보스톤 집값 예측 모델

A반 강연주

추진 배경 및 목적

치솟는 집값, 그 이유는 무엇일까
– 다른 시간, 다른 공간에서 이유를 찾다.

- 2020년, 서울에 위치한 집값의 급격한 상승으로 경제적, 정치적 불안정 야기

- 국내 집 값 변화의 요인을 확인하기 위해, 서울과 같이 높은 집값을 보인 1970년대 Boston의 집값과 영향 요인 탐색

- 탐색적 분석 ➡ 변수의 상관성 파악

  예측 모델 ➡ 정확한 집값을 예측

목표 변수와 설명 변수

- 목표 변수 : 주택 가격 (MEDV)

- 설명변수
  - 범죄율 (CRIM)
  - 주거지 비율 (ZN)
  - 비소매업 비율 (INDUS)
  - 강 조망 여부 (CHAS)
  - 산화질소 농도 (NOX)
  - 주거당 평균 객실 수 (RM)
  - 노후 건물 비율 (AGE)
  - 중심지(노동센터) 접근 거리 (DIS)
  - 고속도로 접근 편이성 지수 (RAD)
  - 재산세율 (TAX)
  - 학생당 교사 비율 (PTRATIO)
  - 흑인 인구 비율 (B)
  - 저소득층 비율 (LSTAT)

가설설정

- 환경적 요건

- 강 조망이 있으면 주택 가격이 높을 것이다.
- 산화 질소 농도가 높으면 주택 가격이 낮을 것이다.

- 접근성 요건

- 주거지 비율이 높을수록 집값은 높을 것이다.
- 중심지 (직업 센터) 접근 거리가 가까울수록 주택 가격이 높을 것이다.
- 방사형 도로 접근성 지수가 높을수록 주택 가격이 높을 것이다.

- 이웃 요건

- 학생/교사 비율이 낮을수록 주택 가격이 높을 것이다.
- 흑인 인구 비율이 높을수록 주택 가격이 낮을 것이다.
- 저소득층 비율이 높을수록 주택 가격이 낮을 것이다.

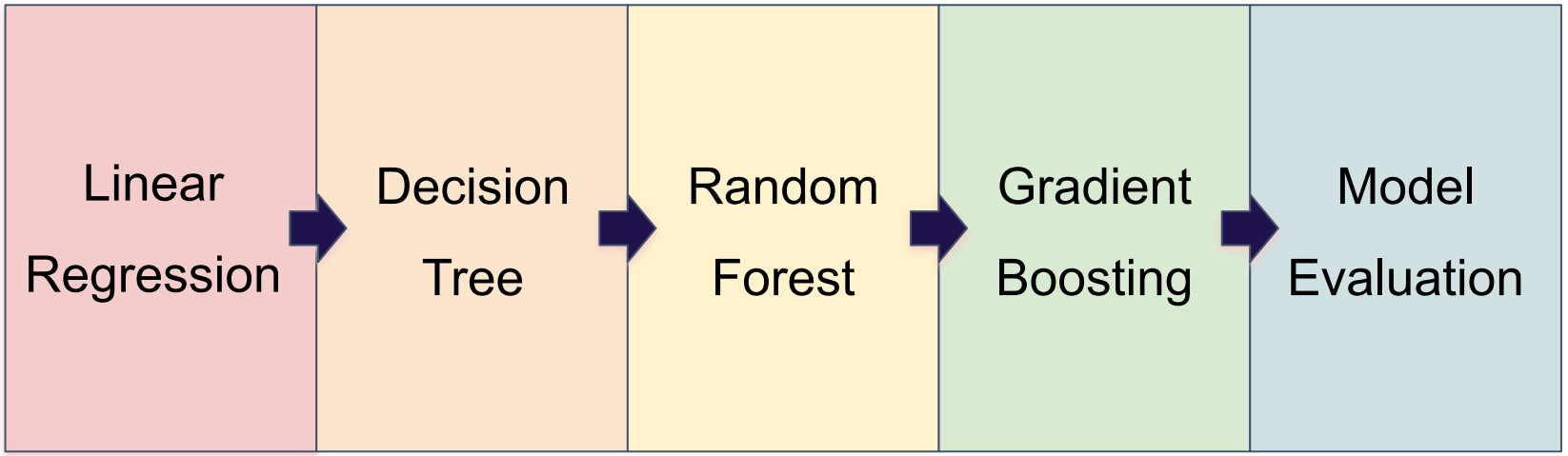
- 사회 경제적 요건

- 1인당 범죄율이 높을수록 주택 가격이 낮을 것이다.
- 자기 소유 집 비율이 높을 수록 주택 가격이 높아질 것이다.
- 재산 세율이 낮으면 주택 가격이 낮을 것이다.

