보스턴 주택 가격 Report

Jongkook Choi
jongkook90@gmail.com

Second order polynomial regression 및 gradient boosting regression을 통해 데이터를 피팅하여 RMSE=6% 정도의 모델을 얻었습니다. SHAP을 통해 모델을 분석한 결과, 하위계층 비율(LSTAT)이 음의 방향으로, 방 개수(+RM)가 양의 방향으로 각각 주택 가격 예측에 가장 큰 영향을 주었으며, 특히 방 개수는 많을 때 model에 impact가 큰 것을 확인하였습니다. Model 생성을 위한 feature selection은 correlation에 기반한 score를 이용하여 greedy하게 선택하였습니다. 추가적으로, deep learning 모델과의 비교를 수행하였습니다.

데이터 로드

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00
mean	3.61	11.36	11.14	0.07	0.55	6.28	68.57	3.80	9.55	408.24	18.46	356.67	12.65	22.53
std	8.60	23.32	6.86	0.25	0.12	0.70	28.15	2.11	8.71	168.54	2.16	91.29	7.14	9.20
min	0.01	0.00	0.46	0.00	0.39	3.56	2.90	1.13	1.00	187.00	12.60	0.32	1.73	5.00
25%	0.08	0.00	5.19	0.00	0.45	5.89	45.02	2.10	4.00	279.00	17.40	375.38	6.95	17.02
50%	0.26	0.00	9.69	0.00	0.54	6.21	77.50	3.21	5.00	330.00	19.05	391.44	11.36	21.20
75%	3.68	12.50	18.10	0.00	0.62	6.62	94.07	5.19	24.00	666.00	20.20	396.23	16.96	25.00
max	88.98	100.00	27.74	1.00	0.87	8.78	100.00	12.13	24.00	711.00	22.00	396.90	37.97	50.00

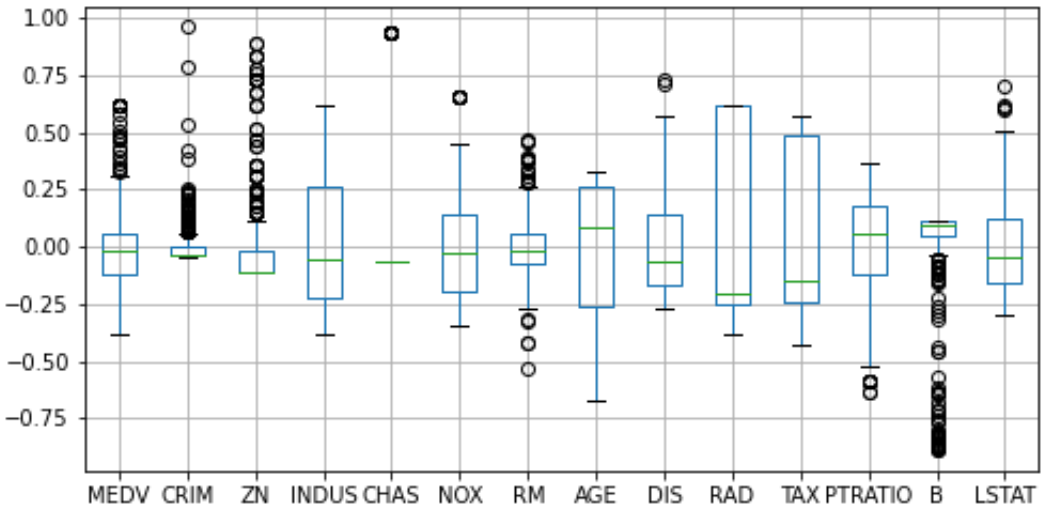
- Harrison (1978)과 평균값, Belsley (1980)와 첫 데이터를 비교하여 이상 없음을 확인했습니다.
- CRIM - 범죄율 (1인당)
- ZN - large-lot zone 비율 (25000sqft 이하의 주택이 금지되는 구역)
- INDUS - 산업지구 비율 (공해도)
- CHAS - 더미 변수 (찰스강 경계에 위치 = 1, 아니면 = 0)
- NOX - 일산화질소 농도 (단위 : 10pphm, parts per 10 million)
- RM - 가구당 방의 개수 (부동산 품질)
- AGE - 오래된 자가주택 비율 (1940년 기준, 구조물 품질에 관련됨)
- DIS - 고용센터까지의 거리 (통근거리, 가중평균, 로그)
- RAD - 고속도로 접근성 (공해도, 통근거리, 로그)
- TAX - 재산세율

