0.668023	0.188933	-0.380051	0.590879	-0.427321
RM	0.225052	0.088753	-0.292048	-0.355501	0.128069	-0.613808	0.695360
AGE	-0.713313	0.019658	0.506456	0.261515	-0.273534	0.602339	-0.376955
DIS	1.000000	0.003030	-0.541369	-0.269140	0.293621	-0.479158	0.264325
RAD	0.003030	1.000000	-0.049221	-0.116969	0.040705	-0.069828	0.113519
TAX	-0.541369	-0.049221	1.000000	0.460853	-0.441808	0.543993	-0.468536
PTRATIO	-0.269140	-0.116969	0.460853	1.000000	-0.177383	0.374044	-0.507787
B	0.293621	0.040705	-0.441808	-0.177383	1.000000	-0.366087	0.333461
LSTAT	-0.479158	-0.069828	0.543993	0.374044	-0.366087	1.000000	-0.737663
MEDV	0.264325	0.113519	-0.468536	-0.507787	0.333461	-0.737663	1.000000

- 상관관계 히트맵 표시



해당 히트맵에서 볼 수 있듯이 타켓 변수인 'MDEV'와의 상관관계 외에도 독립변수간의 상관관계도 함께 두드러짐

```
Target Correlations
RM          0.695360
ZN          0.360445
B           0.333461
DIS         0.264325
CHAS        0.175260
RAD         0.113519
CRIM       -0.362077
AGE        -0.376955
NOX        -0.427321
TAX        -0.468536
INDUS      -0.483725
PTRATIO    -0.507787
LSTAT      -0.737663
Name: MEDV, dtype: float64
```

```
Important Correlations between Figures
    attribute pair  correlation
15     (CRIM, TAX)     0.793392
3     (INDUS, NOX)     0.763651
1       (AGE, NOX)     0.731470
7     (INDUS, TAX)     0.720760
20      (NOX, TAX)     0.668023
18    (AGE, INDUS)     0.644779
5      (CRIM, NOX)     0.634679
14  (INDUS, LSTAT)     0.603800
21    (AGE, LSTAT)     0.602339
11    (LSTAT, NOX)     0.590879
12   (CRIM, INDUS)     0.590822
6        (DIS, ZN)     0.566660
0     (LSTAT, TAX)     0.543993
19      (AGE, TAX)     0.506456
4        (NOX, ZN)    -0.516604
8      (INDUS, ZN)    -0.533828
17      (DIS, TAX)    -0.541369
16       (AGE, ZN)    -0.569537
9      (LSTAT, RM)    -0.613808
13    (DIS, INDUS)    -0.678498
10      (AGE, DIS)    -0.713313
2       (DIS, NOX)    -0.748872
```

