Author: Ludovic Henry

6.1/

a.

The data should be split between training and validation set to avoid overfitting by validating – with the validation set – the performance of the model without using the same set for training and validation.

b.

MEDV = -28.8107 + -0.2607 \* CRIM + 3.7630 \* CHAS + 8.2782 \* RM + ε

c.

MEDV = -28.8107 + -0.2607 \* 0.1 + 3.7630 \* 0 + 8.2782 \* 6 + ε

MEDV = 20.83243 + ε

The error (ε) is the following:

Min 1Q Median 3Q Max

-13.3015 -2.8258 -0.4292 1.8629 24.3484

d.

i.

The INDUS and NOX seems related since Nox is mostly produced by industries. DIS and INDUS also seems related since a lot of work is available close to industries.

INDUS, NOX and TAX are most likely related since industries are often driving taxes low and pollution high.

ii.

Some highly correlated pairs are :

* INDUS, NOX and TAX
* AGE and NOX
* TAX and RAD
* DIS and ZN
* DIS and AGE
* DIS and NOX
* DIS and INDUS

We remove NOX, INDUS and DIS

iii.

Here are the results for each model’s RMSE, MAPE and mean error:

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | MAPE | Mean error |
| Forward | 4.79331 | 0.1671347 | 2.514172 |
| Backward | 4.789369 | 0.16788 | 2.519711 |
| Both | 4.789369 | 0.16788 | 2.519711 |

The best model is thus Forward since it has the lowest MAPE and Mean error and highest RMSE. It also has the lowest R-squared adjusted from the 3 models.

6.2/

6.3/

10.4/