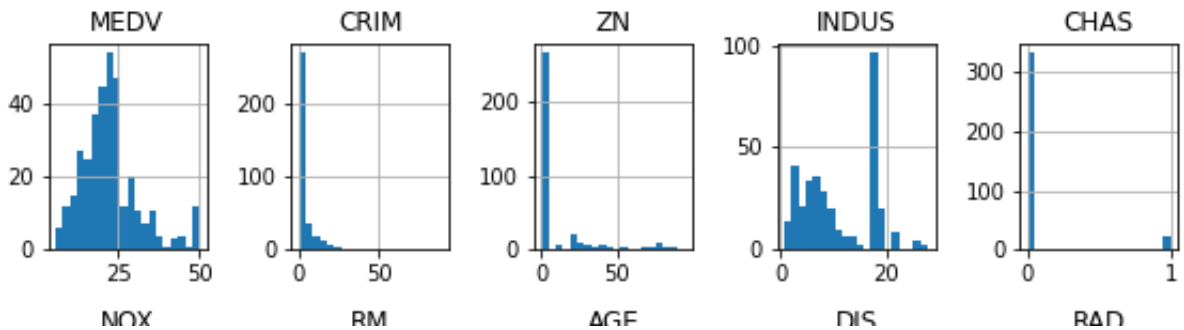
- PTRATIO - 학생-교사 비율
- B - 흑인 비율 (1000(Bk - 0.63)^2)
- LSTAT - 하위계층 %비율
- MEDV - 자가주택 가격 중앙값 (단위 $1000)

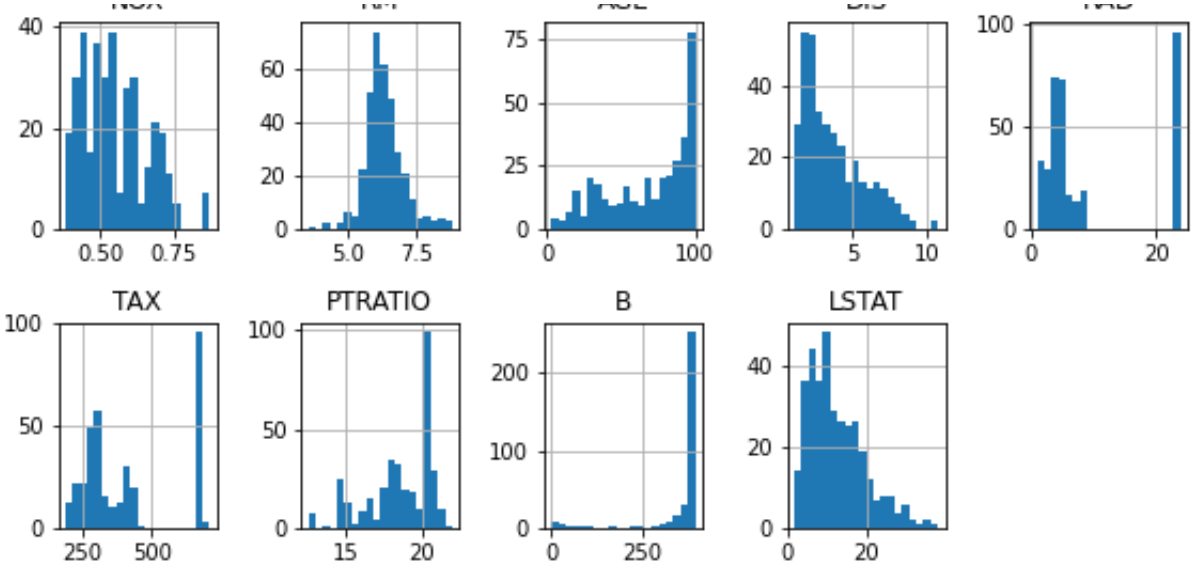
- 논문에서는 변수를 다음과 같이 구분하였습니다.
- 종속변수 : MEDV
- Structural : RM, AGE
- Neighborhood : B, LSTAT, CRIM, ZN, INDUS, TAX, PTRATIO, CHAS
- Accessibility : DIS, RAD
- Air Pollution : NOX, PART

- 데이터 로드 후, 용도별로 분리하였습니다. (shuffle 후 training 70%, test 20%, validation 10% 로 분리)

데이터 확인







- target인 MEDV의 최대값(50)의 빈도가 너무 높습니다. clipping되었거나 문제가 있는 값이라고 판단됩니다. 그 외의 feature도 대체로 분포가 skew되어 있지만, 주택 데이터의 특성상 실제 분포가 그럴 수도 있다고 생각됩니다.
- Category 데이터인 CHAS feature는 0과 1로 이루어져 있습니다.
- RM, LSTAT, AGE, DIS는 대체로 실수형 데이터로 보입니다.
- TAX는 multimodal한 데이터입니다. Harrison (1978)에 따르면 town간 과세평가적용률이 무시된 데이터입니다. 두개의 node가 두 가지 과세평가적용률을 의미할 수 있습니다. 만약 그렇다면, 과세평가적용률 데이터가 있다면 조합하여 교차 특성을 만들면 모델이 향상될 가능성이 있습니다.

데이터 상관관계 확인

- MEDV의 이상치를 제거하고 분석하였습니다.