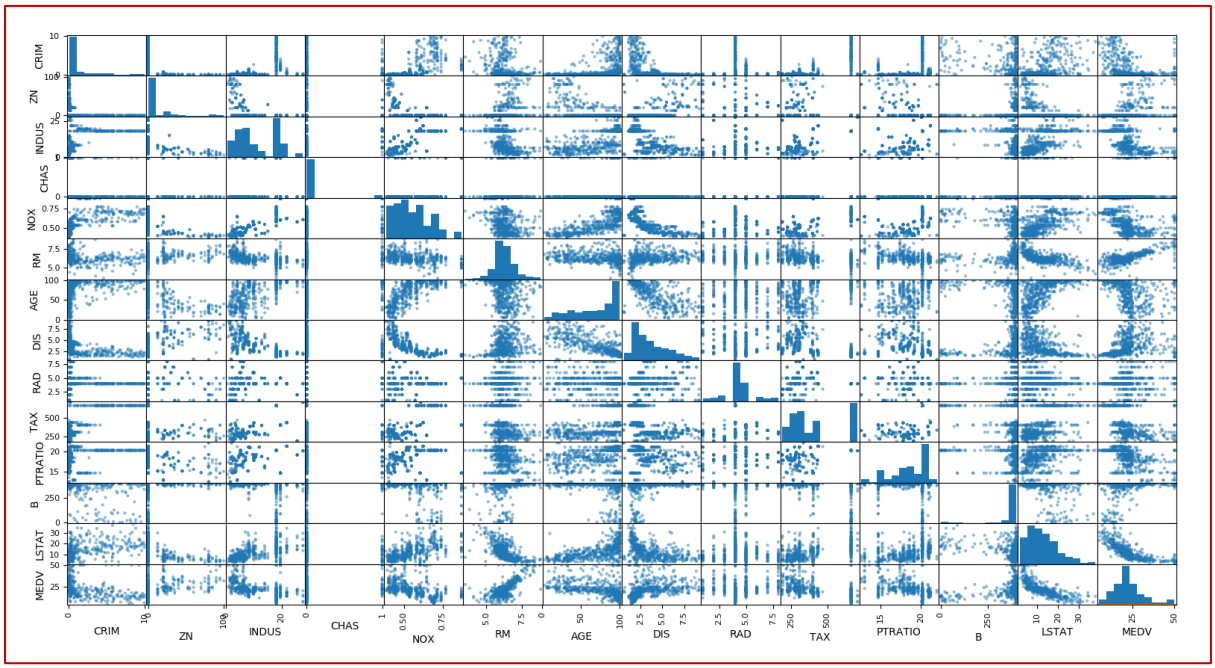
'LSTAT'-저소득 계층이 몰려 있으면 집값이 낮게 형성되며, 'RM'-방의 개수가 많으면 집값이 높게 형성되는 등 큰 상관관계를 보인 변수들에 대해서는 합리적인 설명이 됨

추가적으로 'NOX'-산화 질소 공기오염도와 관련된 부분에서 공장과 가까우면 공기오염도가 함께 올라가거나 'DIS'-업무지역과 멀어지면 공기오염도가 낮아지는 등의 부분들이 다양하게 보임