데이터 분석 방법

- 탐색적 분석을 통한 목표 변수와 설명 변수와의 관계성을 밝혀 가설 검증

- 예측 모델을 활용한 보스턴 집값 예측



데이터 전처리

- 결측치 처리

```
# 결측 처리
df_raw.isnull().sum(axis=0)
```

```
MEDV       0
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
B          0
LSTAT      0
dtype: int64
```

- 타입 변경

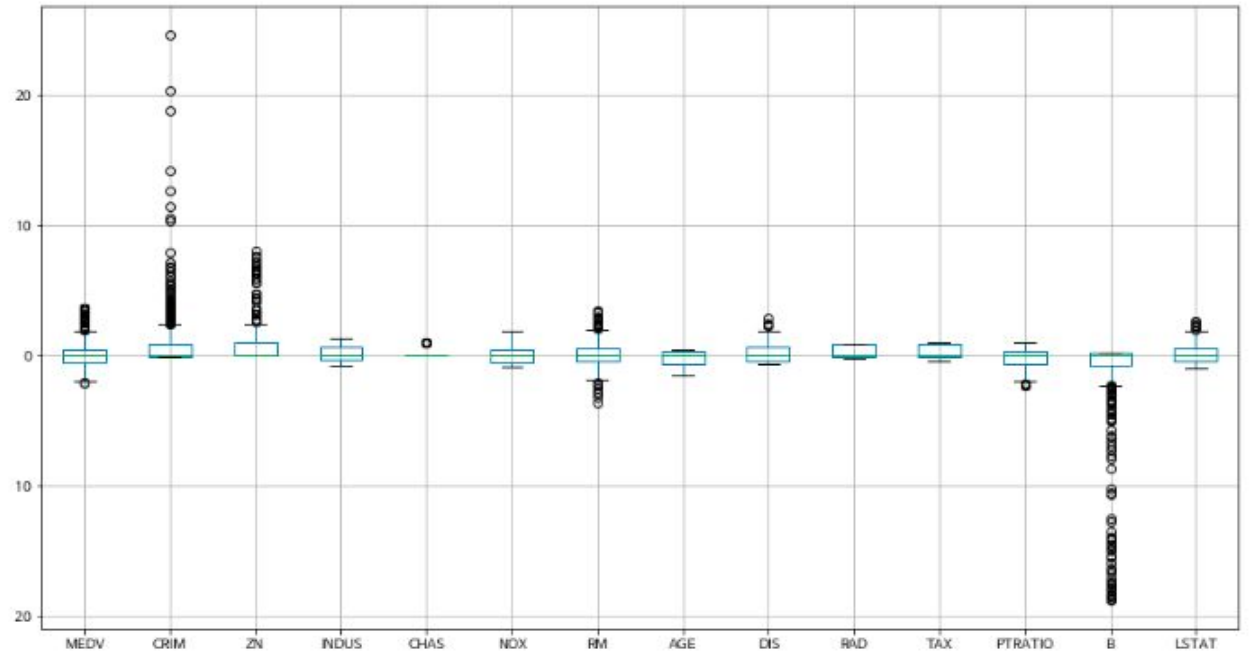
```
# 변수별 타입 분석
df_raw.dtypes
```

```
MEDV       float64
CRIM       float64
ZN         float64
INDUS      float64
CHAS         int64
NOX        float64
RM         float64
AGE        float64
DIS        float64
RAD          int64
TAX          int64
PTRATIO    float64
B          float64
LSTAT      float64
dtype: object
```

```
# CHAS 타입 변경
df_raw=df_raw.astype({'CHAS':object})
```

- 이상치 제거

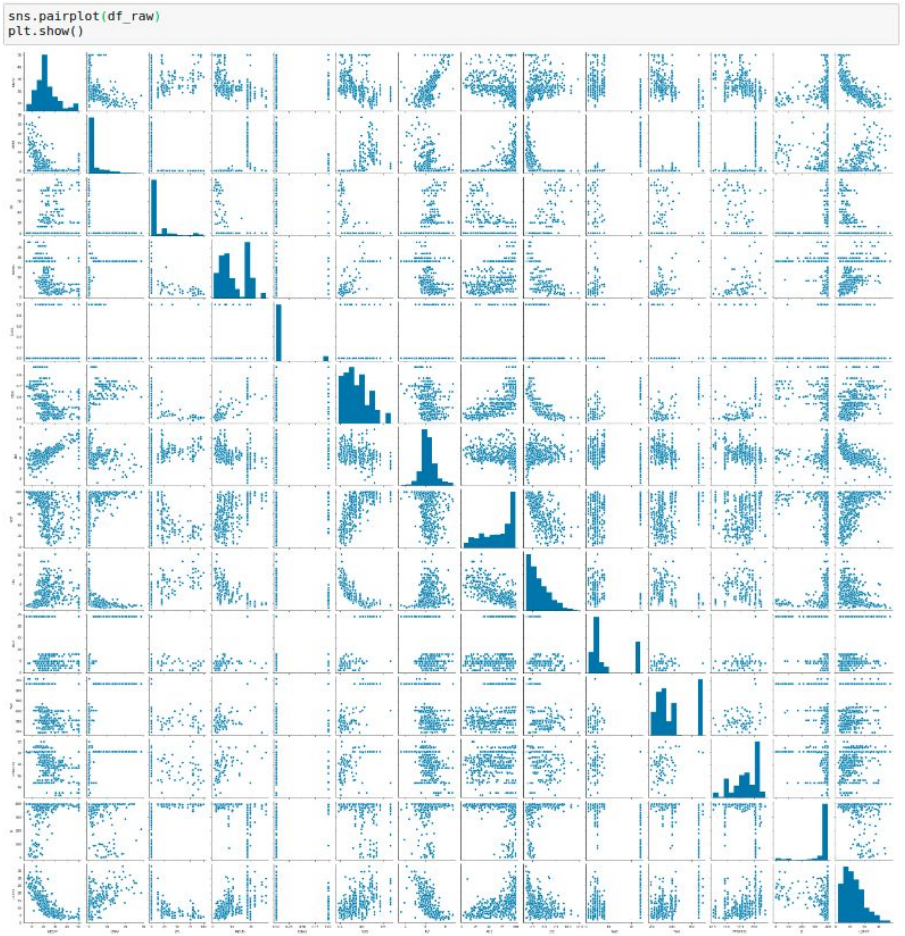
```
# 표준화하여 Boxplot을 그려 이상치 제거
df_raw_1=robust_scale(df_raw_numeric)
df_raw_1=pd.DataFrame(df_raw_1,columns = df_raw.columns)
df_raw_1.boxplot(figsize=(15,8))
```




```
# CRIM 이상치 제거
df_raw=df_raw.drop([380,418,405,410,414,404,398,427])
```