- Correlation matrix(signed, unsigned), hierarchical clustering

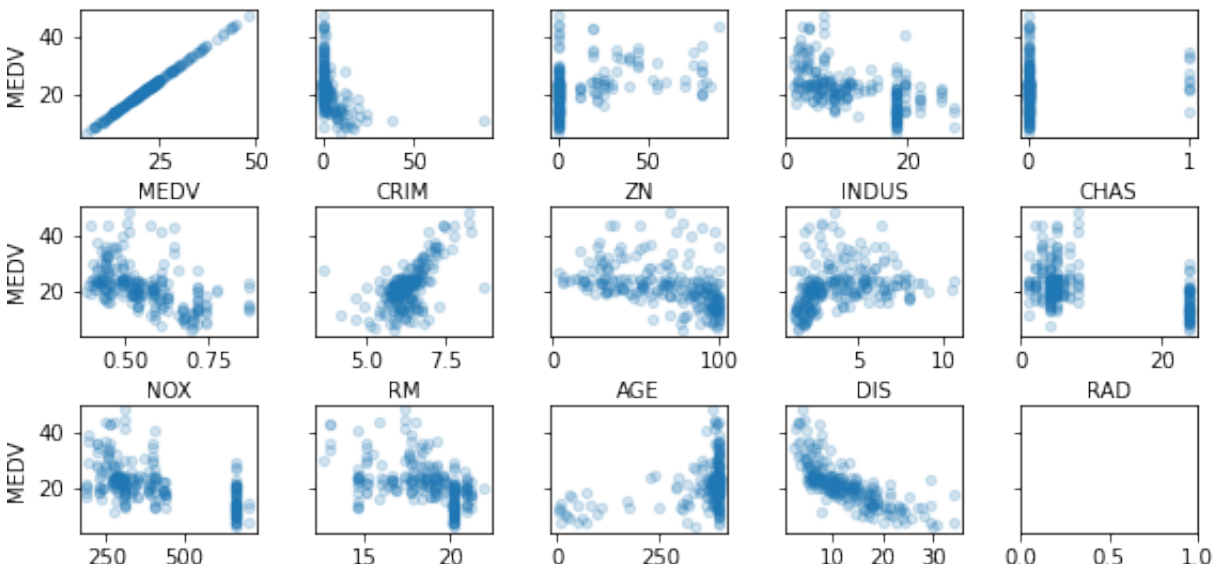
	AGE	CRIM	INDUS	LSTAT	NOX	PTRATIO	RAD	TAX	B	DIS	MEDV	RM	ZN	CHAS
AGE	1.00	0.33	0.63	0.61	0.74	0.24	0.44	0.50	-0.26	-0.72	-0.46	-0.21	-0.54	0.03
CRIM	0.33	1.00	0.39	0.42	0.43	0.27	0.62	0.57	-0.32	-0.38	-0.41	-0.10	-0.20	-0.06
INDUS	0.63	0.39	1.00	0.64	0.76	0.35	0.57	0.67	-0.34	-0.71	-0.55	-0.36	-0.51	0.01
LSTAT	0.61	0.42	0.64	1.00	0.62	0.33	0.53	0.58	-0.41	-0.53	-0.75	-0.54	-0.43	-0.02
NOX	0.74	0.43	0.76	0.62	1.00	0.21	0.63	0.68	-0.38	-0.78	-0.50	-0.26	-0.53	0.02
PTRATIO	0.24	0.27	0.35	0.33	0.21	1.00	0.43	0.40	-0.17	-0.23	-0.49	-0.25	-0.34	-0.13
RAD	0.44	0.62	0.57	0.53	0.63	0.43	1.00	0.90	-0.46	-0.48	-0.48	-0.17	-0.30	-0.06
TAX	0.50	0.57	0.67	0.58	0.68	0.40	0.90	1.00	-0.45	-0.51	-0.57	-0.26	-0.31	-0.09

B	-0.26	-0.32	-0.34	-0.41	-0.38	-0.17	-0.46	-0.45	1.00	0.29	0.38	0.08	0.18	0.06
DIS	-0.72	-0.38	-0.71	-0.53	-0.78	-0.23	-0.48	-0.51	0.29	1.00	0.30	0.16	0.61	-0.06
MEDV	-0.46	-0.41	-0.55	-0.75	-0.50	-0.49	-0.48	-0.57	0.38	0.30	1.00	0.62	0.37	0.10
RM	-0.21	-0.10	-0.36	-0.54	-0.26	-0.25	-0.17	-0.26	0.08	0.16	0.62	1.00	0.27	0.12
ZN	-0.54	-0.20	-0.51	-0.43	-0.53	-0.34	-0.30	-0.31	0.18	0.61	0.37	0.27	1.00	-0.01
CHAS	0.03	-0.06	0.01	-0.02	0.02	-0.13	-0.06	-0.09	0.06	-0.06	0.10	0.12	-0.01	1.00