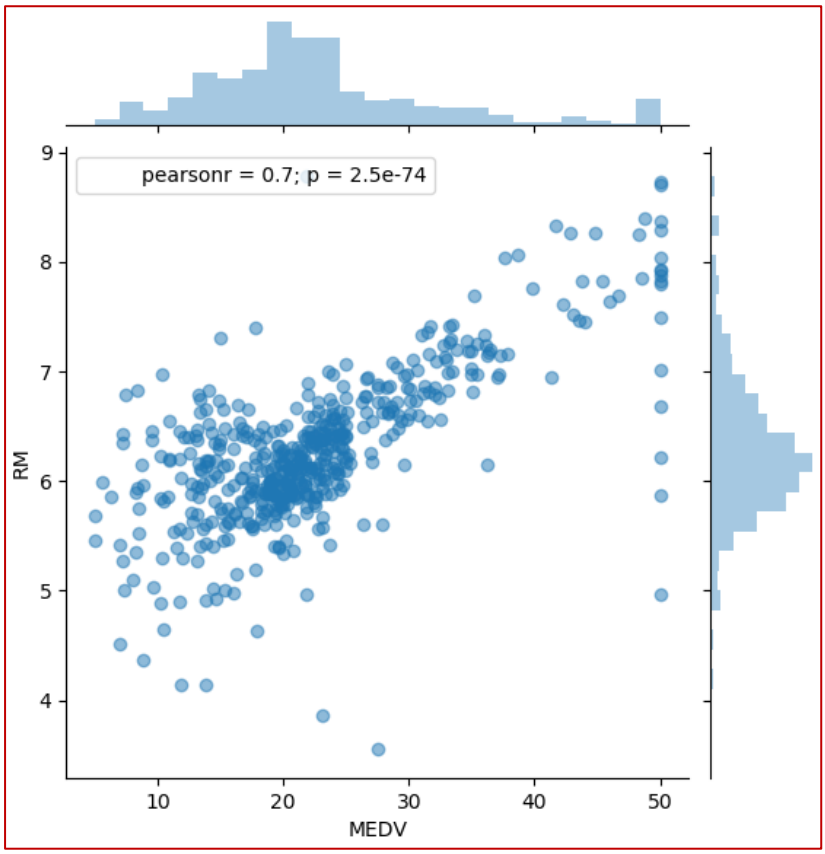
3) 매트릭스 차팅

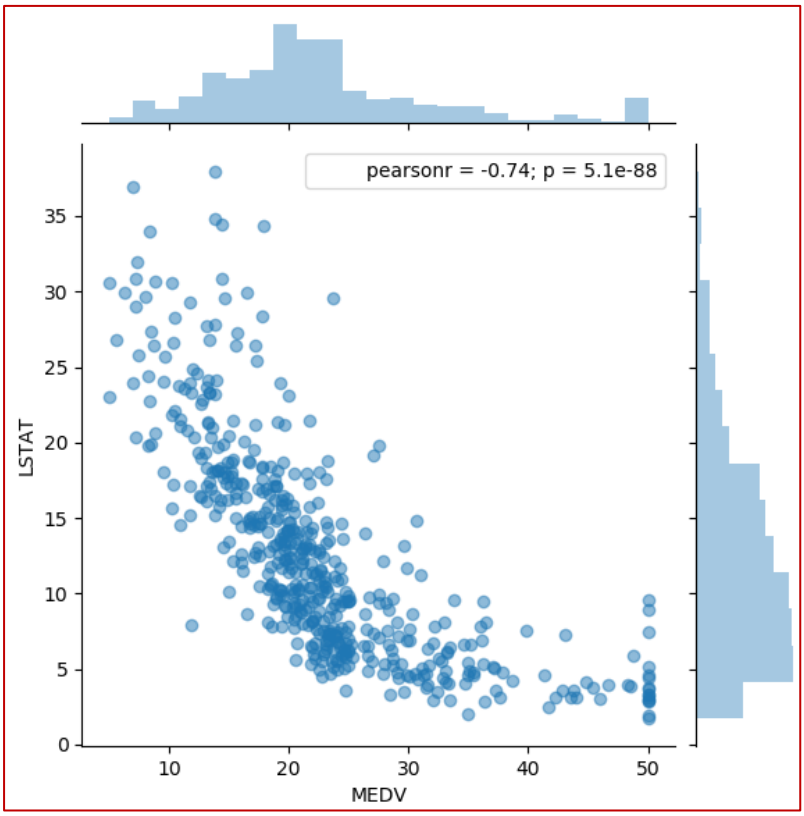
매트릭스 차트를 통해 한번 더 변수 간의 관계성을 확인하고 상위의 상관관계를 보인 주요인자와 타겟 변수와의 선형성을 확인 함