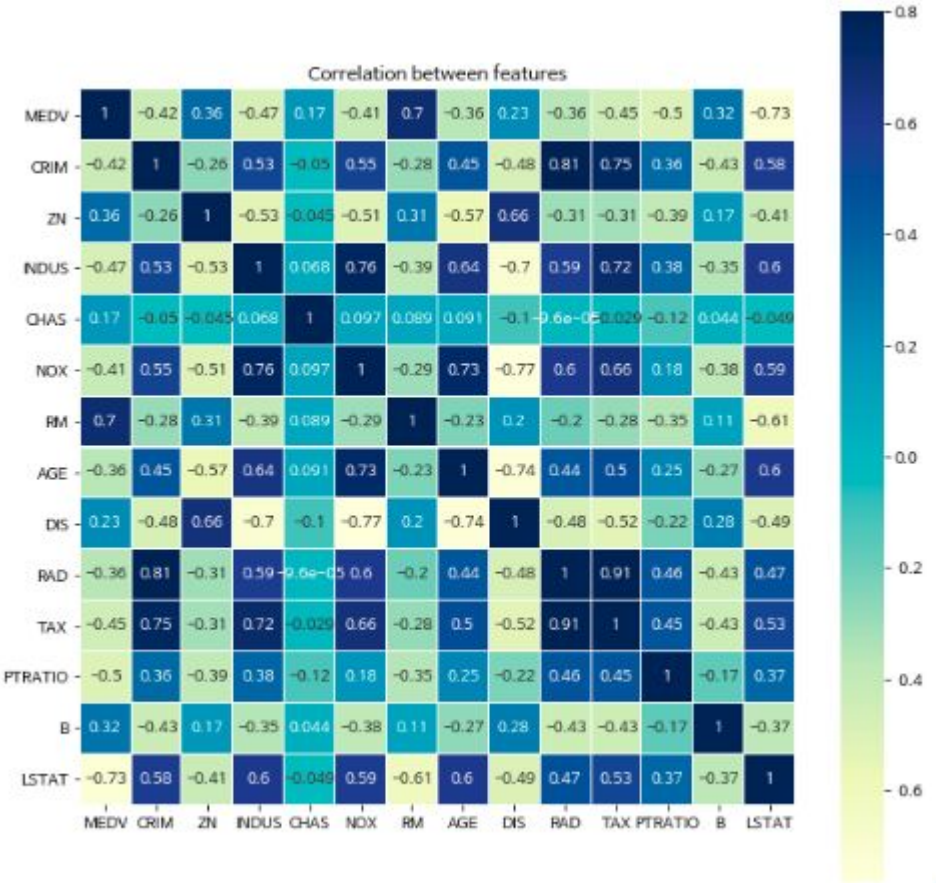
탐색적 분석을 통한 변수간의 상관 관계 파악

- Scatter plot



변수들의 상관 분포

- Hit Map

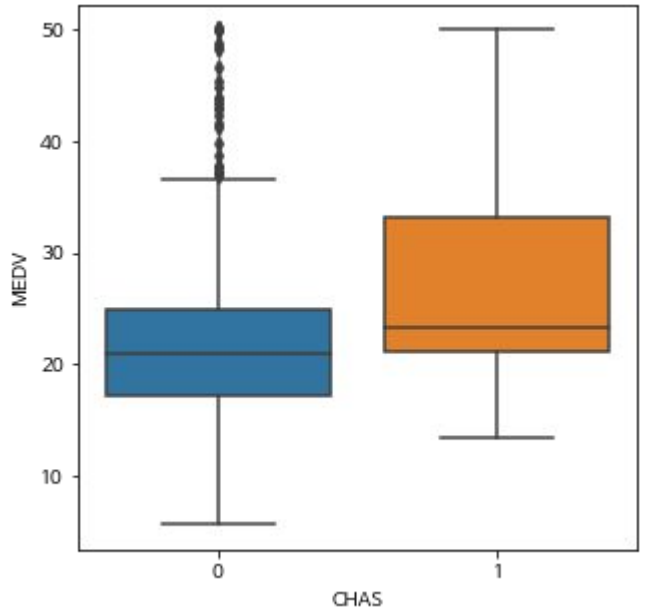


변수들의 상관 계수

탐색적 분석을 통한 가설 검증

- 환경적 요건

  - 강 조망이 있으면 주택 가격이 높을 것이다.