	AGE	**DIS**	**INDUS**	**NOX**	**ZN**	**CRIM**	**LSTAT**	**MEDV**	**RAD**	**TAX**	**B**	**PTRATIO**	**RM**	**CHAS**
AGE	1.00	0.72	0.63	0.74	0.54	0.33	0.61	0.46	0.44	0.50	0.26	0.24	0.21	0.03
DIS	0.72	1.00	0.71	0.78	0.61	0.38	0.53	0.30	0.48	0.51	0.29	0.23	0.16	0.06
INDUS	0.63	0.71	1.00	0.76	0.51	0.39	0.64	0.55	0.57	0.67	0.34	0.35	0.36	0.01
NOX	0.74	0.78	0.76	1.00	0.53	0.43	0.62	0.50	0.63	0.68	0.38	0.21	0.26	0.02
ZN	0.54	0.61	0.51	0.53	1.00	0.20	0.43	0.37	0.30	0.31	0.18	0.34	0.27	0.01
CRIM	0.33	0.38	0.39	0.43	0.20	1.00	0.42	0.41	0.62	0.57	0.32	0.27	0.10	0.06
LSTAT	0.61	0.53	0.64	0.62	0.43	0.42	1.00	0.75	0.53	0.58	0.41	0.33	0.54	0.02
MEDV	0.46	0.30	0.55	0.50	0.37	0.41	0.75	1.00	0.48	0.57	0.38	0.49	0.62	0.10
RAD	0.44	0.48	0.57	0.63	0.30	0.62	0.53	0.48	1.00	0.90	0.46	0.43	0.17	0.06
TAX	0.50	0.51	0.67	0.68	0.31	0.57	0.58	0.57	0.90	1.00	0.45	0.40	0.26	0.09
B	0.26	0.29	0.34	0.38	0.18	0.32	0.41	0.38	0.46	0.45	1.00	0.17	0.08	0.06
PTRATIO	0.24	0.23	0.35	0.21	0.34	0.27	0.33	0.49	0.43	0.40	0.17	1.00	0.25	0.13
RM	0.21	0.16	0.36	0.26	0.27	0.10	0.54	0.62	0.17	0.26	0.08	0.25	1.00	0.12
CHAS	0.03	0.06	0.01	0.02	0.01	0.06	0.02	0.10	0.06	0.09	0.06	0.13	0.12	1.00

- MEDV와 양의 상관관계를 갖는 데이터와 음의 상관관계를 갖는 데이터가 구분됩니다.
- MEDV와 음의 상관관계를 가지는 feature가 8개, 양의 상관관계를 가지는 feature가 4개, 상관관계가 없는 feature가 1개 있습니다.

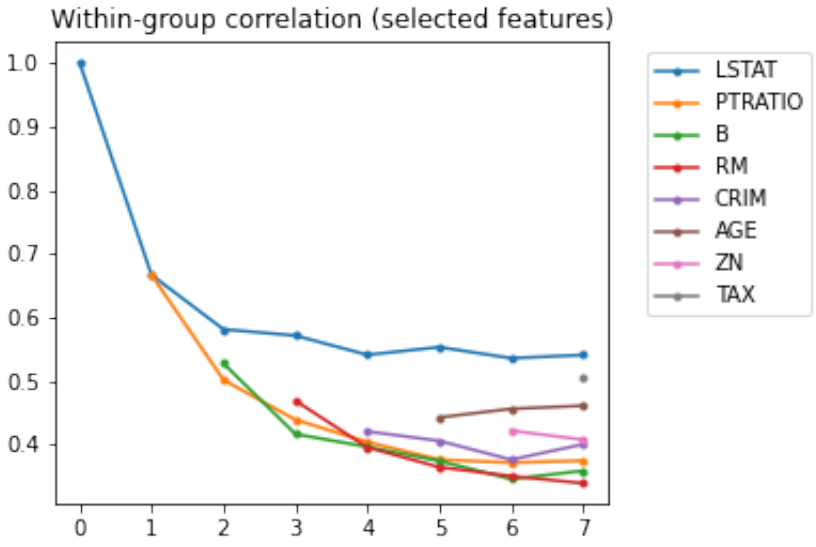
- 유사한 데이터셋이 관찰됩니다. correlation이 0.7 이상인 (AGE,DIS,NOX,INDUS), (LSTAT,MEDV), (RAD,TAX) 는 거의 동일한 데이터입니다.
- 더미 데이터가 관찰됩니다. MEDV와의 상관관계가 0.2 이하인 CHAS는 더미 데이터입니다.



