Boston Housing을 통한 집값 분석

A반 1조 최우혁

Boston Housing MEDV

가설

1. 방이 많으면 비쌀 것이다.
2. 저소득층이 많으면 저렴할 것이다.
3. 중심부에 가까울수록 비쌀 것이다.
4. 환경이 나쁘면 저렴할 것이다.
5. 범죄율이 높으면 저렴할 것이다.
   (= 저렴한 곳의 범죄율이 높을 것이다.)

-> 예상되는 주요 Parameter:
RM, LSTAT, DIS, NOX, CRIM

1. 데이터 확인

이상치 및 결측치 확인, 필요시 수정 / 제거
그래프를 통한 데이터 경향성 확인

➔ **이상치**

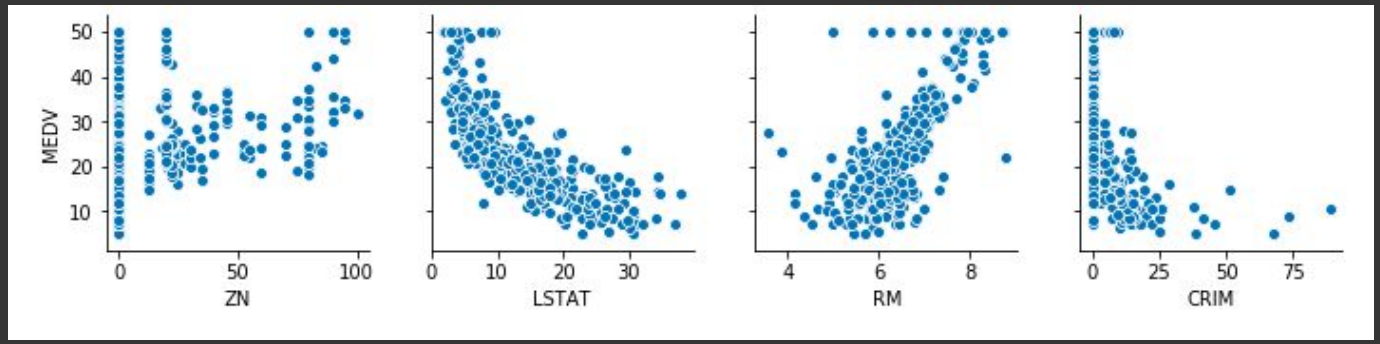
유명한 데이터 셋으로 에러로 인한
이상치는 없을 것, 주어진 기존 데이터
최대한 유지

➔ **결측치**

결측치가 있을 시, 가설에 의거하여
비슷한 조건의 평균값으로 대체

➔ **그래프**

Pairplot으로 대략적인 Parameter 유추
의심되는 Parameter는 확대 조사



	MEDV	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
MEDV	1.00	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74

그래프와 상관관계 분석

가설에 부합되는 경향 (LSTAT, RM 등)
보임, 실제 모델로 검증 필요

Comment

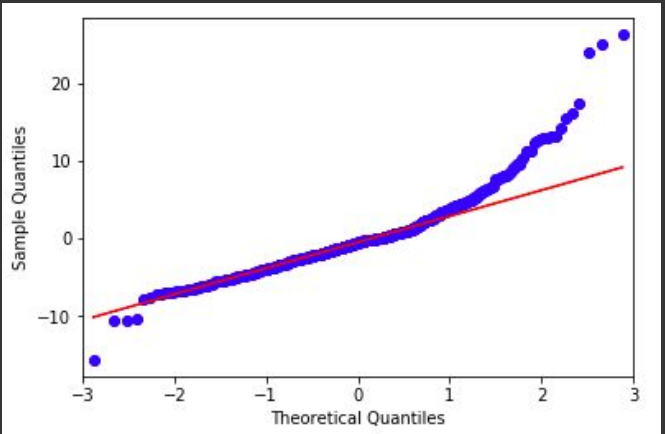
RM, CRIM에서 한 X에 대해 수많은 Y가 분포하거나, 그 반대 경우를 확인 가능. 하지만 과반수 이상은 선형적인 관계를 따르는 것으로 '보임'. 이상치인지, 데이터의 부족으로 인해 확인되지 않는 패턴인지 알 수 없으므로 두 방향으로 분석.