0(비조망)에 비해 1(조망)이 높은 집값 분포를 보임

  - 산화 질소가 높으면 주택 가격이 낮을 것이다.

OLS Regression Results

Dep. Variable:	NOX	R-squared:	0.170
Model:	OLS	Adj. R-squared:	0.169
Method:	Least Squares	F-statistic:	101.9
Date:	Wed, 25 Nov 2020	Prob (F-statistic):	6.48e-22
Time:	03:45:22	Log-Likelihood:	414.33
No. Observations:	498	AIC:	-824.7
Df Residuals:	496	BIC:	-816.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6723	0.013	52.730	0.000	0.647	0.697
MEDV	-0.0053	0.001	-10.094	0.000	-0.006	-0.004

Omnibus:	46.408	Durbin-Watson:	0.215
Prob(Omnibus):	0.000	Jarque-Bera (JB):	57.048
Skew:	0.817	Prob(JB):	4.09e-13
Kurtosis:	3.284	Cond. No.	66.2

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

회귀식 : [MEDV]=-0.01*[NOX]+0.67
설명력 : 17%

Prob (F-statistic) < 0.05 , P>|t| < 0.05이므로 유의함

⇨ 산화 질소 농도와 주택 가격은 매우 약한 양의 상관관계를 보임

탐색적 분석을 통한 가설 검증

- 접근성 요건

  - 주거지 비율이 높을수록 집값이 높을 것이다.

OLS Regression Results

Dep. Variable:	ZN	R-squared:	0.127
Model:	OLS	Adj. R-squared:	0.125
Method:	Least Squares	F-statistic:	72.07
Date:	Wed, 25 Nov 2020	Prob (F-statistic):	2.43e-16
Time:	03:45:22	Log-Likelihood:	-2243.8
No. Observations:	498	AIC:	4492.
Df Residuals:	496	BIC:	4500.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-9.3620	2.652	-3.530	0.000	-14.573	-4.151
MEDV	0.9189	0.108	8.489	0.000	0.706	1.132

Omnibus:	180.896	Durbin-Watson:	0.471
Prob(Omnibus):	0.000	Jarque-Bera (JB):	495.424
Skew:	1.813	Prob(JB):	2.63e-108
Kurtosis:	6.275	Cond. No.	66.2

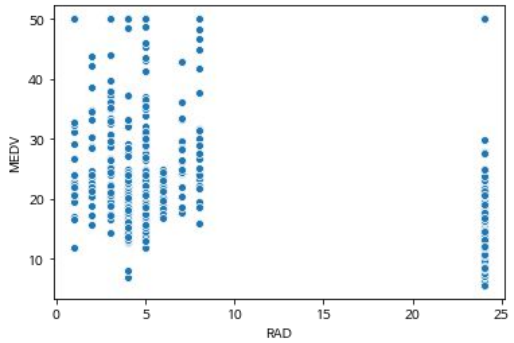
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

회귀식 : [MEDV]=0.91*[ZN]-9.36
설명력 : 12.7%

Prob (F-statistic) < 0.05 , P>|t| < 0.05이므로 유의함

⇨ 주거지의 비율과 주택가격은 강한 양의 상관관계를 보임

  - 방사형 도로 접근성 지수가 높을수록 주택 가격이 높을 것이다.



방사형 도로 접근성 지수와 주택 가격의 상관성이 거의 없음

탐색적 분석을 통한 가설 검증

- 이웃 요건

  - 학생/교사 비율이 낮을수록 주택 가격이 높을 것이다.

OLS Regression Results

Dep. Variable:	PTRATIO	R-squared:	0.258
Model:	OLS	Adj. R-squared:	0.256
Method:	Least Squares	F-statistic:	175.1
Date:	Wed, 25 Nov 2020	Prob (F-statistic):	1.61e-34
Time:	03:58:10	Log-Likelihood:	-1032.9
No. Observations:	506	AIC:	2070.
Df Residuals:	504	BIC:	2078.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	21.1489	0.220	96.216	0.000	20.717	21.581
MEDV	-0.1195	0.009	-13.233	0.000	-0.137	-0.102

Omnibus:	39.733	Durbin-Watson:	0.389
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47.057
Skew:	-0.705	Prob(JB):	6.05e-11
Kurtosis:	3.495	Cond. No.	64.5

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

회귀식 : [MEDV]=-0.12*[PIRAITO]+21.15
설명력 : 25.6%

Prob (F-statistic) < 0.05 , P>|t| < 0.05이므로 유의함

⇨ 교육적 요소를 고려하여 학생/교사 비율 주택가격은 약한 음의 상관관계를 가진다.

  - 저소득층 비율이 높을수록 주택 가격이 낮을 것이다.