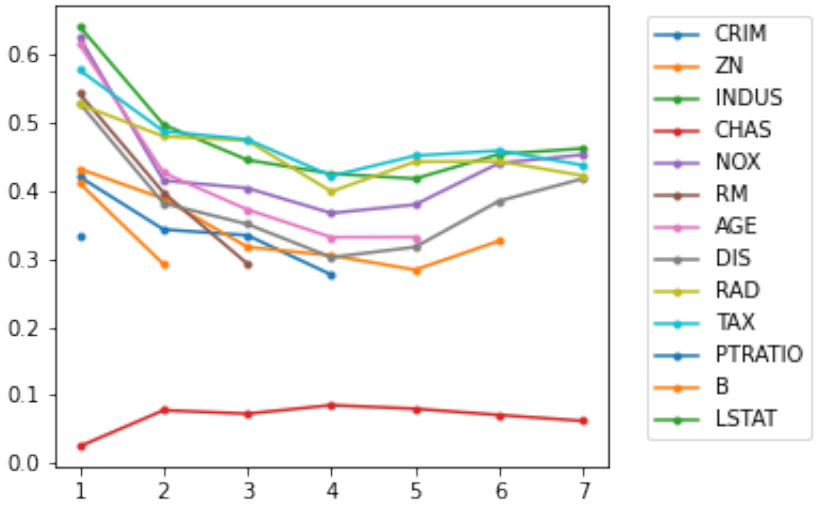
TAX PTRATIO B LSTAT

- RM, LSTAT이 MDEV과 강한 상관관계가 있음을 확인할 수 있습니다.
- DIS의 경우 MDEV<20 구간에서만 잘 예측하는 특징이 있습니다.

Feature 선택



Between-group correlation (selected features, rejected features)



- selected_features: ['LSTAT', 'PTRATIO', 'B', 'RM', 'CRIM', 'AGE', 'ZN', 'TAX']
- rejected_features : ['INDUS', 'CHAS', 'NOX', 'DIS', 'RAD']

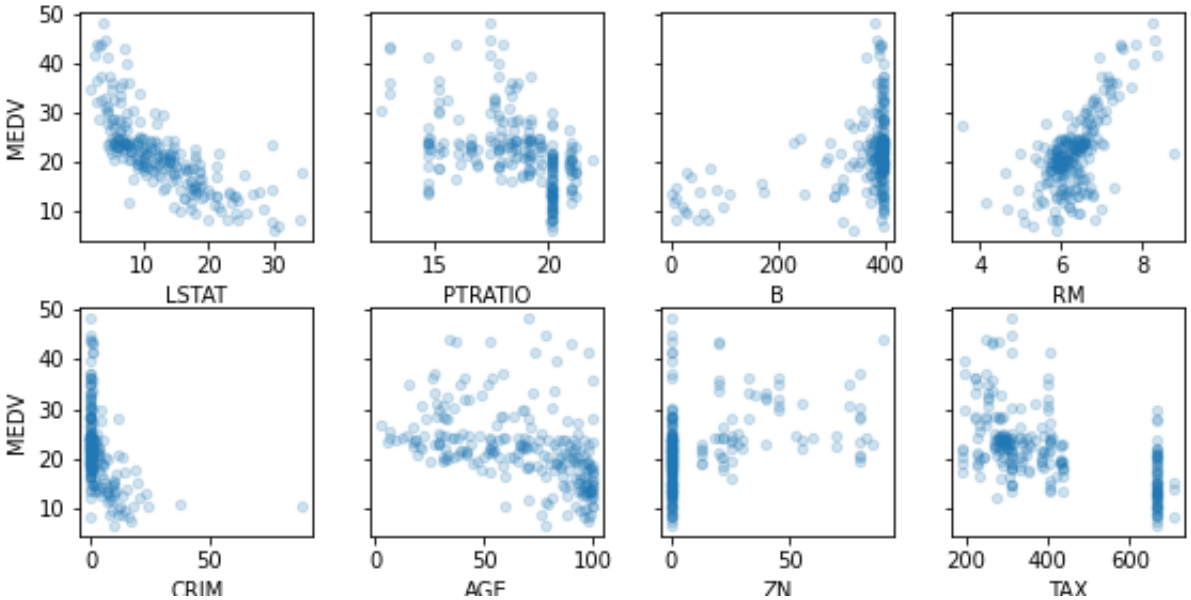
- Within-group correlation : 선택한 feature set의 intrinsic dimension이 증가함을 의미합니

다. (Intrinsic dimension : F Camastra 2002)

