Chapter 4

Example 4.3.1

# Code to use for binning

# 4.1.1 woe\_months\_to\_maturity

train$woe\_months\_to\_maturity<- rep(NA, length(train$months\_to\_maturity))

train$woe\_months\_to\_maturity[which(is.na(train$months\_to\_maturity))] <- -0.6796

train$woe\_months\_to\_maturity[which(train$months\_to\_maturity <= 48)] <- 1.0169

train$woe\_months\_to\_maturity[which(train$months\_to\_maturity > 48 & train$months\_to\_maturity <= 132)] <- -0.2135

train$woe\_months\_to\_maturity[which(train$months\_to\_maturity > 132)] <- -0.6796

test$woe\_months\_to\_maturity<- rep(NA, length(test$months\_to\_maturity))

test$woe\_months\_to\_maturity[which(is.na(test$months\_to\_maturity))] <- -0.6796

test$woe\_months\_to\_maturity[which(test$months\_to\_maturity <= 48)] <- 1.0169

test$woe\_months\_to\_maturity[which(test$months\_to\_maturity > 48 & test$months\_to\_maturity <= 132)] <- -0.2135

test$woe\_months\_to\_maturity[which(test$months\_to\_maturity > 132)] <- -0.6796

# 4.1.2 ltv\_utd

train$woe\_ltv\_utd<- rep(NA, length(train$ltv\_utd))

train$woe\_ltv\_utd[which(is.na(train$ltv\_utd))] <- -0.6243

train$woe\_ltv\_utd[which(train$ltv\_utd <= 0.6078)] <- 1.3041

train$woe\_ltv\_utd[which(train$ltv\_utd > 0.6078 & train$ltv\_utd <= 0.7617)] <- 0.1313

train$woe\_ltv\_utd[which(train$ltv\_utd > 0.7617 & train$ltv\_utd <= 0.9196)] <- -0.5696

train$woe\_ltv\_utd[which(train$ltv\_utd > 0.9196)] <- -0.6243

test$woe\_ltv\_utd<- rep(NA, length(test$ltv\_utd))

test$woe\_ltv\_utd[which(is.na(test$ltv\_utd))] <- -0.6243

test$woe\_ltv\_utd[which(test$ltv\_utd <= 0.6078)] <- 1.3041

test$woe\_ltv\_utd[which(test$ltv\_utd > 0.6078 & test$ltv\_utd <= 0.7617)] <- 0.1313

test$woe\_ltv\_utd[which(test$ltv\_utd > 0.7617 & test$ltv\_utd <= 0.9196)] <- -0.5696

test$woe\_ltv\_utd[which(test$ltv\_utd > 0.9196)] <- -0.6243

# 4.1.3 tob

train$woe\_tob<- rep(NA, length(train$tob))

train$woe\_tob[which(is.na(train$tob))] <- 0.7888

train$woe\_tob[which(train$tob <= 117)] <- -0.6278

train$woe\_tob[which(train$tob > 117 & train$tob <= 137)] <- -0.2228

train$woe\_tob[which(train$tob > 137 & train$tob <= 153)] <- 0.2290

train$woe\_tob[which(train$tob > 153)] <- 0.7888

test$woe\_tob<- rep(NA, length(test$tob))

test$woe\_tob[which(is.na(test$tob))] <- 0.7888

test$woe\_tob[which(test$tob <= 117)] <- -0.6278

test$woe\_tob[which(test$tob > 117 & test$tob <= 137)] <- -0.2228

test$woe\_tob[which(test$tob > 137 & test$tob <= 153)] <- 0.2290

test$woe\_tob[which(test$tob > 153)] <- 0.7888

# 4.2. Multivariate analysis

##Extract WoE variables and check for multicollinearity

library(dplyr)

woe\_vars<- train %>%

 dplyr::select(starts\_with("woe"))

#Calculate Spearman rank correlation, since we're interested in solving a binary classification problem (1/0)

woe\_corr<- cor(as.matrix(woe\_vars), method = 'spearman')

class(woe\_corr)

View(woe\_corr)

#Inspect graphically:

library(corrplot) #Plotting correlations/matrices

library(corrgram) #Correlation calculations

corrplot(woe\_corr, method = 'number')

#A threshold can be set (e.g. 0.70) above which correlated variables need to excluded: here max\_arrears\_balance\_6m is excluded:

# woe\_vars\_clean<- woe\_vars %>%

# dplyr::select( -woe\_max\_arrears\_bal\_6m)