Salwa Niaz Preet (sxp240086)

Business Analytics With R

Assignment 2

(a)

Because we have to predict prices for new data, it is best to partition data into training and validation sets. Partitioning data into training and validation sets ensures that a model generalizes well to new, unseen data. The main purpose of the training set is to teach the model. The model learns the underlying patterns, relationships, and structures in the data. The parameters of the model are based on this data and aim is to minimize the error. However, only using the training set can lead to overfitting, where the model performs very well on training data but fails to generalize to new data. The validation data is used to evaluate the model's performance with new, unseen data but is not used to update the model's parameters. If the model does well on the training set but badly on the validation set, it means the model might be memorizing rather than learning general patterns. Thus, validation set helps point out overfitting. Additionally. if the model stops improving on the validation set even though it keeps improving on the training set, it’s a signal that continuing training might cause overfitting. The validation set is also used to tell when to stop training.

(b)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2964 -2.3914 -0.0215  3.0246  9.4428 

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)    
(Intercept) -45.83892    5.91108  -7.755       0.000000000196 ***
CRIM          0.08188    0.69056   0.119                0.906    
CHAS         -0.41219    1.79545  -0.230                0.819    
RM           11.03200    0.92467  11.931 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.155 on 56 degrees of freedom
Multiple R-squared:  0.7323,	Adjusted R-squared:  0.718 
F-statistic: 51.06 on 3 and 56 DF,  p-value: 0.000000000000004912
```

MEDV = -45.83892 + 0.08188 CRIM – 0.41219 CHAS + 11.03200 RM

(c)

Using the estimated regression model, what median house price is predicted for a tract in the Boston area that does not bind the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6? What is the prediction error? (5 points)

CRIM = 0.1

CHAS = 0

RM = 6

MEDV = -45.83892 + 0.08188*0.1 – 0.41219*0 + 11.03200*6

=20.36127

Prediction error: 0.7093308

(d)

i. Which predictors are likely to be measuring the same thing among the 13 predictors? Discuss the relationships among INDUS, NOX, and TAX. (10 points)

CRIM - Per capita crime rate by town
ZN - Proportion of residential land zoned for lots over 25,000 ft2
INDUS - Proportion of nonretail business acres per town
CHAS - Charles River dummy variable (= 1 if tract bounds river; = 0 otherwise)
NOX - Nitric oxide concentration (parts per 10 million)
RM - Average number of rooms per dwelling
AGE - Proportion of owner-occupied units built prior to 1940
DIS - Weighted distances to five Boston employment centers
RAD - Index of accessibility to radial highways
TAX - Full-value property-tax rate per $10,000
PTRATIO - Pupil/teacher ratio by town
LSTAT - Percentage lower status of the population

DIS, RAD and TAX: DIS and RAD are factors measuring urban accessibility, which in turn influences property taxes. Areas with better access to highways or employment centers may have higher property taxes.

INDUS and NOX: Both are related to industrialization. Areas with more industrial business land might have higher pollution levels, contributing to higher NOX concentrations.

