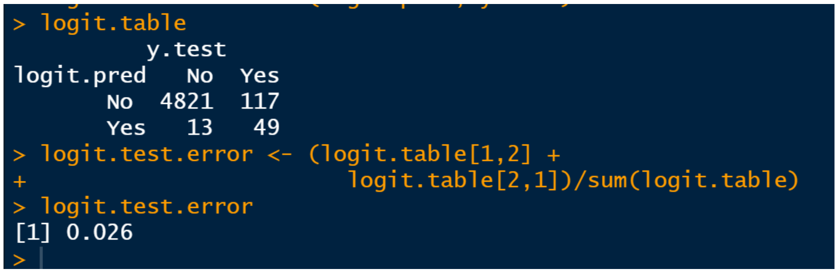
5. In Chapter 4, we used logistic regression to predict the probability of default using income and balance on the Default data set. We will now estimate the test error of this logistic regression model using the k-fold cross validation approach. Use set.seed(1018) before solving problems. (You must copy your R code together with your answer)

(a) Fit a logistic regression model that uses income and balance to predict default.

```
set.seed(1018)
dim(Default)[1]
train <- sample(1:dim(Default)[1],5000)
Default.test <- Default[-train, ]
y.test <- Default$default[-train]
logistic.fit <- glm(default ~ balance + income, data = Default,
                    subset = train, family = binomial)
logit.prob <- predict(logistic.fit, Default.test,
                      type = "response")
logit.pred <- rep("No", dim(Default.test)[1])
logit.pred[logit.prob > 0.5] = "Yes"
logit.table <- table(logit.pred, y.test)
logit.table
logit.test.error <- (logit.table[1,2] +
                       logit.table[2,1])/sum(logit.table)
logit.test.error
```



```
> logit.table
          y.test
logit.pred   No  Yes
       No  4821  117
       Yes   13   49
> logit.test.error <- (logit.table[1,2] +
+                        logit.table[2,1])/sum(logit.table)
> logit.test.error
[1] 0.026
>
```

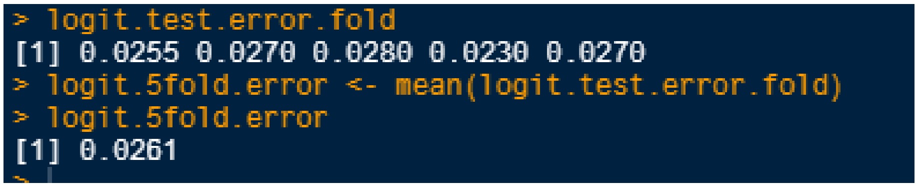
(b) Using the 5-fold cross validation, estimate the test error of the model that uses income and balance to predict default.

```
kfolds <- 5
folds <- rep_len(1:kfolds, 10000)
folds <- sample(folds, 10000)
logit.test.error.fold <- rep(0, kfolds)

for(k in 1:kfolds){
  fold <- which(folds == k)
  Default_train <- Default[-fold, ]
  Default_test <- Default[fold, ]

  logistic.fit <- glm(default ~ balance + income, data = Default_train,
                      family = binomial)
  logit.prob <- predict(logistic.fit, Default_test,
                        type = "response")
  logit.pred <- rep("No", dim(Default_test)[1])
  logit.pred[logit.prob > 0.5] = "Yes"
  logit.table <- table(logit.pred, Default_test$default)
  logit.test.error <- (logit.table[1,2] +
                         logit.table[2,1])/sum(logit.table)
  logit.test.error.fold[k] <- logit.test.error
}

logit.test.error.fold
logit.5fold.error <- mean(logit.test.error.fold)
logit.5fold.error
```



```
> logit.test.error.fold
[1] 0.0255 0.0270 0.0280 0.0230 0.0270
> logit.5fold.error <- mean(logit.test.error.fold)
> logit.5fold.error
[1] 0.0261
>
```

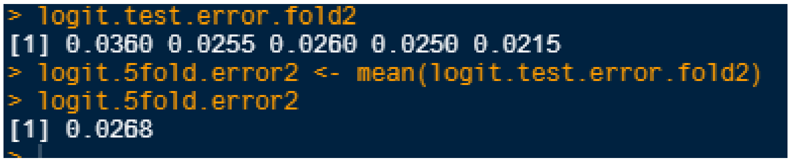
(c) Now consider a logistic regression model that predicts the probability of default using income, balance, and student. Estimate the test error for this model using the 5-fold cross validation. Comment on whether or not including student leads to an improvement in the test error rate.

```
kfolds2 <- 5
folds2 <- rep_len(1:kfolds2, 10000)
folds2 <- sample(folds2, 10000)
logit.test.error.fold2 <- rep(0, kfolds)

for(k in 1:kfolds2){
  fold2 <- which(folds2 == k)
  Default_train2 <- Default[-fold2, ]
  Default_test2 <- Default[fold2, ]

  logistic.fit2 <- glm(default ~ balance + income +
                       student, data = Default_train2,
                     family = binomial)
  logit.prob2 <- predict(logistic.fit2, Default_test2,
                       type = "response")
  logit.pred2 <- rep("No", dim(Default_test2)[1])
  logit.pred2[logit.prob2 > 0.5] = "Yes"
  logit.table2 <- table(logit.pred2, Default_test2$default)
  logit.test.error2 <- (logit.table2[1,2] +
                        logit.table2[2,1])/sum(logit.table)
  logit.test.error.fold2[k] <- logit.test.error2
}

logit.test.error.fold2
logit.5fold.error2 <- mean(logit.test.error.fold2)
logit.5fold.error2
```



```
> logit.test.error.fold2
[1] 0.0360 0.0255 0.0260 0.0250 0.0215
> logit.5fold.error2 <- mean(logit.test.error.fold2)
> logit.5fold.error2
[1] 0.0268
>
```

Including the “student” status variable does not lead to an improvement in the test error rate.