Exercise 2

Before we vary the decision threshold, lets see if we can find the default 50% threshold in our model.

To do that, let's look at the posterior class probabilities and count how many of them are higher than 50%. These are the observations that are classified YES.

We plot these probabilties together with true class values and compare it with the plot p1 from exercise 1 – just to see if we are right.

The plots also help us to see whats going on after we will have changed the threshold in the next step.

```
# Varying the threshold

   # Recreating the above prediction manually from the probabilies using the same 50% threshold the LDA uses
   head(lda.pred$posterior)
   sum(lda.pred$posterior[,2] >= 0.5) # how many observations are classified Yes
   sum(lda.pred$posterior[,2] < 0.5)  # how many observations are classified No

   # Plotting the predicted classes and probabilities once more
   lda.prob.df <- data.frame(balance=test.data$balance,lda.prob=lda.pred$posterior[,2]) # make it a data frame for plotting
   p2 <- ggplot() + geom_point(data = lda.prob.df, aes(x=balance, y=lda.prob, col=test.data$default), size=5) +
     geom_hline(yintercept = 0) + geom_hline(yintercept = 1) + geom_hline(yintercept = 0.5, linetype="dashed") + ylim(0,1)
   grid.arrange(p1, p2, nrow = 1)
   par(mfrow=c(1,1))
```

Try for yourself !

Inspect the code and plot, and try to interpret it!

Exercies 2

Now we change the threshold from 50% to 20% - we have to do this manually:

- First we find all the observations with posterior probability of >=0.2, and see how many we have here.
- We then manually reclassifiy them: We create a new variable `lda.reclassified`, that we use instead of `lda.pred$class. (lda.pred$class` is where the predict function had stored the class predictions in exercise 1.)
- We set it to „No“ for all observations except the ones with posterior >=0.2.
- We then do the same plots as before, but with the manually reclassified

```
# Imposing a lower threshold for Yes
sum(lda.pred$posterior[,2] >= 0.2) # how many observations are classified Yes with a 20% threshold
# reclassify
  lda.reclassified <- rep("No", length(lda.class))
  lda.reclassified[lda.pred$posterior[,2] >= 0.2] <- "Yes"
  # Plotting the new classification
  lda.reclassified.df <- data.frame(balance=test.data$balance,lda.reclassified=lda.reclassified) # make a data frame for plotting
  p3 <- ggplot() + geom_point(data = lda.reclassified.df, aes(x=balance, y=lda.reclassified, col=test.data$default), size=5)
  lda.prob.df <- data.frame(balance=test.data$balance,lda.prob=lda.pred$posterior[,2]) # make a data frame for plotting
  p4 <- ggplot() + geom_point(data = lda.prob.df, aes(x=balance, y=lda.prob, col=test.data$default), size=5) +
    geom_hline(yintercept = 0) + geom_hline(yintercept = 1) + geom_hline(yintercept = 0.2, linetype="dashed") + ylim(0,1)
  grid.arrange(p3, p4, nrow = 1)
  par(mfrow=c(1,1))
```

Try for yourself and inspect the plots!

Particularly, compare the plots produced now with the plots produced with the original class predictions of step 1.