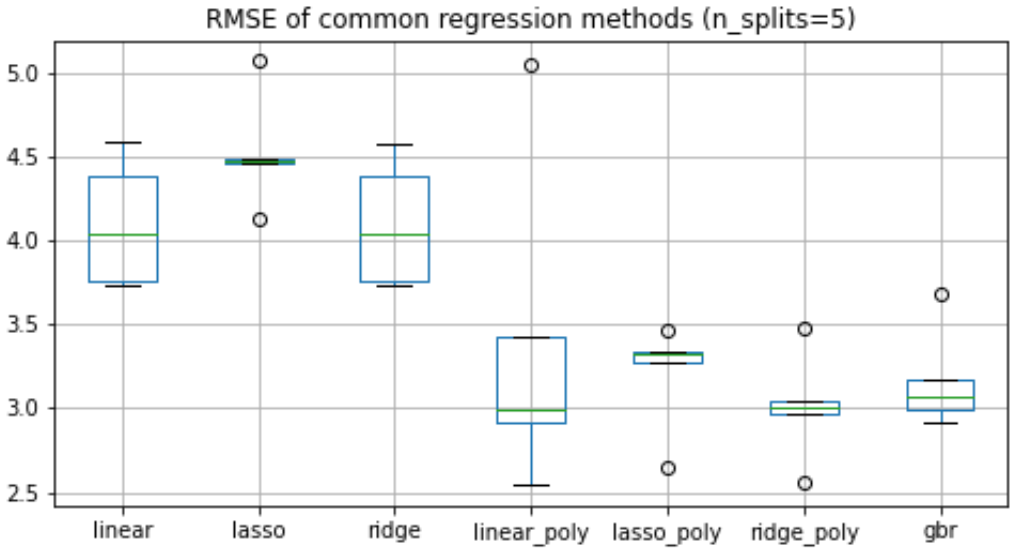
- Between-group correlation : 데이터를 5개 선택하면 나머지 feature를 예측하기 시작함을 의미합니다.

- Correlation을 기준으로 예측력이 높으면서 중복되지 않는 데이터를 greedy하게 선택하여, dummy가 나올때까지 반복했습니다.
- 위의 차트로부터 적절한 feature의 개수는 5개~8개 정도로 예상됩니다. 8개 사용하기로 하였습니다.

- 선택된 feature에 대한 Scatter Plot



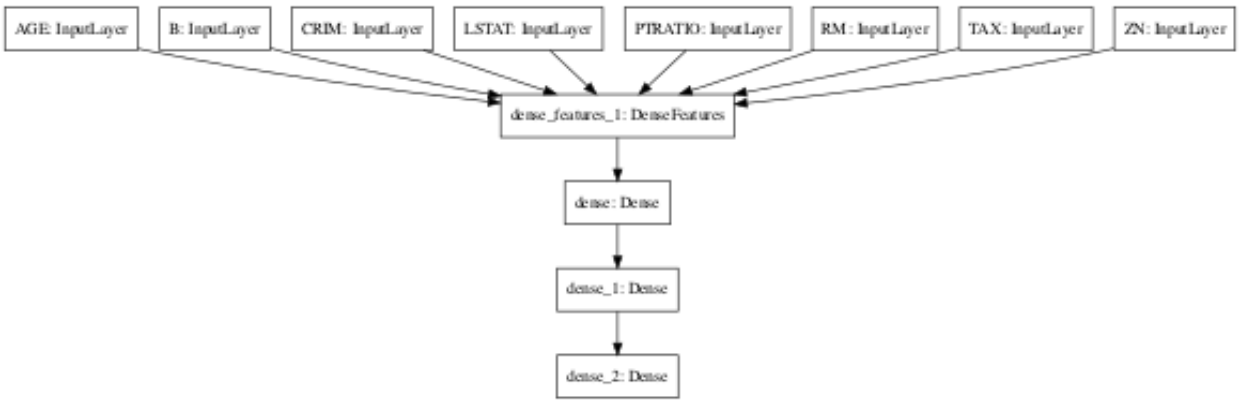
- 일반적인 regression method 적용

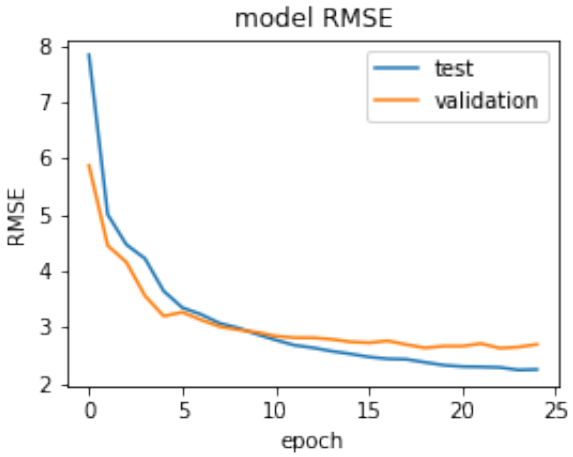


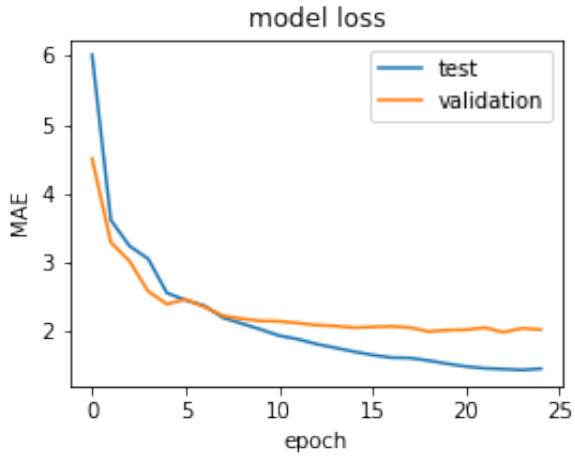
- linear, polynomial model에 대하여 여러 가지 regression method를 적용하였습니다.
- 전반적으로 Polynomial model이 Linear model에 비해 더 좋은 성능을 보입니다.
- ensemble method인 gbr을 적용해도 큰 차이가 없는 것으로 보아, feature의 nonlinearity가 주로 2차 정도라고 판단됩니다