OLS Regression Results

Dep. Variable:	LSTAT	R-squared:	0.544
Model:	OLS	Adj. R-squared:	0.543
Method:	Least Squares	F-statistic:	601.6
Date:	Wed, 25 Nov 2020	Prob (F-statistic):	5.08e-88
Time:	04:00:14	Log-Likelihood:	-1513.5
No. Observations:	506	AIC:	3031.
Df Residuals:	504	BIC:	3039.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	25.5589	0.568	44.980	0.000	24.442	26.675
MEDV	-0.5728	0.023	-24.528	0.000	-0.619	-0.527

Omnibus:	87.432	Durbin-Watson:	0.901
Prob(Omnibus):	0.000	Jarque-Bera (JB):	143.457
Skew:	1.059	Prob(JB):	7.06e-32
Kurtosis:	4.524	Cond. No.	64.5

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

회귀식 : [MEDV]=-0.57*[PIRAITO]+25.56
설명력 : 54.4%

Prob (F-statistic) < 0.05 , P>|t| < 0.05이므로 유의함

⇨ 저소득층 비율이 높을수록 주택 가격이 낮다.

탐색적 분석을 통한 가설 검증

- 사회 경제적 요건

  - 1인당 범죄율이 높을수록 주택 가격이 낮을 것이다.

OLS Regression Results

Dep. Variable:	CRIM	R-squared:	0.180
Model:	OLS	Adj. R-squared:	0.178
Method:	Least Squares	F-statistic:	108.6
Date:	Wed, 25 Nov 2020	Prob (F-statistic):	3.95e-23
Time:	03:45:23	Log-Likelihood:	-1467.0
No. Observations:	498	AIC:	2938.
Df Residuals:	496	BIC:	2947.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.1735	0.557	14.663	0.000	7.078	9.269
MEDV	-0.2371	0.023	-10.422	0.000	-0.282	-0.192

Omnibus:	216.887	Durbin-Watson:	0.449
Prob(Omnibus):	0.000	Jarque-Bera (JB):	807.208
Skew:	2.036	Prob(JB):	5.21e-176
Kurtosis:	7.724	Cond. No.	66.2

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

회귀식 : [MEDV]=-0.24*[CRIM]+8.17
설명력 : 18%

Prob (F-statistic) < 0.05 , P>|t| < 0.05이므로 유의함

⇨ 범죄율이 높은 지역일수록 주택 가격이 낮다.

  - 주거당 평균 객실 수가 높으면 주택 가격이 높을 것이다.

OLS Regression Results

Dep. Variable:	RM	R-squared:	0.484
Model:	OLS	Adj. R-squared:	0.483
Method:	Least Squares	F-statistic:	471.8
Date:	Wed, 25 Nov 2020	Prob (F-statistic):	2.49e-74
Time:	04:03:00	Log-Likelihood:	-371.73
No. Observations:	506	AIC:	747.5
Df Residuals:	504	BIC:	755.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.0876	0.060	85.492	0.000	4.971	5.205
MEDV	0.0531	0.002	21.722	0.000	0.048	0.058

Omnibus:	123.606	Durbin-Watson:	1.160
Prob(Omnibus):	0.000	Jarque-Bera (JB):	931.463
Skew:	-0.840	Prob(JB):	5.44e-203
Kurtosis:	9.431	Cond. No.	64.5

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

회귀식 : [MEDV]=0.05*[RM]+5.09
설명력 : 48.4%

Prob (F-statistic) < 0.05 , P>|t| < 0.05이므로 유의함

⇨ 주거당 평균 객실 수와 주택 가격은 약한 양의 상관관계를 갖는다.

Linear Regression (다중 회귀 분석)을 통한 예측 모델링

- 최종 모델링

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:	Wed, 25 Nov 2020	Prob (F-statistic):	6.72e-135
Time:	03:46:22	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
C(CHAS)[T.1]	2.6867	0.862	3.118	0.002	0.994	4.380
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
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

Omnibus:	178.041	Durbin-Watson:	1.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783.126
Skew:	1.521	Prob(JB):	8.84e-171
Kurtosis:	8.281	Cond. No.	1.51e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.51e+04. This might indicate that there are strong multicollinearity or other numerical problems.

*범죄율, 비소매업 비율, 노후 건물 비율 에서 **P> 0.05***

➪ 모델의 수정 필요

- 최종 모델링
(후진제거법으로 중요변수 추출)

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.708
Model:	OLS	Adj. R-squared:	0.705
Method:	Least Squares	F-statistic:	242.6
Date:	Wed, 25 Nov 2020	Prob (F-statistic):	3.67e-131
Time:	03:47:41	Log-Likelihood:	-1528.7
No. Observations:	506	AIC:	3069.
Df Residuals:	500	BIC:	3095.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	37.4992	4.613	8.129	0.000	28.436	46.562
LSTAT	-0.5811	0.048	-12.122	0.000	-0.675	-0.487
NOX	-17.9966	3.261	-5.519	0.000	-24.403	-11.590
RM	4.1633	0.412	10.104	0.000	3.354	4.973
DIS	-1.1847	0.168	-7.034	0.000	-1.516	-0.854
PTRATIO	-1.0458	0.114	-9.212	0.000	-1.269	-0.823