LSTAT and CRIM: Areas with a higher percentage of lower status of the population might have higher crime rates.

AGE and TAX: Places with older houses may have different tax rates due to historical and heritage preservation and infrastructure costs which can impact taxes.

TAX, INDUS, and NOX: More industrialized areas might require higher taxes for infrastructure and maintenance, and increased industrial activity can raise NOX levels.

ii) Compute the correlation table for the 12 numerical predictors and search for highly correlated pairs. These have potential redundancy and can cause multi-collinearity. Choose which ones to remove based on this table

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT
CRIM	1.0000	-0.2005	0.4066	-0.05589	0.4210	-0.2192	0.3527	-0.3797	0.62551	0.5828	0.290	0.4556
ZN	-0.2005	1.0000	-0.5338	-0.04270	-0.5166	0.3120	-0.5695	0.6644	-0.31195	-0.3146	-0.392	-0.4130
INDUS	0.4066	-0.5338	1.0000	0.06294	0.7637	-0.3917	0.6448	-0.7080	0.59513	0.7208	0.383	0.6038
CHAS	-0.0559	-0.0427	0.0629	1.00000	0.0912	0.0913	0.0865	-0.0992	-0.00737	-0.0356	-0.122	-0.0539
NOX	0.4210	-0.5166	0.7637	0.09120	1.0000	-0.3022	0.7315	-0.7692	0.61144	0.6680	0.189	0.5909
RM	-0.2192	0.3120	-0.3917	0.09125	-0.3022	1.0000	-0.2403	0.2052	-0.20985	-0.2920	-0.356	-0.6138
AGE	0.3527	-0.5695	0.6448	0.08652	0.7315	-0.2403	1.0000	-0.7479	0.45602	0.5065	0.262	0.6023
DIS	-0.3797	0.6644	-0.7080	-0.09918	-0.7692	0.2052	-0.7479	1.0000	-0.49459	-0.5344	-0.232	-0.4970
RAD	0.6255	-0.3119	0.5951	-0.00737	0.6114	-0.2098	0.4560	-0.4946	1.00000	0.9102	0.465	0.4887
TAX	0.5828	-0.3146	0.7208	-0.03559	0.6680	-0.2920	0.5065	-0.5344	0.91023	1.0000	0.461	0.5440
PTRATIO	0.2899	-0.3917	0.3832	-0.12152	0.1889	-0.3555	0.2615	-0.2325	0.46474	0.4609	1.000	0.3740
LSTAT	0.4556	-0.4130	0.6038	-0.05393	0.5909	-0.6138	0.6023	-0.4970	0.48868	0.5440	0.374	1.0000

Identifying only high correlation:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT
CRIM	NA	NA	NA	NA	NA	NA	NA	NA	0.626	0.583	NA	NA
ZN	NA	NA	-0.534	NA	-0.517	NA	-0.570	0.664	NA	NA	NA	NA
INDUS	NA	-0.534	NA	NA	0.764	NA	0.645	-0.708	0.595	0.721	NA	0.604
CHAS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NOX	NA	-0.517	0.764	NA	NA	NA	0.731	-0.769	0.611	0.668	NA	0.591
RM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.614
AGE	NA	-0.570	0.645	NA	0.731	NA	NA	-0.748	NA	0.506	NA	0.602
DIS	NA	0.664	-0.708	NA	-0.769	NA	-0.748	NA	NA	-0.534	NA	NA
RAD	0.626	NA	0.595	NA	0.611	NA	NA	NA	NA	0.910	NA	NA
TAX	0.583	NA	0.721	NA	0.668	NA	0.506	-0.534	0.910	NA	NA	0.544
PTRATIO	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
LSTAT	NA	NA	0.604	NA	0.591	-0.614	0.602	NA	NA	0.544	NA	NA

RAD and TAX have a high correlation (=0.910) so either if these two can be removed
NOX and INDUS (=0.764)
DIS and NOX (= - 0.769)
DIS and AGE (= - 0.748)
AGE and NOX (=0.731)
TAX and INDUS (=0.721)

We can remove TAX, NOX and DIS.

d(iii):

Best model for forward:

```
Step:  AIC=142
MEDV ~ RM + PTRATIO + AGE + DIS + NOX

        Df Sum of Sq RSS AIC
<none>               525 142
+ TAX    1     16.38 508 142
+ CHAS   1     15.09 509 142
+ LSTAT  1     10.18 514 143
+ RAD    1      4.42 520 144
+ CRIM   1      1.53 523 144
+ INDUS  1      0.60 524 144
+ ZN     1      0.34 524 144
```

```
> # predicting forward with validation set
> house.lm.step.predf <- predict(house.lm.stepf, valid.df1)
> accuracy(house.lm.step.predf, valid.df1$MEDV)
            ME RMSE  MAE  MPE MAPE
Test set 0.471 3.45 2.69 1.09   11
```

Best model for backward:

```
Step:  AIC=141
MEDV ~ CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO

