Business Analytics with R – Assignment 2

Name – Amruth Pai Thukaram

Net ID – AXT230147

Date – 4th October 2024

A]

data should be partitioned into training and validation sets because it helps in models performance on unseen data. Below are the few reasons

Prevent overfitting: - by training the model on training set we reduce the risk of the model overfitting to noise or specific patters in the data.

Assess model performance: - validation set provides an unbiased evaluation of the model. We can add or remove parameters based on performance of the model. If the model is performing well on validation set then it generalizes well to other datasets

Model selection: - different model configurations can be compared based on their performance on the validation set.

We use training set to fit the data to the model, to learn the patterns, relationships and weights based on predictors and target variable. Also optimize the model parameters to minimize the errors to improve the prediction accuracy.

We use validation set to check the performance of the model on unseen data, tune the data like choosing the number of features during model building and ensure the model doesn’t overfit on training data.

B]

A computer code with black text

Description automatically generated

A white background with black text

Description automatically generated

MEDV= -25.019 -0.301\*CRIM + 5.238\*CHAS + 7.644\*RM

C]

A computer screen shot of a math equation

Description automatically generated

A computer code with numbers and symbols

Description automatically generated

D]

i]

Below predictors are likely to be measuring the same thing among the 13 predictors because of high correlation (>0.7)

TAX, INDUS = 0.72 correlation

NOX, INDUS = 0.76 correlation

NOX, AGE = 0.73 correlation

TAX, RAD = 0.91

A computer code with numbers and symbols

Description automatically generated

A close up of numbers

Description automatically generated

we see high correlation between INDUS, NOX and TAX which means when we increase one factor another factor also increases.

TAX, NOX = if we increase Full-value property-tax rate per $10,000 then Nitric oxide concentration also increases

TAX, INDUS = if we increase Full-value property-tax rate per $10,000 then Proportion of nonretail business acres per town also increases

NOX, INDUS = if we increase Nitric oxide concentration then Proportion of nonretail business acres per town also increases

D: ii]

A computer code with black and white text

Description automatically generated

A red and blue squares with numbers

Description automatically generated

From the above correlations heatmap we can observe that there are below predictors that are highly correlated (>0.7)

MEDV, RM = more the average number of rooms per dwelling more the MEDV

TAX, INDUS = more nonretail business acres per town more the tax rate per property

AGE, NOX = more the proportion of owner-occupied more the nitric oxide concentration

NOX, INDUS = more the proportion of nitric oxide concentration more the proportion of nonretail business

TAX, RAD = more the Full-value property-tax rate per $10,000 more the Index of accessibility to radial highways

We can remove NOX, INDUS and TAX

D iii]

A screenshot of a computer program

Description automatically generated

A computer code with text

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated

Performance of each model after removing highly correlated variables

A close-up of a computer code

Description automatically generated

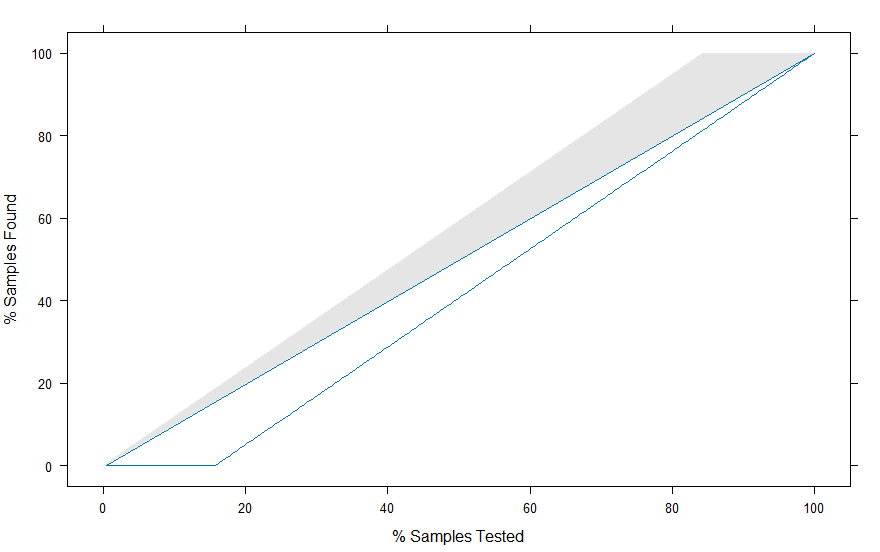
Lift charts: -

Backwards:

A graph with a line

Description automatically generated

Forward:



Both:

A graph with a line

Description automatically generated

All the 3 models i.e. forward, backwards and both give the same results for RMSE, MAPE and MEAN error which concludes that we have selected strong predictors, and we have low multicollinearity among the predictor variables.