	variable	VIF
4	CHAS	1.074
12	B	1.349
1	CRIM	1.792
11	PTRATIO	1.799
6	RM	1.934
2	ZN	2.299
13	LSTAT	2.941
7	AGE	3.101
8	DIS	3.956
3	INDUS	3.992
5	NOX	4.394
9	RAD	7.484
10	TAX	9.009
0	const	585.265

OLS Regression Results

Dep. Variable:	MEDV	**R-squared:**	0.741
Model:	OLS	**Adj. R-squared:**	0.734
Method:	Least Squares	**F-statistic:**	108.1
Date:	Sat, 04 May 2019	**Prob (F-statistic):**	6.72e-135
Time:	09:08:57	**Log-Likelihood:**	-1498.8
No. Observations:	506	**AIC:**	3026.
Df Residuals:	492	**BIC:**	3085.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	36.4595	5.103	7.144	0.000	26.432	46.487
CRIM	-0.1080	0.033	-3.287	0.001	-0.173	-0.043
ZN	0.0464	0.014	3.382	0.001	0.019	0.073
INDUS	0.0206	0.061	0.334	0.738	-0.100	0.141
CHAS	2.6867	0.862	3.118	0.002	0.994	4.380
NOX	-17.7666	3.820	-4.651	0.000	-25.272	-10.262
RM	3.8099	0.418	9.116	0.000	2.989	4.631
AGE	0.0007	0.013	0.052	0.958	-0.025	0.027
DIS	-1.4756	0.199	-7.398	0.000	-1.867	-1.084
RAD	0.3060	0.066	4.613	0.000	0.176	0.436
TAX	-0.0123	0.004	-3.280	0.001	-0.020	-0.005
PTRATIO	-0.9527	0.131	-7.283	0.000	-1.210	-0.696
B	0.0093	0.003	3.467	0.001	0.004	0.015
LSTAT	-0.5248	0.051	-10.347	0.000	-0.624	-0.425



Comment

다중공선성 확인 결과 모든 파라미터는 수용 가능 확인

회귀 분석 결과, INDUS와 AGE는 선형적으로 유의하지 않음

잔차 분석 결과, Theoretical Quantities = -2 ~ 1까지는 정규분포를 따르지만, 그 밖에서는 y = x^3 그래프처럼 휘는 것을 볼 수 있음

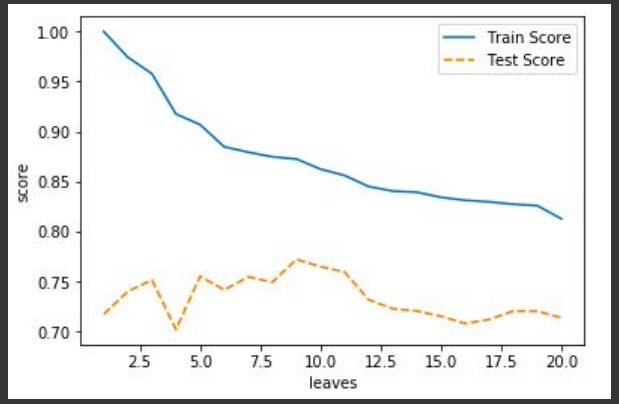
2-1. Decision Tree

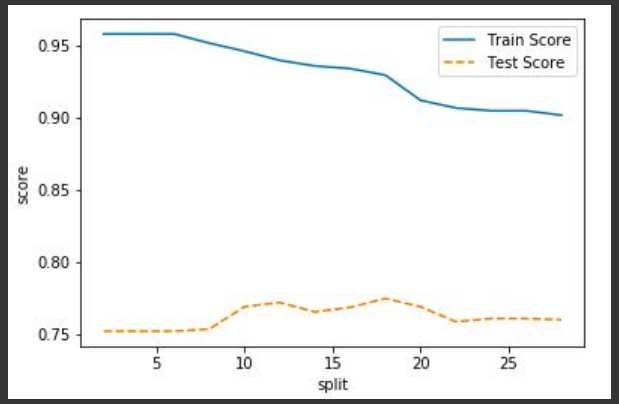
현재 가진 정보로는 회귀식을 만들고 결론을
내릴 수 없으므로, 다른 모델을 사용하여 분석