          Df Sum of Sq  RSS AIC
<none>                  465 141
- NOX      1        17  482 141
- RAD      1        21  486 142
- CHAS     1        29  494 142
- TAX      1        39  504 144
- AGE      1        82  547 149
- DIS      1       103  568 151
- PTRATIO  1       153  618 156
- RM       1      1558 2023 227
```

```
> # predicting backward with validation set
> house.lm.step.predb <- predict(house.lm.stepb, valid.df1)
> accuracy(house.lm.step.predb, valid.df1$MEDV)
            ME RMSE MAE   MPE MAPE
Test set 0.328 3.51 2.7 0.651   11
>
```

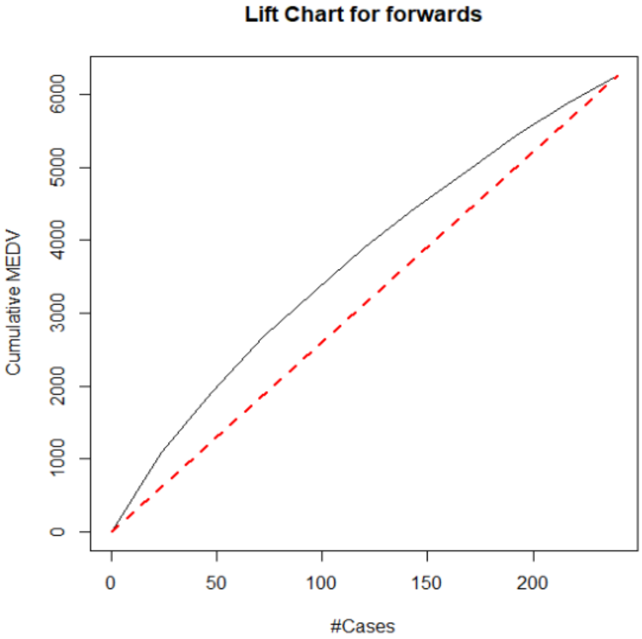
Best model for both:

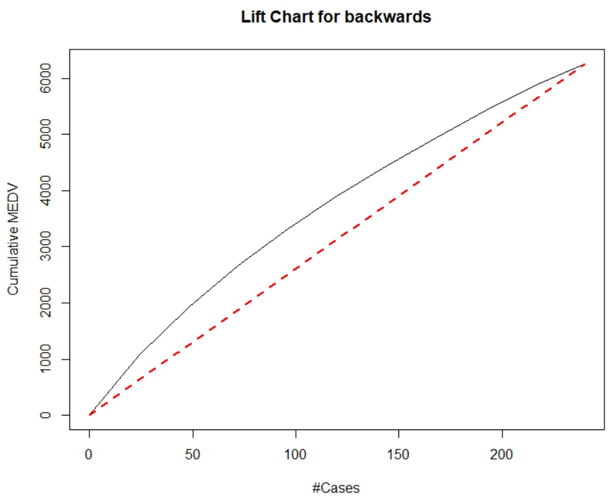
```
Step:  AIC=141
MEDV ~ CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO

          Df Sum of Sq  RSS AIC
<none>                  465 141
- NOX      1        17  482 141
- RAD      1        21  486 142
+ LSTAT    1         6  459 142
+ INDUS    1         6  459 142
+ CRIM     1         5  460 142
+ ZN       1         4  461 142
- CHAS     1        29  494 142
- TAX      1        39  504 144
- AGE      1        82  547 149
- DIS      1       103  568 151
- PTRATIO  1       153  618 156
- RM       1      1558 2023 227
```

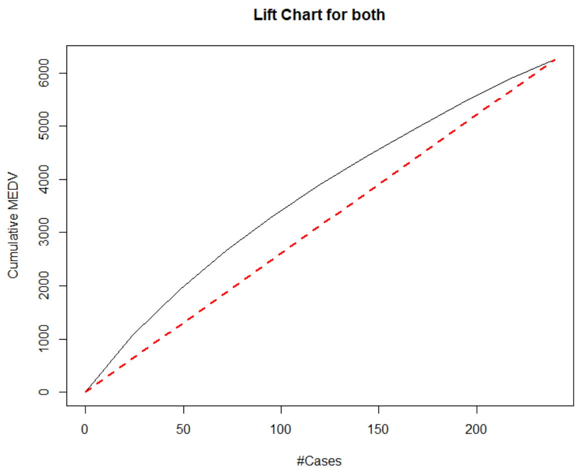
```
> # predicting both with validation set
> house.lm.step.pred <- predict(house.lm.stepfb, valid.df1)
> accuracy(house.lm.step.pred, valid.df1$MEDV)
            ME RMSE MAE   MPE MAPE
Test set 0.328 3.51 2.7 0.651   11
```

Salwa Niaz Preet (sxp240086)

