Exercise 1: Regression

Using the dataset `CRM_class.csv` develop a regression model for the prediction of sales from the other variables except the variable `Train`.

Points to consider:

1. Descriptive summaries for the variables
2. Use a splitting of the data into a training and test sample (new random selection or using the `Train` specification
3. Select a model using variable selection
4. Apply the trained model to the test sample and evaluate the errors.

Exercise 2: Credit Data

The data set Credit.csv contains cases of 1000 credit applications. The meaning of the variables is as follows:

`Default:` Whether the credit was defaulted (1) or not (0)
`Duration:` Runtime of credit in month
`Amount:` Credit amount
`Installment:` Installment rate as percentage of disposable income
`Age:` Age of the person in years
`ForeignWorker:` indicator whether the person is a foreign worker (1) or not (0)
`Rent:` whether the person lives in a rented house (1) or not (0)
`HistoryPoor:` existing credit or credits in the past had delay in paying off (1) or not (0)
`HistoryTerrible:` critical account/person has other credits (1) or not (0)

Perform the following analyses and answer the questions

a) Compute descriptive statistics for the variables duration, age, amount and installment. Are there any visible differences in the variables between persons whose credit defaulted or not?
b) Split the data randomly in a training set of 800 cases and a test set of 200 cases. Compute the model from the training data and predict the values for the test data. Compare the classification tables for training and test set.
c) How many persons have good history (not poor and not terrible) in the training and the test set?
d) Display a cross tabulation of rent and foreign workers for training and test set.
e) Apply at least two classification methods for the training set.
f) Determine the ROC-curves for the models and the areas under the curve.
g) Display the classification matrix for the data with a threshold of 0.5.
h) Suppose that lending to a person who defaults has costs three times the costs of not lending to a person which would be a good debtor. What are the costs of the decisions according to the estimated model? Can you improve costs by changing the threshold for misclassification? If yes, which threshold you would suggest?