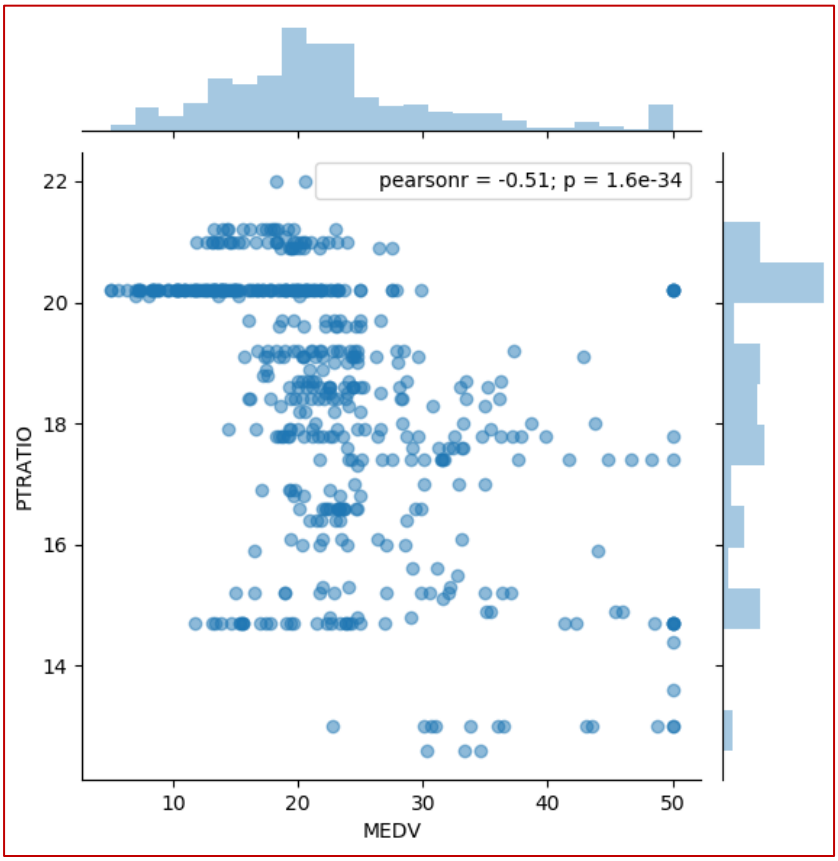
일반적인 선형성을 보이는 변수도 있었으나 의외로 로그,지수 곡선을 그리는 변수들 존재하였음. 회귀모형을 주로 한다면 해당 변수들에 대해 별도의 식을 세워 파생 변수로 선형성을 확보하는 작업을 진행하면 더 좋을 것 같았으나 이미 앞서 독립변수 간에 상관관계가 큰 것을 확인했으므로 결정트리를 주로 사용할 것이기에 해당 부분은 생략함