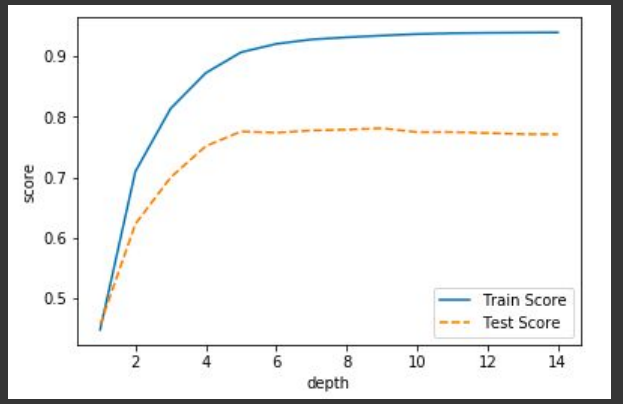
해당 분석에서는 이상치를 경향성의 연장선에
있다고 가정하여, 이상치를 제거하지 않고 분석

➔ **의사결정나무**

결과 해석에 용이
자료를 가공할 필요가 없음







의사결정나무 모델 세팅

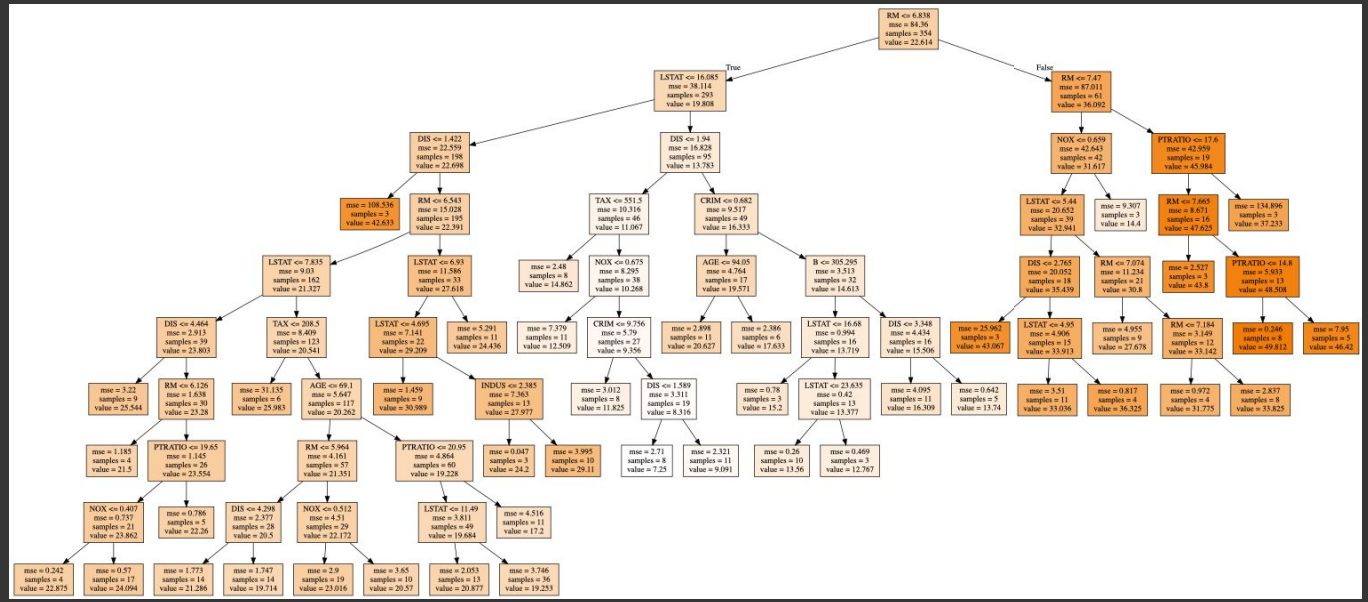
max_depth = 9

min_samples_split = 12

Min_samples_leaf = 3

Comment

Train Score와 Test Score 모두 높고, 기울기가 갑작스럽게 변하거나, Score가 증가 -> 감소로 바뀌는 지점으로 모델을 세팅



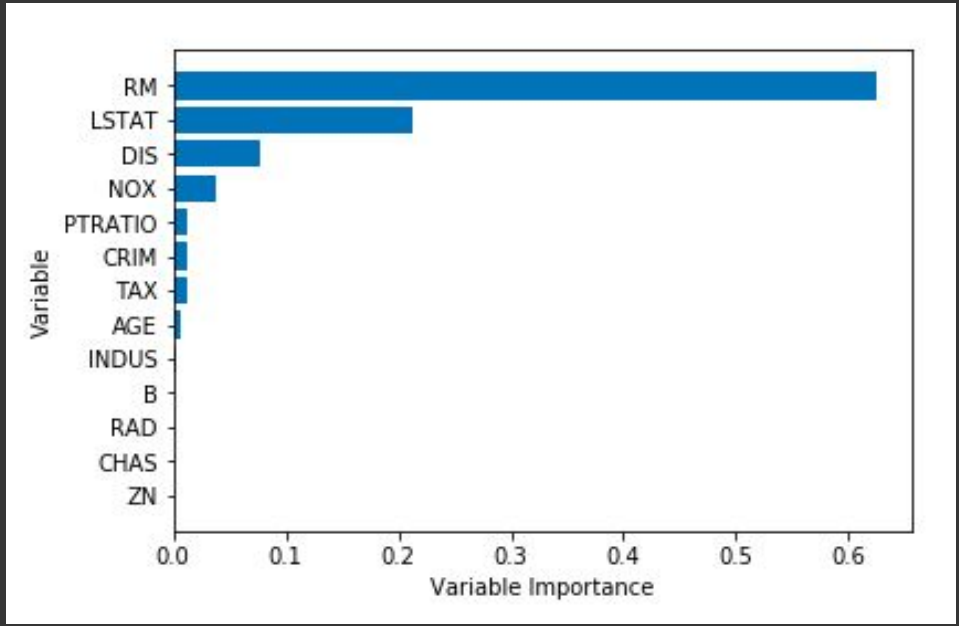
Training score: 0.934

Test score: 0.781

Comment

Test Score가 0.781로 비교적 높은 수치를 보이지만, 고작 506개의 데이터 잎사귀 노드 24개로 나누는 것은 과적합이라고 판단됨.

	Feature	Importance
5	RM	0.626
12	LSTAT	0.212
7	DIS	0.078
4	NOX	0.039
10	PTRATIO	0.013
0	CRIM	0.012
9	TAX	0.012
6	AGE	0.006
2	INDUS	0.002
11	B	0.001
1	ZN	0.000
3	CHAS	0.000
8	RAD	0.000



Parameter

RM, LSTAT, DIS, NOX, PTRATIO, CRIM

Comment

분석 결과 MEDV에 영향을 많이 주는 Parameter는 RM, LSTAT, DIS, NOX, PTRATIO, CRIM로 대부분 기존의 가설과 일치함