```r
actual = valid.df1$MEDV

#lift for forward
gain1 = gains(actual,
              house.lm.step.predf,
              group = 10)

plot(c(0, gain1$cume.pct.of.total*sum(actual))~c(0, gain1$cume.obs), type = "l",
     xlab = "#Cases", ylab = "Cumulative MEDV", main = "Lift Chart for forwards")
segments(0, 0, nrow(valid.df1), sum(actual), lty = "dashed", col = "red", lwd = 2)

#lift for backwards
gain2 = gains(actual,
              house.lm.step.predb,
              group = 10)

plot(c(0, gain2$cume.pct.of.total*sum(actual))~c(0, gain2$cume.obs), type = "l",
     xlab = "#Cases", ylab = "Cumulative MEDV", main = "Lift Chart for backwards")
segments(0, 0, nrow(valid.df1), sum(actual), lty = "dashed", col = "red", lwd = 2)


#lift for both
gain3 = gains(actual,
              house.lm.step.predfb,
              group = 10)

plot(c(0, gain3$cume.pct.of.total*sum(actual))~c(0, gain3$cume.obs), type = "l",
     xlab = "#Cases", ylab = "Cumulative MEDV", main = "Lift Chart for both")
segments(0, 0, nrow(valid.df1), sum(actual), lty = "dashed", col = "red", lwd = 2)
```







The result comes out to be the same for backward and both. These have the following model:

```
Call:
lm(formula = MEDV ~ CHAS + NOX + RM + AGE + DIS + RAD + TAX + 
    PTRATIO, data = train.df1)

Residuals:
   Min     1Q Median     3Q    Max 
-6.732 -1.709 -0.268  1.690  8.665 

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)    
(Intercept) -4.03244   10.47936  -0.385            0.701987    
CHAS        -2.41353    1.36431  -1.769            0.082866 .  
NOX         -9.64486    7.06064  -1.366            0.177931    
RM           9.71426    0.74331  13.069 < 0.0000000000000002 ***
AGE         -0.06130    0.02043  -3.000            0.004166 ** 
DIS         -1.25940    0.37461  -3.362            0.001474 ** 
RAD          0.47640    0.31462   1.514            0.136144    
TAX         -0.01629    0.00792  -2.057            0.044839 *  
PTRATIO     -0.89559    0.21871  -4.095            0.000151 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.02 on 51 degrees of freedom
Multiple R-squared:  0.8712,	Adjusted R-squared:  0.851 
F-statistic: 43.11 on 8 and 51 DF,  p-value: < 0.00000000000000022
```

MEDV = -4.03244 - 2.41353 CHAS – 9.64486 NOX + 9.71426 RM -0.06130 AGE -1.25940 DIS + 0.47640 RAD -0.01629 TAX -0.89559 PTRATIO

This model is the preferred one because it has lower AIC (=141) than the best model chosen by forward step (=142). If we compare errors, backward and both steps gave same errors and they are not highly different from errors reported for ‘forward step’. ME and MPE (forward) are higher than those for backward/both. MAPE is the same for all three models. RMSE and MAE for forward are slightly lower than those for backward/both but because AIC is a better metric to compare models, the model above is the preferred one.