Deep learning

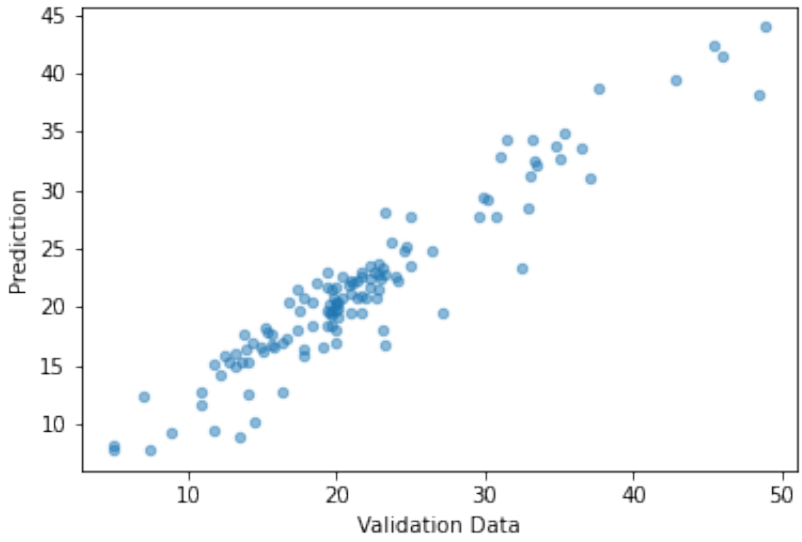
- tensorflow.org/tutorials을 참고하여 128 크기의 hidden layer 2개를 이용한 deep learning 모델을 다음과 같이 작성 후 train 하였습니다.







- Outlier가 있을 수 있다고 생각해 MAE를 loss function으로 잡고 25 epoch training 하였습니다.
- Second order polinomial fitting과 유사한 결과를 얻었습니다.
- 더 많은 데이터를 얻거나, feature간의 관계를 좀더 잘 파악하면 결과를 개선할 수 있을 것입니다.

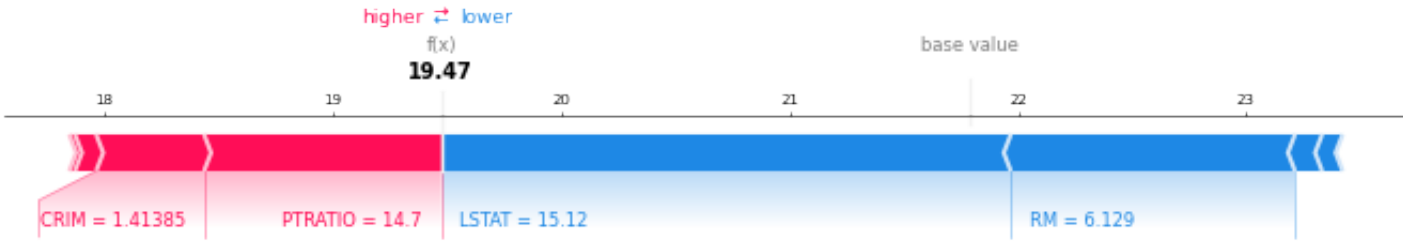


- Validation set을 이용한 model의 prediction이 실제 값과 잘 일치하는 것이 확인됩니다.

