

6.1 보스턴 주택가격 예측 BonstonHousing.csv 파일은 미국 인구통계조사국에 의해 수집된 메사추세츠주 보스턴 지역의 주택정보를 담고 있다. 이 데이터 세트는 보스턴 지역의 506개의 인구통계조사 주택단지 정보를 포함한다. 데이터 수집의 목표는 범죄율, 공해, 방의 개수에 대한 정보에 근거하여 새로운 주택단지의 주택가격의 중앙값(median)을 예측하는 것이다. 데이터 세트는 13개의 예측변수와 반응(목표)변수인 주택가격의 중앙값(MEDV)을 포함한다.

d. 예측 변수의 수를 감소하시오:

I. 13개의 예측변수들 중에서 어떠한 예측변수들이 동일한 것을 측정할 것 같은가? INDUS, NOX, TAX 간의 관계에 대하여 논하시오.

1) 13개의 예측변수들 중에서 INDUS와 NOX가 동일한 것을 측정할 것 같다.

2) INDUS, NOX TAX간의 상관관계

```
> cor(BHouse.df[, c("INDUS", "NOX", "TAX")])
```

```
          INDUS      NOX      TAX
INDUS  1.0000000  0.7636514  0.7207602
NOX     0.7636514  1.0000000  0.6680232
TAX     0.7207602  0.6680232  1.0000000
```

=> 3개의 예측변수(INDUS, NOX, TAX) 중에서 가장 높은 상관관계를 가지고 있는 변수는 예상대로 NOX와 INDUS다. 둘 중 하나를 제거해야 한다.

II. 12개의 수치형 예측변수들에 대한 상관관계 표를 계산하고 높은 상관관계를 갖는 변수의 쌍을 찾으시오. 이들은 잠재적으로 중복성을 가지며 다중공선성의 원인이 될 수 있다. 상관관계 표를 보고 어떤 변수들을 제거할지 고르시오.

1) 두 변수의 상관계수의 절대값이 0.75 이상인 변수들을 확인하기 위해 데이터를 변환

```
> over0.75
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV
CRIM	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
ZN	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
INDUS	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
CHAS	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
NOX	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
RM	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
AGE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
DIS	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
RAD	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
TAX	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
PTRATIO	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
LSTAT	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
MEDV	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

상관계수가 0.75 이상 & -0.75 이하 변수 쌍

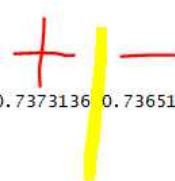
{INDUS, NOX} -> NOX를 제거(∵ 다른 변수와도 상관관계가 높음)

{RAD, TAX} -> RAD를 제거

III. 나머지 예측변수들을 감소시키기 위해서 전역 탐색방법을 사용하시오: 첫째, 성능이 좋은 상위 3개 모델을 선택하시오. 그다음으로 이 모델들을 학습 세트에 대하여 각각 구축한 후, 검증 세트에 대하여 예측정확도를 비교하시오. RMSE, 평균오차, 리프트 도표를 사용하여 비교하시오. 마지막으로 가장 좋은 모형을 선택하고 이에 대하여 기술하시오.

+ 상관계수가 높은 변수를 제거하고 전역탐색을 실시한다.

```
> sum$which
(Intercept) CRIM ZN INDUS CHAS RM AGE DIS TAX PTRATIO LSTAT
1 TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
2 TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE
3 TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE
4 TRUE FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE TRUE TRUE
5 TRUE FALSE FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE TRUE TRUE
6 TRUE FALSE FALSE TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE TRUE
7 TRUE FALSE TRUE TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE TRUE
8 TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE TRUE
9 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE
10 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> sum$adjr2
[1] 0.5649423 0.6507048 0.6923658 0.7125634 0.7235649 0.7301488 0.7354984 0.7371446 0.7373136 0.7365141
```



1) 성능이 좋은 상위 3개의 모델 선택

> 예측변수 9개인 모델( $R^2_{adj}=0.7373136$ )

```
> summary(house.lm.first)
```

```
Call:
lm(formula = MEDV ~ ., data = train.df)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.8687  -2.9771  -0.7654   2.0883  25.8976
```

```
Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  27.89532     5.06006   5.513 0.000000077503 ***
CRIM         -0.06656     0.03822  -1.742   0.08264 .
ZN           0.05178     0.01868   2.773   0.00592 **
INDUS       -0.17317     0.06475  -2.674   0.00790 **
CHAS         4.61888     1.12991   4.088 0.000056288628 ***
RM           3.67032     0.52772   6.955 0.000000000023 ***
AGE         -0.01865     0.01710  -1.090   0.27640
DIS         -1.52572     0.25538  -5.974 0.000000006691 ***
PTRATIO     -0.66000     0.15391  -4.288 0.000024490172 ***
LSTAT      -0.61682     0.06654  -9.270 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.736 on 293 degrees of freedom
Multiple R-squared:  0.7451,    Adjusted R-squared:  0.7373
F-statistic: 95.18 on 9 and 293 DF,  p-value: < 0.00000000000000022
```

> 예측변수 8개인 모델( $R^2_{adj}=0.7371446$ )

> summary(house.lm.second)