- 'MEDV' 와 주요 독립 변수(RM, LSTAT, PTRATIO) 분포




pearsonr = -0.74; p = 5.1e-88
LSTAT
MEDV
0
5
10
15
20
25
30
35
10
20
30
40
50

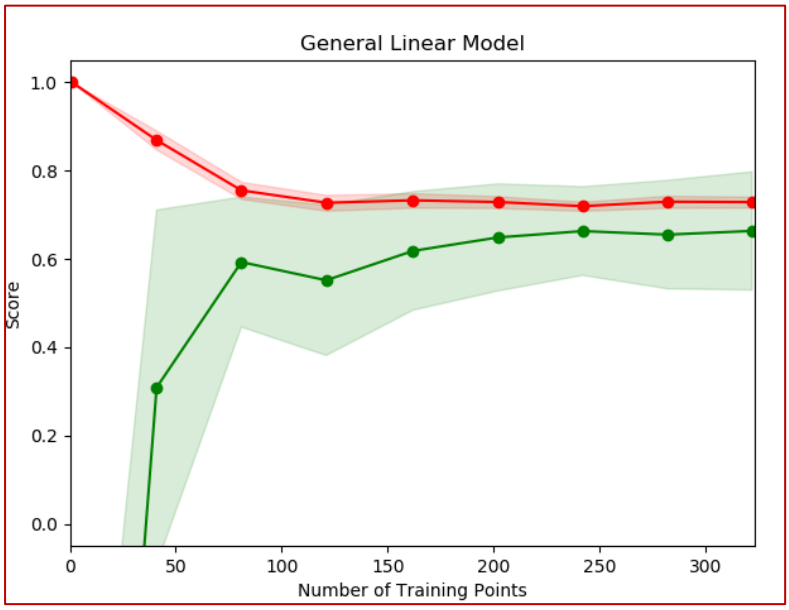

pearsonr = -0.51; p = 1.6e-34
PTRATIO
MEDV
14
16
18
20
22
10
20
30
40
50

4. 모델링 알고리즘 테스트

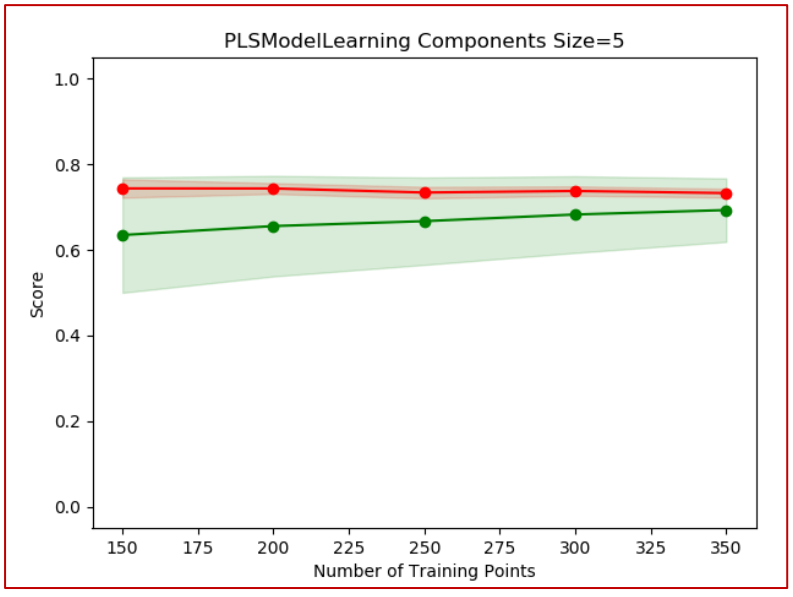
모든 알고리즘은 트레인셋과 테스트셋을 나누어 크로스벨리데이션 과정을 적용하였고 스코어는 R2 설정하여 해당 성능을 비교하기 쉽게 차트로 표현하였음

일반회귀 : 성능이 좋지 하겠지만 기준을 세우기 위해 일반 회귀 모형에서의 상태 확인 차 확인
PLS(PCA+Regression) : 상관관계가 독립변수간에 존재하기에 데이터 압축하여 회귀 모형 적용
Descision Tress : 독립변수간의 상관관계가 존재할 뿐아니라 선형적이지 않은 부분도 존재하기에 결정트리 사용
GBoost : 부스팅 기법을 적용하여 좀 더 결정트리의 성능을 높인 모델을 찾도록 함