2-2. Gradient Boosting

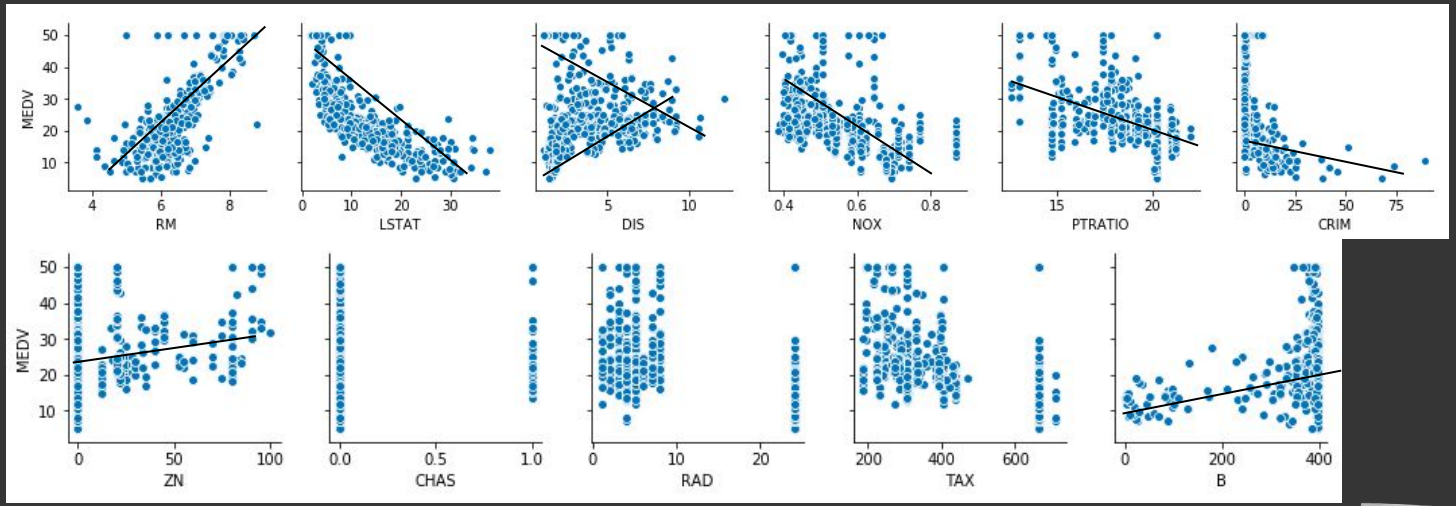
눈에 띄는 이상치만 제거한 모델을 만들어서,
이상치가 있을 때의 모델과 비교. 의사결정 나무
모델 사용시 0.7대에 머물렀던 성능을
끌어올리기 위해 Gradient Boosting 사용.
이상치를 제거한 데이터를 사용할 것이므로 2-1
모델과 직접적인 비교는 어렵지만, 대략적인
차이를 보기 위함

➔ **이상치 제거**

중요하게 생각되는 Parameter의
산점도를 그려서 경향에서 **크게** 벗어나는
이상치만 제거

➔ **Gradient Boosting**

먼저 학습된 결과가 다음 학습에 영향을
주면서 성능을 최대화하는 학습 기법



이상치 제거

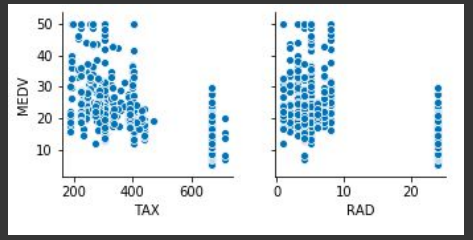
1. **RAD > 20 && MEDV > 50**
2. **TAX > 600 && MEDV > 50**

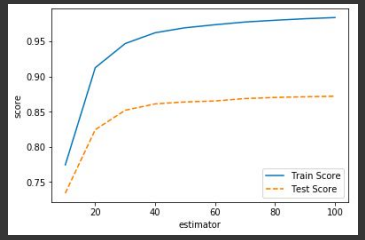
이상치의 개수가 적어서 크게 영향을 끼치지 않을 것이라고 기대됨, 예상과 다르게 모델 2-1과 비교해도 괜찮을 것으로 기대

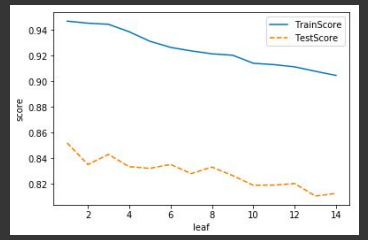
Comment

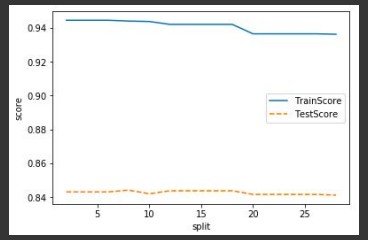
주요 Parameter에서 이상치를 찾으려했지만, 대부분 경향에서 크게 벗어나지 않으며, RM 같은 경우 현재 데이터에서만 이상치로 확인될 수 있다고 판단