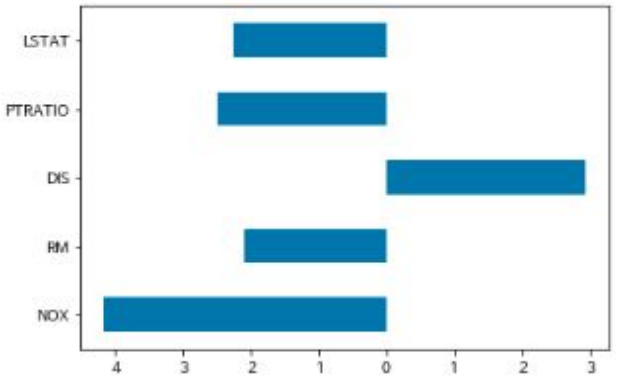
Omnibus:	187.456	Durbin-Watson:	0.971
Prob(Omnibus):	0.000	Jarque-Bera (JB):	885.498
Skew:	1.584	Prob(JB):	5.21e-193
Kurtosis:	8.654	Cond. No.	545.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

*모든 변수에서 **p-value < 0.05***

➪ 적합 모델

- 표준화 회귀 계수
(변수의 중요도 파악)

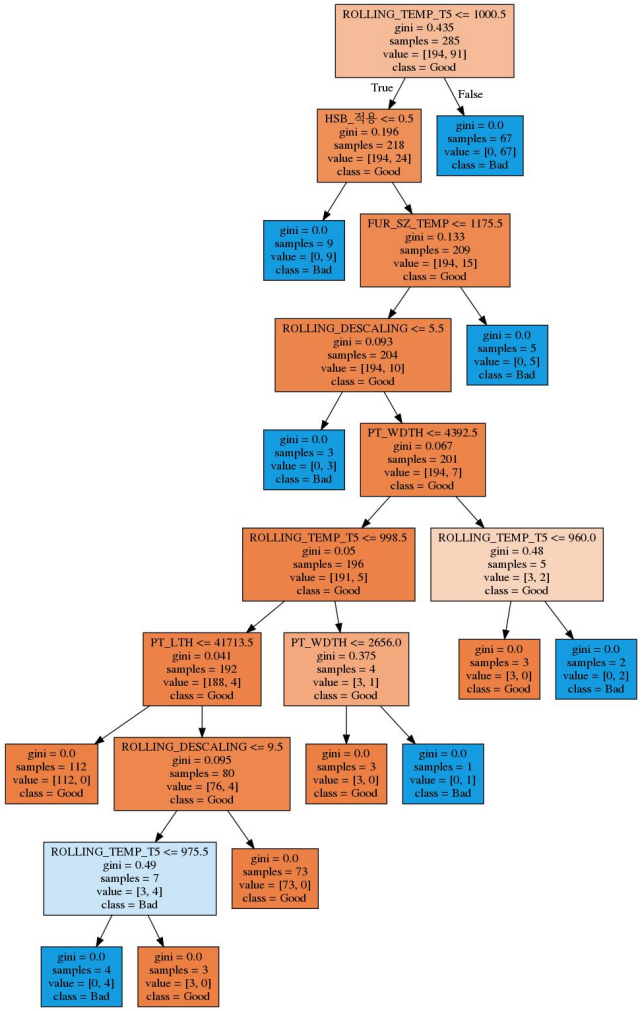


변수 중요도

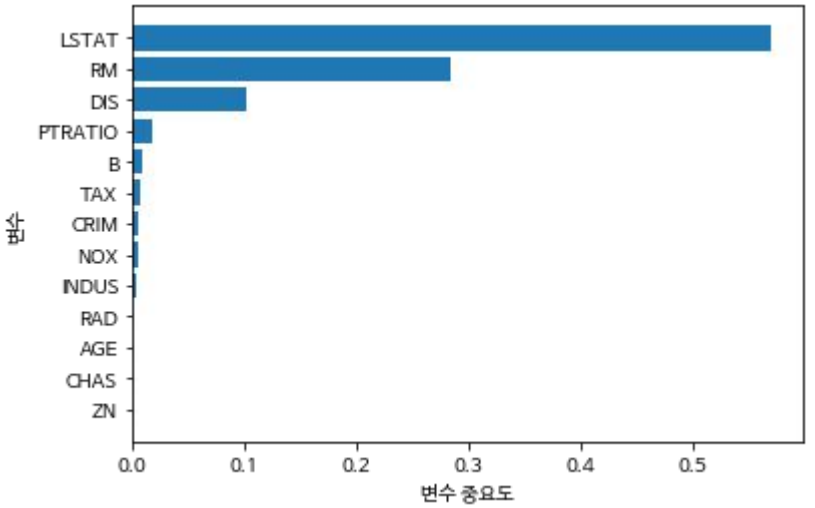
산화질소 농도 > 중심지 (노동센터) 접근거리 > 학생당 교사 비율 > 저소득층 비율 > 주거당 평균 객실 수