- 일반 회귀 결과 차트

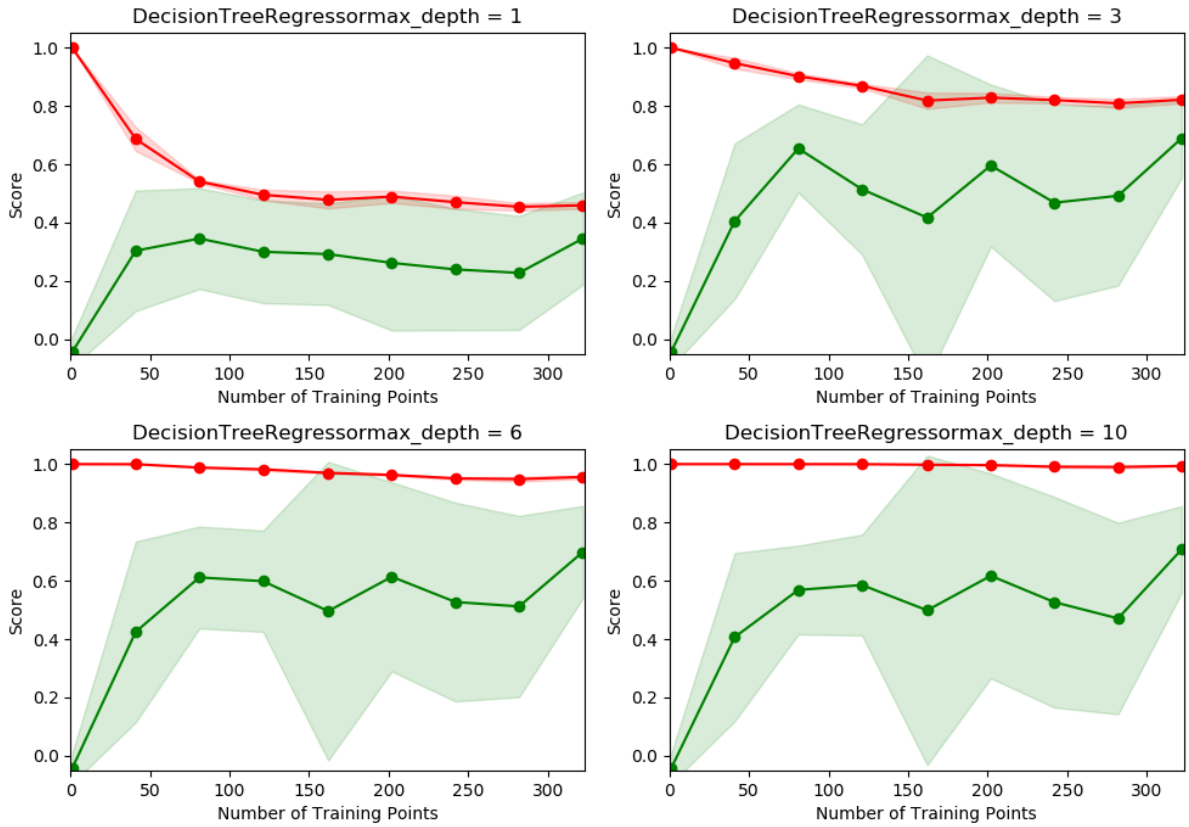


-PLS 결과 차트



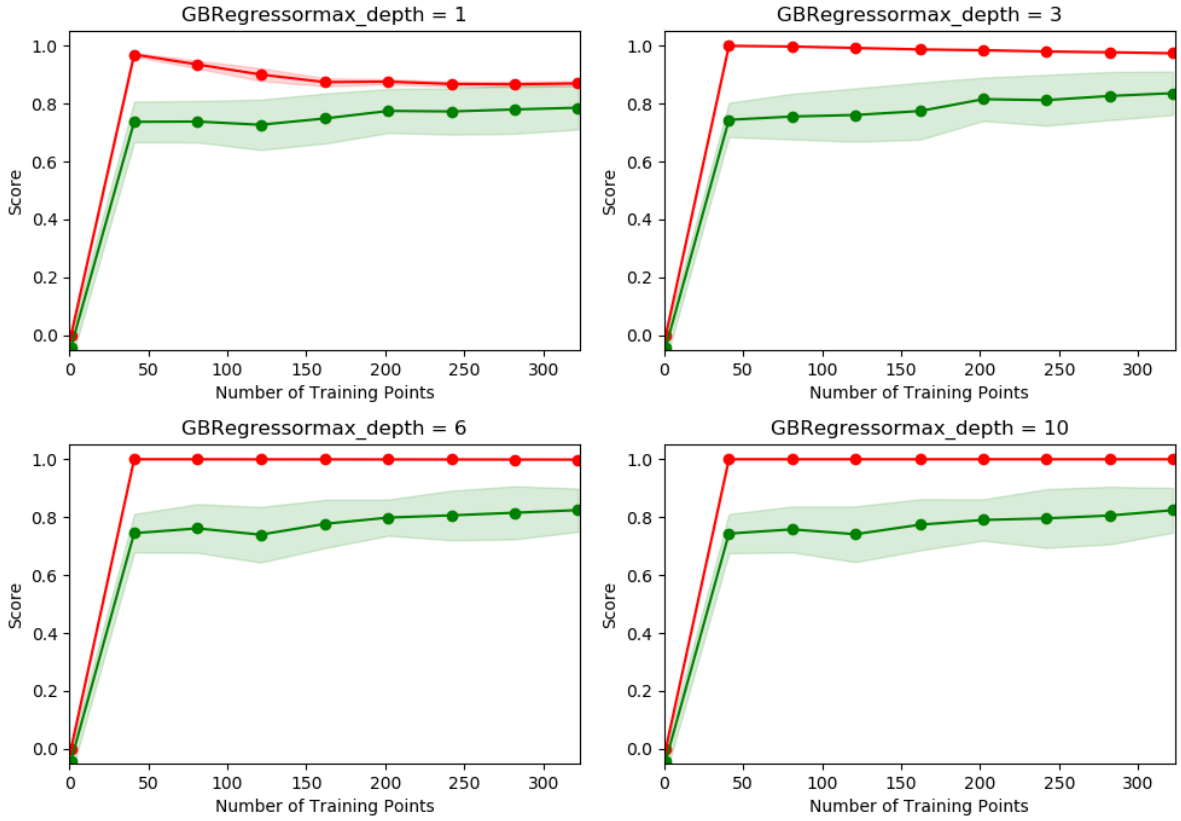
일반 회귀 모형 보다 PLS 가 안정적인 성능을 보여줌, 주성분 개수 3개 이상 부터는 큰 변동 폭이 없었음