(중심지의 고급 스튜디오 등)










```
Score on training: 0.9963352570272895
Score on test: 0.8458540717674071
```

```
1 gb_fin
```

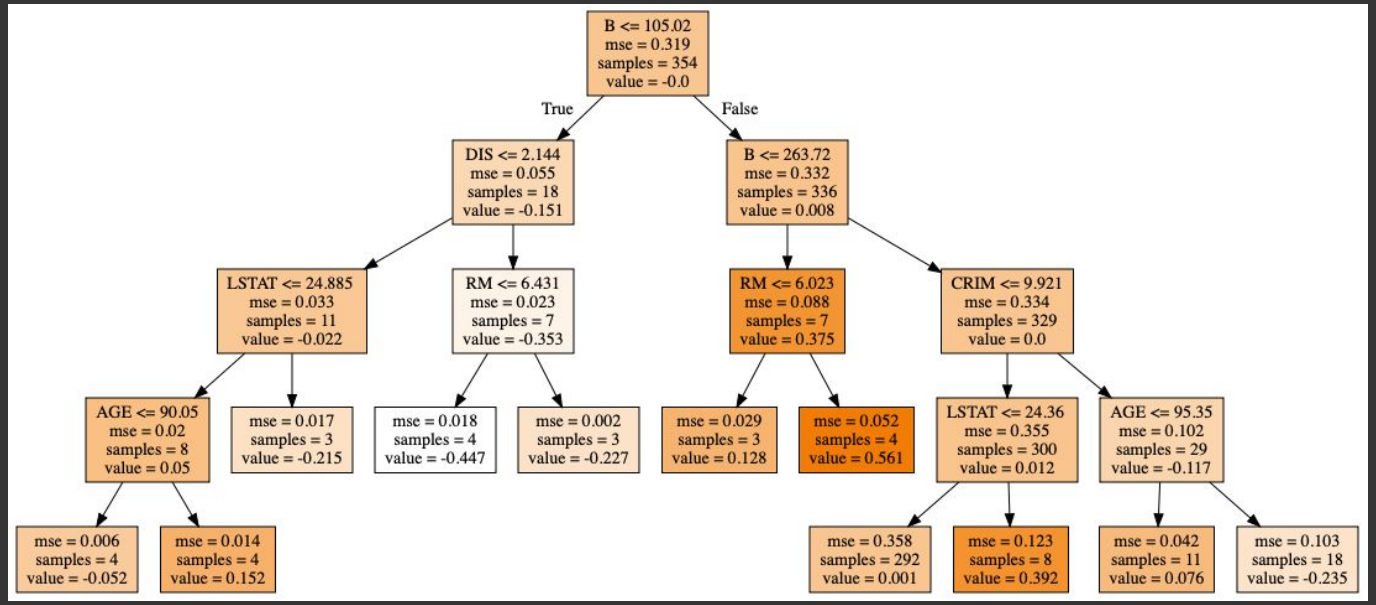
```
GradientBoostingRegressor(alpha=0.9, criterion='mse', init=None,
             learning_rate=0.5, loss='ls', max_depth=4, max_features=None,
             max_leaf_nodes=None, min_impurity_decrease=0.0,
             min_impurity_split=None, min_samples_leaf=3,
             min_samples_split=2, min_weight_fraction_leaf=0.0,
             n_estimators=30, n_iter_no_change=None, presort='auto',
             random_state=777, subsample=1.0, tol=0.0001,
             validation_fraction=0.1, verbose=0, warm_start=False)
```

이상치 제거, 모델 제작

Learning rate = 0.5, Max_depth = 4
Min_samples_split = 2, Min_samples_leaf = 3

Comment

Decision Tree보다 깊이, 잎사귀의 수는 적으면서 스코어는 10% 가량 더 높아짐.