Decision Tree (의사 결정 나무)를 통한 예측 모델링

- 최종 모델



- 설명 변수 중요도

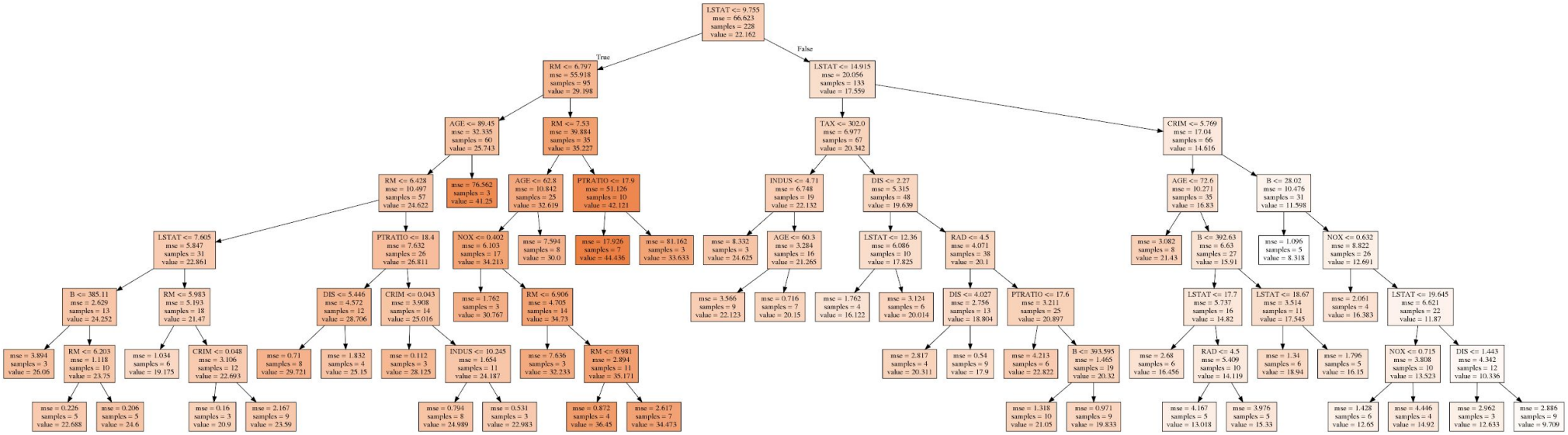


⏫ *저소득층 비율 > 주거당 평균 객실 수 > 중심지 (노동센터) 접근 거리 > 학생당 교사 비율*

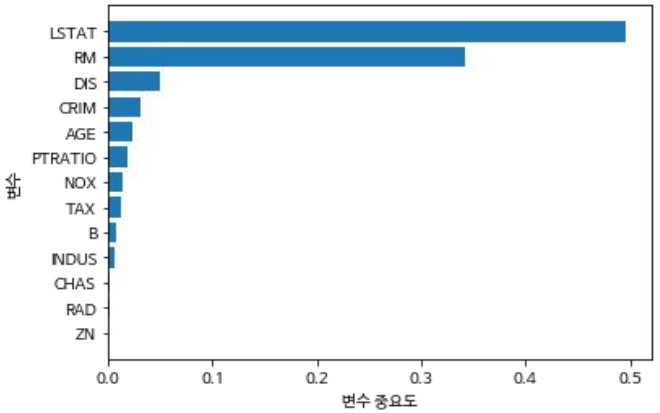
⏪ 최종 모델 결과
train set의 정확도 : 91.2%
test set의 정확도 " 85.5%
Train set의 정확도와 모델의 적합 개선 필요

Random Forest 를 통한 예측 모델링

- 최종 모델



- 설명 변수 중요도



⏫ 최종 모델 결과

train set의 정확도 : 89.2%

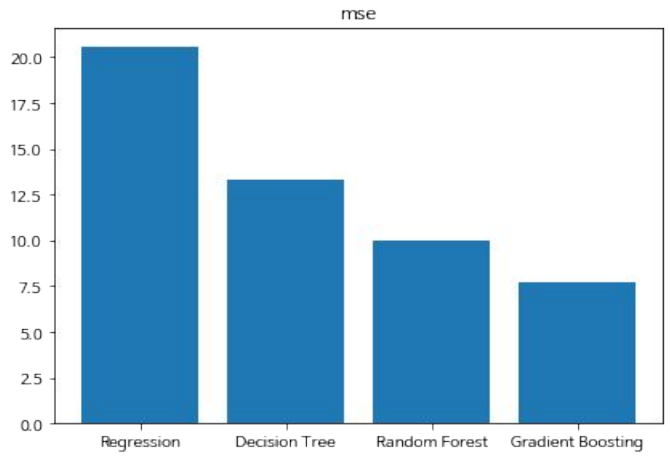
test set의 정확도 : 89.2%

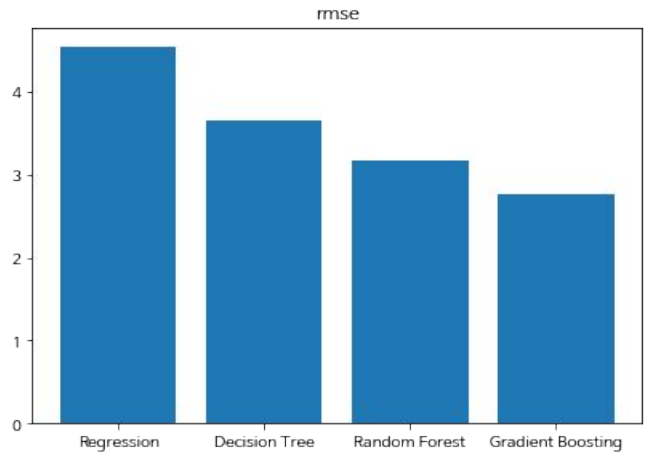
Train/test 정확도를 고려하였을 때, 정확도는 낮지만 적합한 모델