- 결정트리



트리 깊이를 5 이상 부터는 비슷한 성능을 보이며, 트레이닝 셋에서는 일반 회귀보나 훨씬 좋은 성능을 보이나 오버피팅이 많이 일어남. 해당 부분은 샘플링 다시 하거나하여 어느정도 조정 가능함

-그라디언트 부스팅



일반 회귀 모형 또는 일반 결정 트리 보다 훨씬 더 좋은 성능을 테스트 셋과 트레인셋 양쪽다 보두 보이며 본 분석에서는 그라디언트 부스팅 알고리즘을 통해 모델 최적화를 수행하기로 함

5. 모델링 최적화 선정 및 분석 결과 도출

크로스벨리데이션을 적용함과 동시에 sklearn 패키지에서 제공하는 RandomizedSearchCV 를 활용하여, 각 모델의 파라메타 (max_depth, learning_rate, gamma, reg_lambda) 을 주어진 범위안에서 무작위로 탐색하여 최적의 모델을 찾도록 함

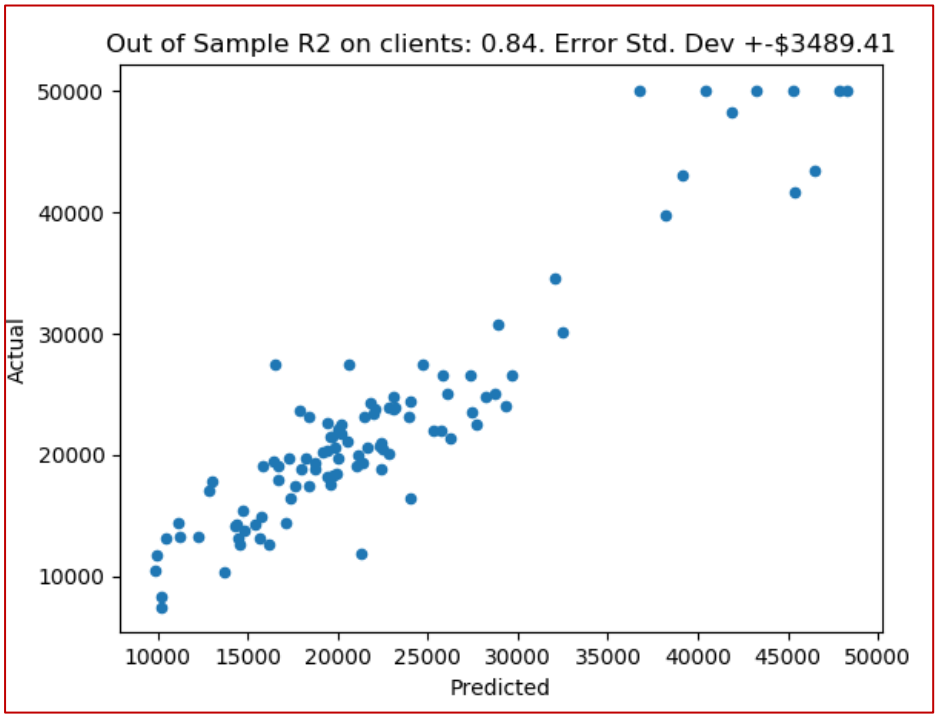
코드상에서는 보다 좋은 모델을 찾기 위해 옵션을 조절하여 찾을 가능성이 있음