```
Call:
lm(formula = MEDV ~ ., data = train.df)

Residuals:
    Min       1Q   Median       3Q      Max
-11.023  -2.906  -0.680   2.030  25.581

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.53940    5.05115   5.452 0.0000001055636 ***
CRIM        -0.06443    0.03818  -1.687   0.09257 .
ZN          0.05417    0.01855   2.920   0.00377 **
INDUS      -0.17882    0.06456  -2.770   0.00597 **
CHAS        4.52803    1.12720   4.017 0.0000748961191 ***
RM          3.54872    0.51597   6.878 0.0000000000366 ***
DIS        -1.41591    0.23477  -6.031 0.0000000048805 ***
PTRATIO    -0.67039    0.15367  -4.363 0.0000178130013 ***
LSTAT      -0.64542    0.06117 -10.551 < 0.000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.737 on 294 degrees of freedom
Multiple R-squared:  0.7441,    Adjusted R-squared:  0.7371
F-statistic: 106.9 on 8 and 294 DF,  p-value: < 0.00000000000000022
```

> 예측변수 7개인 모델( $R^2_{adj}=0.7354984$ )

> summary(house.lm.third)

```
Call:
lm(formula = MEDV ~ ., data = train.df)

Residuals:
    Min       1Q   Median       3Q      Max
-11.2835  -2.9266  -0.6783   2.0806  25.4213

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.86243    5.00553   5.766 0.0000000204499 ***
ZN          0.04833    0.01829   2.643   0.00865 **
INDUS      -0.18895    0.06448  -2.930   0.00365 **
CHAS        4.64095    1.12873   4.112 0.0000509791912 ***
RM          3.47159    0.51555   6.734 0.0000000000863 ***
DIS        -1.33919    0.23104  -5.796 0.0000000174062 ***
PTRATIO    -0.71567    0.15178  -4.715 0.0000037299030 ***
LSTAT      -0.67377    0.05901 -11.419 < 0.000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.752 on 295 degrees of freedom
Multiple R-squared:  0.7416,    Adjusted R-squared:  0.7355
F-statistic: 121 on 7 and 295 DF,  p-value: < 0.00000000000000022
```

2) 각 모형의 평균오차와 RMSE

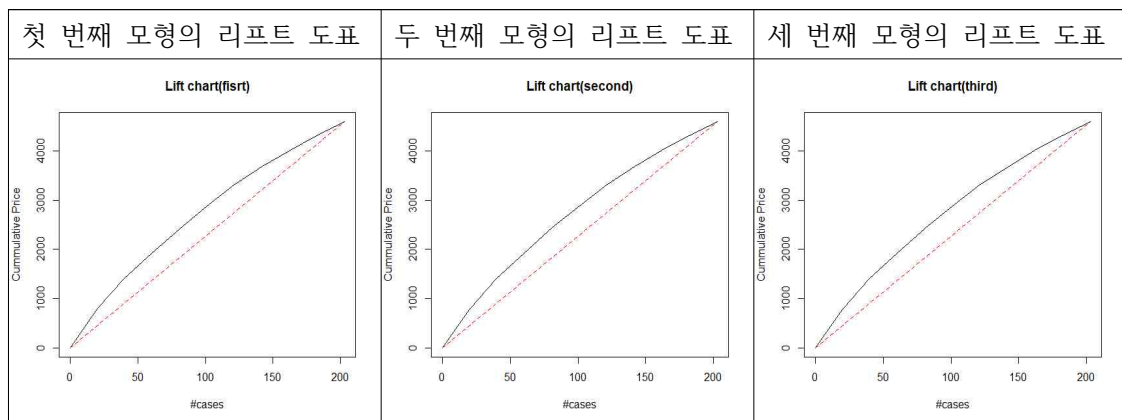
```
> accuracy(house.lm.first.pred, valid.df$MEDV)
      ME      RMSE      MAE      MPE      MAPE
Test set -0.283562  5.376751  3.771344 -6.778014 19.79887

> accuracy(house.lm.second.pred, valid.df$MEDV)
      ME      RMSE      MAE      MPE      MAPE
Test set -0.2524309  5.381714  3.790971 -6.634051 19.87957

> accuracy(house.lm.third.pred, valid.df$MEDV)
      ME      RMSE      MAE      MPE      MAPE
Test set -0.2435528  5.448385  3.854537 -6.929669 20.63523
```

-> 세 번째 모형의 RMSE이 제일 크다. 첫 번째 모형의 평균오차(ME)가 제일 작다.

3) 각 모형의 리프트 도표



> 리프트 도표의 차이점은 가시적이 않다.