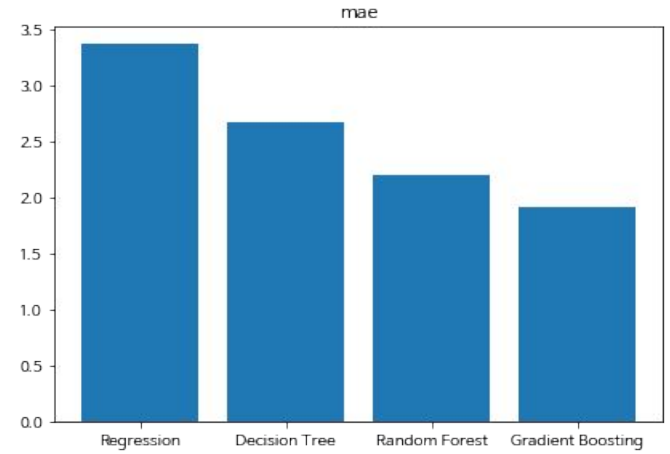
⏪ 저소득층 비율 > 주거당 평균 객실수 > 중심지(노동센터) 접근 거리 > 범죄율 > 노후 건물 비율

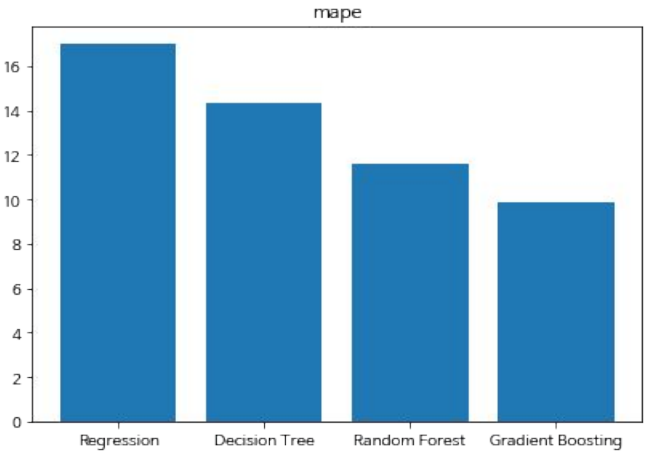
모델 예측 및 평가

Linear Regression, DTR, RFR, GBR 모델 비교









4종류의 오차를 평가하였을 때, Gradient Boosting에서 오차가 가장 작다.

⇨ Gradient Boosting 모델의 예측 정확도가 가장 높다

집값에 영향을 미치는 요인들과 집값의 예측

- 강 조망권, 주거 밀집율, 주거당 평균 객실 수가 높을 수록 주택 가격은 높아진다.
- 반면, 산화 질소의 농도, 저소득층의 비율, 1인당 범죄율, 중심지로 부터의 접근 거리가 클수록 주택 가격은 낮아진다.
- Gradient Boosting 모델을 통한 예측이 가장 높은 정확도를 보인다.

➡ 과거 집 값에 영향을 미친 요인은 현재에도 유효한 것을 확인

➡ 서울과 유사한 해외 지역의 집값 예측 모델을 활용하여 국내 집값 예측 모델 정확한 구상에 기여 가능

실습 과정을 통해 배운 또는 느낀 통찰, 아이디어, 애로사항 등을 정리합니다

- 탐색적 방법을 활용하여 목표 변수와 설명 변수의 상관 관계를 파악하며 가설을 검정하는 것이 흥미로움
- 국내에서 발생한 문제를 해외 사례를 적용하여 해결 방안을 제시할 수 있음을 암시
- 제약적 프로그래밍 활용 역량으로 다양한 방법으로 데이터 분석을 하지 못한 것이 아쉬움

핵심인자 정리 템플릿(참조용)

핵심인자 선정을 위한 분석 과정에서 나온 결과를 순위 등으로 종합 정리합니다.
각자 필요한 형식으로 변경해서 사용하세요(엑셀 파일 제공)

변수	변수 설명	변수 역할	변수 형태	분석 제외 사유	탐색적 기법			모델링 기법							총점	선정 (순위, 사유)
					그래프	검정	상관분석	회귀분석	DT	RF	GB	...	KNN	사례연구		
MEDV	주택가격(중앙값)	목표변수	연속형													
CRIM	범죄율	설명변수	연속형						7	4	4				15	4
ZN	주거지 비율	설명변수	연속형						13	13	13				39	13
INDUS	비소매업 비율	설명변수	연속형						9	10	10				29	10
CHAS	강 조망 여부(1-조망,0-비조망)	설명변수	이산형						12	11	11				34	11
NOX	산화질소 농도	설명변수	연속형					1	8	7	7				22	7
RM	주거당 평균 객실 수	설명변수	연속형					5	2	2	2				6	2
AGE	노후 건물 비율	설명변수	연속형						11	5	5				21	6
DIS	중심지(노동센터) 접근 거리	설명변수	연속형					2	3	3	3				9	3
RAD	고속도로 접근 편이성 지수	설명변수	연속형	상관성이 낮음					10	12	12				34	11
TAX	재산세율	설명변수	연속형						6	8	8				22	7
PTRATIO	학생당 교사 비율	설명변수	연속형		-			3	4	6	6				16	5
B	흑인 인구 비율	설명변수	연속형	상관성이 낮음					5	9	9				23	9
LSTAT	저소득층 비율	설명변수	연속형					4	1	1	1				3	1