Code:

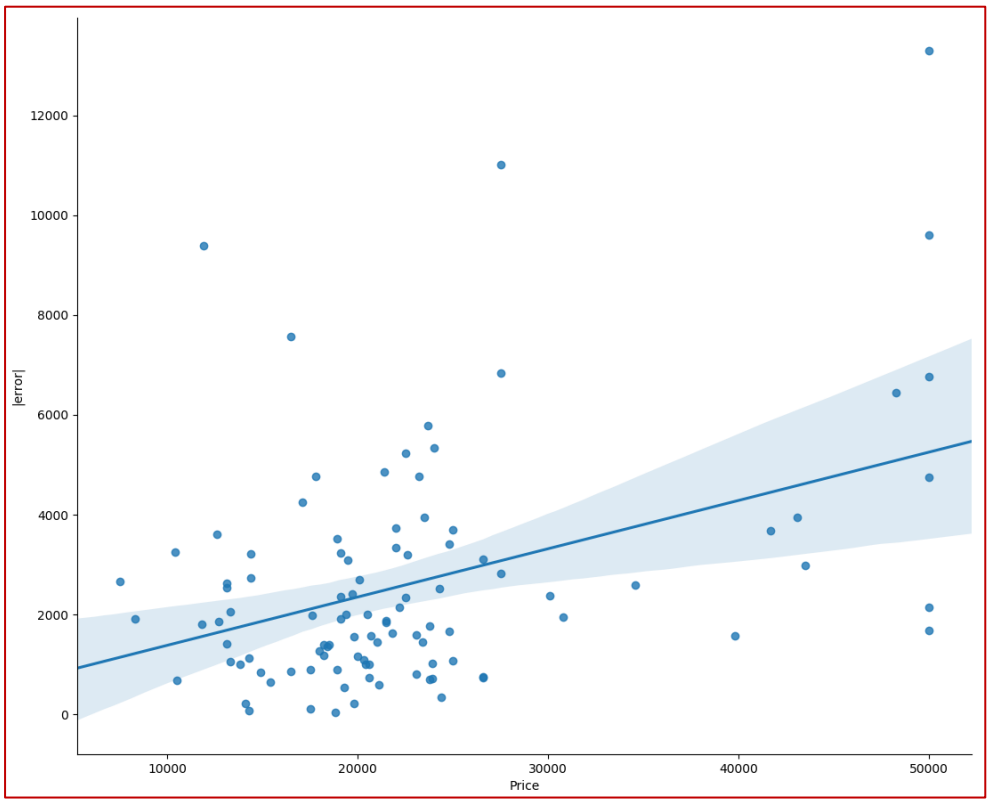
def get_optimal_GBModel(X, y):

상기 함수에 구현 됨

- 최적 모델 탐색 결과



테스트 셋으로 남겨둔 데이터로 최적 모델 탐색 결과 R2 : 0.84 로 앞서 테스트한 모델들에 비워 좋은 성능의 모델을 찾음



그라디언트 부스팅을 사용하여 해당 집값을 예측하는 최적의 모델을 찾을 수 있었으며, 변수간의 관계를 결정트리를 확인하는 과정을 통해 다시한번 확인이 가능하였음

결정트리를 확인하게 되면 타겟 변수와 관계성이 높았던 독립변수에 의해 트리가 생성됨