Packages

All packages that are used in this script are described and loaded in. The used functions of the packages are noted in brackets (). For more information see the corresponding help files.

- *QuantPsych*: used for extracting the standardized coefficients (lm.beta)
- *dplyr*: used for general data transformation (select, mutate, rename, inner_join)
- *glmnet*: used for conducting the lasso method (cv.glmnet, glmnet)
- *gam*: used for constructing a generalized additive model containing a smoothing spline (gam)
- *gbm*: used for conducting a gradient boosting machine (gbm)
- *caret*: used for fitting and training the gradient boosting machine (trainControl, expand.grid, train)
- *car*: used for plotting the residuals (residualPlot, redidualsPlot)
- *bestglm*: used for finding the best subset selection of a logistic regression (bestglm)

thesisFunctions.R file

Other packages and functions are loaded in via "thesisFunctions.R" file. These packages are used in the created functions and described below. The used functions of the package are again noted in brackets (). For more information see the corresponding help files.

- *psych*: used for descriptive statistics, a correlation test and for measuring Cronbach's alpha (describe, corr.test, alpha)
- *glmnet*: Used for conducting the lasso method and for plotting the shrinkage of the coefficients (glmnet)
- *corrplot*: used for a visualisation of the correlations (corrplot)
- *leaps*: used for conducting a best subset selection method of a linear regression model (regsubsets)
- *splines*: used for conducting a smoothing spline (smooth.spline)
- *ROCR*: used for plotting a ROC curve and accuracy/F-score vs. threshold plot (prediction, performance)

Created functions loaded in via "thesisFunctions.R". For more information see the corresponding file.

- *plotBoxHist*: plots a standardized boxplot and a histogram in one figure. The histogram includes a mean and median line. Used in the function describeVariable.
- *plotBoxBar*: plots a standardized boxplot and a bar plot in one figure. Used in describeVariable.
- *plotBox*: plots a standardized bar plot. Used in describeVariable.
- *describeVariable*: describes a variable and returns the proper descriptive statistics. The function changes the description and plots based on whether the type of variable is a factor, discrete- or a continuous variable. In case of the latter, the parameter histogram = TRUE needs to be supplied.
- *plotCor*: plots a detailed correlation plot, if pvalues = FALSE, all correlations are coloured gradually from -1 to +1, negative relations are coloured reddish and positive relationships are coloured blueish. If pvalues = TRUE, all significant relations are coloured gradually, all other insignificant relations will have a white background. Returns the correlation table and if pvalues = TRUE also the p-values table.
- *cbAlpha*: calculates the Cronbach's alpha score, once for the normal input and once for the z-scaled variables of this input. Returns detailed scores and the correlations between the variables.
- *plotBoxes*: plots standardized boxplots for all preferences in one figure, split by a binary factor variable.

- *baseModelcv*: creates a base model and applies cross-validation. If a classification model is appropriate, then the majority class is taken as baseline and accuracy is measured. When a regression model is proper, RMSE is measured and the mean is used as a baseline. Returns the cross-validated score for every fold and the mean of these folds.
- *bestSubsetcv*: applies best subset selection and cross-validates scores for every best subset corresponding with that number of variables. In case of a classification, logistic regression is conducted, and all best models are supplied as a parameter. These are externally found via the bestglm function. In case of a regression model, linear regression is used, and all best models are internally found using regsubsets and the R2 score. Returns a matrix with cross-validated scores for every fold with that number variables, for all number of variables. Also returns the standard deviation, the mean for every number of variables, the 'true' best number of variables. and the corresponding score and standard deviation.
- *plotR2Subs*: plots the best R2 score for every number of variables and the variables corresponding with these scores in one figure. Returns the summary of regsubset.
- *plotModels*: plots the output scores of cross-validated models and places a red dot at the best performing model in the plot and returns these scores.
- *plotThreshold*: plots an Accuracy- and F-score- vs. threshold plot in one figure. If plotcontent = "ROC", a ROC curve is plotted, and the AUC is returned.
- *plotShrinkage*: generates a glmnet model and then plots the shrinkage the coefficients. The red dotted line corresponds with the best lambda and a blue dotted line corresponds with the best lambda plus one standard error.
- *resultsLasso*: returns the results of the lassocv in a standardized way as in the other created cv functions. Returns the 'true' best lambda and the corresponding score and standard deviation and the best lambda plus one standard error. Takes a measure input and adjusts the output in case of a general linear model.
- *plotBivariate*: plots a bivariate plot with a x variable and y variable and draws the general linear model or linear model best fit line. The plot and fit are adjusted if y == binary variable.
- *nonLinearcv*: cross-validation is applied on non-linear models. The non-linear models should be supplied. Returns the standardized output as in bestsubsetcv.
- *smoothSplinecv*: cross-validation is applied on smoothing splines for a range of supplied degrees of freedom from 2 until a supplied parameter of maxdegree. Returns the standardized output as in bestsubsetcv, but with the degrees of freedom included.