Training score: 0.996

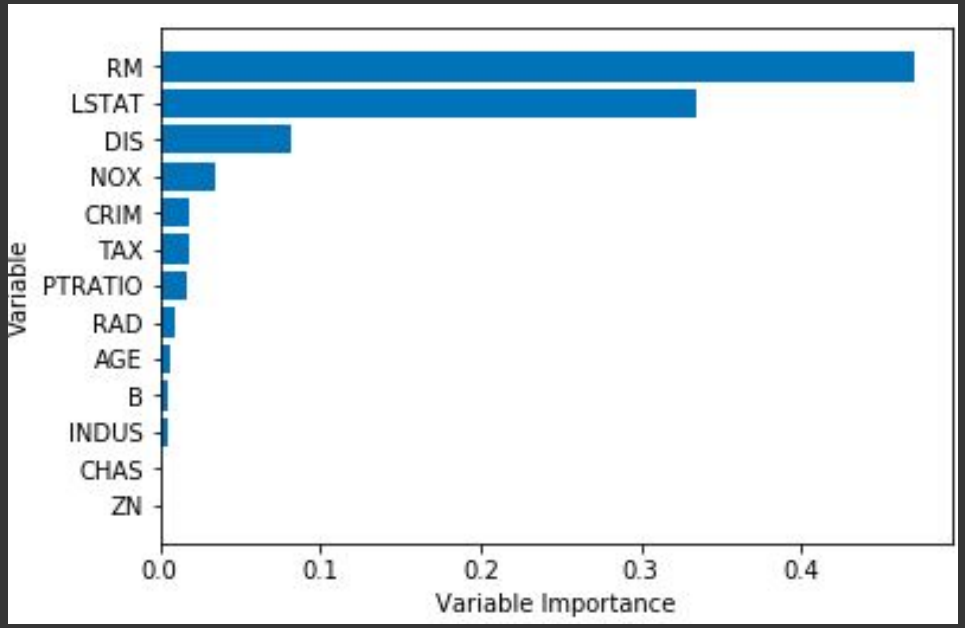
Test score: 0.846

Comment

깊이와 잎사귀 수 모두 줄어들은 것을 알 수 있음

Test Score 역시 10% 가량 증가

	Feature	Importance
5	RM	0.471
12	LSTAT	0.334
7	DIS	0.081
4	NOX	0.034
0	CRIM	0.018
9	TAX	0.018
10	PTRATIO	0.016
8	RAD	0.009
6	AGE	0.007
11	B	0.005
2	INDUS	0.004
3	CHAS	0.001
1	ZN	0.000



Parameter
RM, LSTAT, DIS, NOX, CRIM, TAX

Comment

분석 결과 MEDV에 영향을 많이 주는 Parameter는 RM, LSTAT, DIS, NOX, PTRATIO, CRIM로 대부분 기존의 가설과 일치함

2-1모델과도 대부분 일치

3. 결론

➔ **Decision Tree**

이상치 제거 X
분석 결과 test 데이터에 대해서 상당히 높은 스코어를 보이지만, Depth와 잎사귀 노드 수를 볼 때, 모델이 과적합되어 이상이 있을 것으로 보임

➔ **Gradient Boosting**

이상치 제거
분석 결과 training 데이터에 0.99의 스코어, test 데이터에 0.84의 스코어를 보임, Depth, Leaves 모두 양호

주어진 데이터의 잔차 분석을 볼 때, MEDV는 Parameter가 작아질수록 기울기가 더 작아지고, 커질수록 기울기가 더 커지는 비선형적 관계를 가졌을 것으로 의심됨

개선 방안

1. 이상치 확인을 위해 더 많은 데이터가 필요함

(크롤링, 설문조사)

2.MEDV와 선형적인 관계가 있는 파생 변수가 있다면 더 나은 회귀식을 얻을 수 있을 것

(도메인 전문가)

3.비선형적인 목표변수를 설명할 수 있는 방법이 필요함