Homework 6

The homework is about using classical methods for classification.

Homework uses real data that was available from https://www.bondora.com/en/public-reports#dataset-file-format in 2023 (just selected variables and reduced number of observations)

Use the data from `HW6_Loan.csv`. The aim is to predict probability of default of new applications. Note that again `Education`,`Gender` and `OccupationArea` are actually nominal variables. + Use the command `set.seed(your_study_book_nuber)` to fix randomness (`your_study_book_number` should be equal to the numerical part of your study book number). Split the data into training and model comparison parts (equal size) + Fit the best logistic regression model you can find in reasonable amount of time to the training data and describe what you tried and why the final model is the best of the ones you compared. + Produce a box plot, which shows predicted probabilities for defaulting loans and for non-defaulting loans on model comparison set. + Produce confusion matrix when using the model for classification the observations in the model comparison set with cutoff corresponding to probability of default (default=Yes) equal to 0.3. Can defaulting loans be accurately predicted? + Fit LDA model to the training data leaving out `OccupationArea` and using all other variables as numeric variables and produce box plot of probabilities of default for defaulting loans and non-defaulting loans in the case of model comparison set(predicted probabilities for classes can be obtained by `predict(model)$posterior`) + Compute confusion matrix for predictions of LDA model for observations in the model comparison set in the case of standard cutoff and when the cutoff probabilty 0.3 of default=Yes is used + Repeat the last two steps when using QDA model + Summarize your findings - which model performed the best on model comparison data?