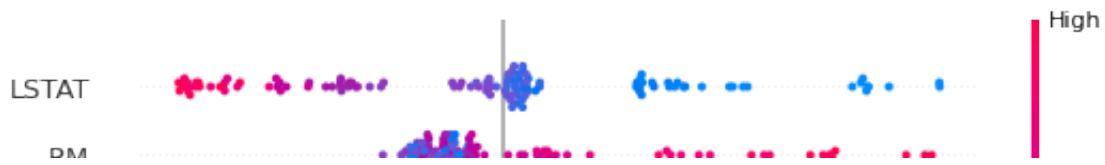
모델 설명

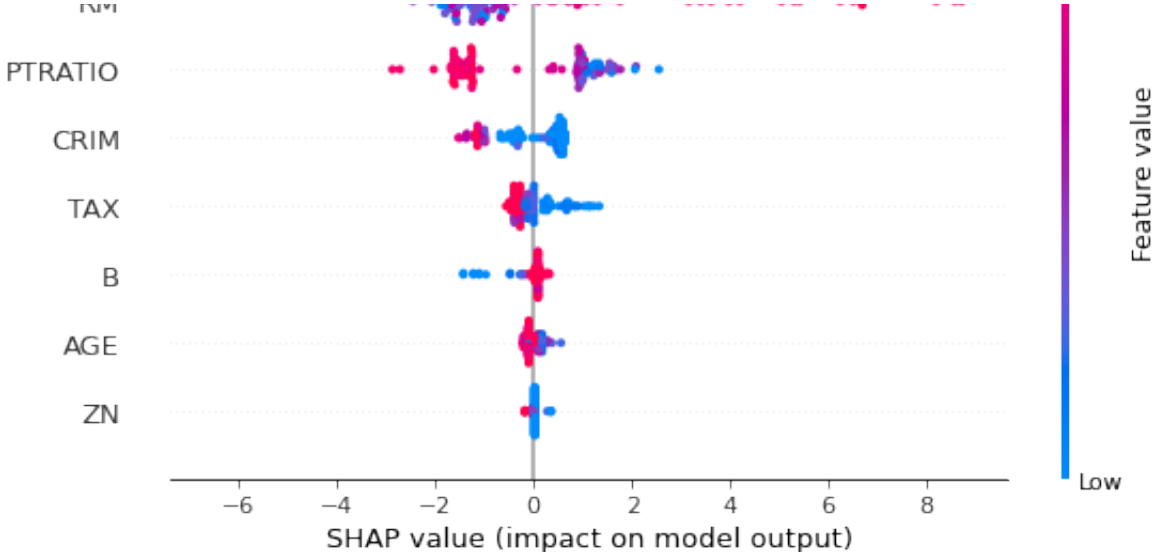
- Feature 중요도 파악 등을 위해, GBR 모델의 피팅 결과를 feature별로 관찰하겠습니다.

- 임의의(154번째) 데이터의 SHAP (각 feature가 준 영향, model이 보는 feature의 중요도.)



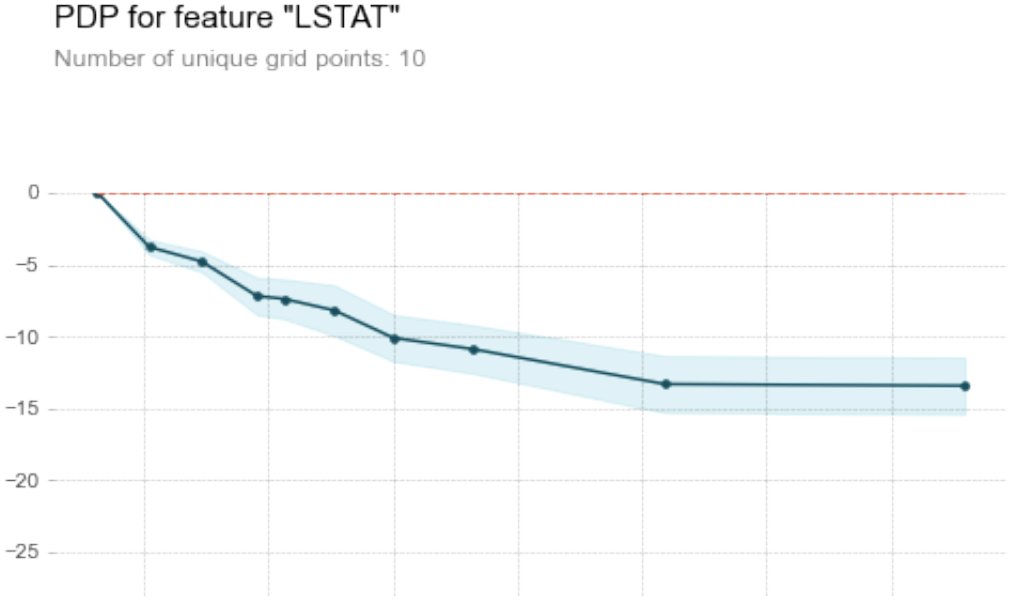
- SHAP summary plot. model은 일부 데이터에 대해 LSTAT에 많이 의존하여 가격을 결정하였습니다.

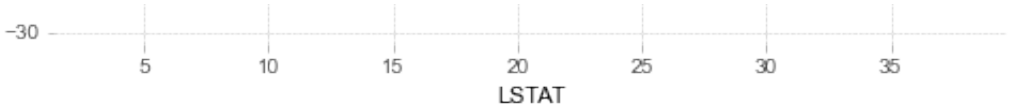




- Permutation plot과 Partial plot

Weight	Feature
0.5131 ± 0.1792	LSTAT
0.2295 ± 0.0366	RM
0.0711 ± 0.0196	PTRATIO
0.0160 ± 0.0106	CRIM
0.0155 ± 0.0118	TAX
0.0056 ± 0.0084	B
0.0036 ± 0.0045	AGE
0.0005 ± 0.0013	ZN





- LSTAT은 일관된 partial plot을 보여주면서 높은 permutation을 보여주므로, overfitting이 아니라 실제로 중요한 feature임을 알 수 있습니다.