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/

a.

b.

These two relations do not look linear



c.

ii.

Spending = 10.142833 + 91.689375 \* FREQ + -0.007005 \* LAST\_UPDATE + 10.829880 \* WEB + -3.211925 \* GENDER(Male) + -74.767837 \* ADDRESS\_RES + -11.478364 \* ADDRESS\_US + ε

iii.

The people most likely to spend are frequent buyers who purchase through the web and which address is non-residential and outside the US. The gender and time to last update has little influence.

iv.

The first predictor to be dropped is GENDER(MALE), followed by ADDRESS\_US.

v.

The first record in the validation partition is:

FREQ: 1, LAST\_UPDATE: 1008, WEB: 0, GENDER(MALE): 0, ADDRESS\_US: 1, ADDRESS\_RES: 0, and Spending: 0

The predicted spending is:

10.142833 + 91.689375 \* 1 + -0.007005 \* 1008 + 10.829880 \* 0 + -3.211925 \* 0 + -74.767837 \* 0 + -11.478364 \* 1 = 83.29

The predicted error is:

Predicted Spending – Spending = 83.29

vi.

The RMSE is 139.6446.

vii.

The distribution seems to follow a normal distribution with the median between 0 and 100.

6.3/

a.

DISTANCE is the most correlated predictor with FARE

b.

SW has the biggest difference between its 2 values

c.

ii.

10.4/

a.

b.

c.

d.

e.

f.

g.

h.

i.

j.