assignment5-part2

yunzhi wang

1.split the data

```
set.seed(600)
library(caret)

## Warning: package 'caret' was built under R version 3.4.3

## Loading required package: lattice

## Loading required package: ggplot2

## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone
'zone/tz/2018c.
## 1.0/zoneinfo/America/Chicago'

library(MASS)
data("GermanCredit")
mydata <- GermanCredit
mypredictors <-
c(10,1,2,3,9,11,12,13,15,16,17,20,22,23,25,26,29,31,32,38,43,46,47,53,57,59)
mydata <- mydata[,mypredictors]

mydata.split <- sample(1:nrow(mydata), size = 0.7 * nrow(mydata))
Train <- mydata[mydata.split,]
Holdout <- mydata[-mydata.split,]
```

2.build LDA and QDA analysis

LDA model

```
set.seed(123)
model.lda <- lda(Class ~ ., data = Train, CV= FALSE)
lda.holdout <- predict(model.lda, newdata = Holdout)$class
cm.lda <- table(lda.holdout, Holdout[,1])

prop.table(cm.lda)

##
## lda.holdout        Bad       Good
##        Bad  0.15333333 0.08666667
##        Good 0.19333333 0.56666667
```

Interpretation: 72% of the test dataset are correctly predicted. False positive rate is 8.7%.

QDA model

```r
set.seed(123)
model.qda <- qda(Class ~ ., data = Train, CV = FALSE)
qda.holdout <- predict(model.qda, newdata = Holdout)$class
cm.qda <- table(qda.holdout, Holdout[, 1])

prop.table(cm.qda)
```

```
##
## qda.holdout       Bad      Good
##        Bad  0.1966667 0.1666667
##        Good 0.1500000 0.4866667
```

Interpretation: 69% of the test sample are accurately predicted. False positive rate is 17%. The accuracy decreases by 3% compared to LDA, false positive prediction also increases by 8.3%. Clearly LDA works better than QDA.

3. Ensembel

```r
set.seed(123)
#Load the previous data
tree.pred.train <- read.csv('./treePredTrain.csv', header
                            = FALSE)
tree.pred.test <- read.csv('./treePredTest.csv', header =
                             FALSE)
lr.pred.train <- read.csv('./lrPredTrain.csv', header =
                            FALSE)
lr.pred.test <- read.csv('./lrPredTest.csv', header =
                           FALSE)

lda.pred.train <- predict(model.lda)$class
lda.pred.test <- predict(model.lda, newdata = Holdout)$class
qda.pred.train <- predict(model.qda)$class
qda.pred.test <- predict(model.qda, newdata = Holdout)$class

ensemble.train <- data.frame(matrix(nrow = 700, ncol = 5))
ensemble.test <- data.frame(matrix(nrow = 300, ncol = 5))

colnames(ensemble.train) <- c("Tree", "LR", "LDA", "QDA", "Ensemble")
colnames(ensemble.test) <- c("Tree", "LR", "LDA", "QDA", "Ensemble")

ensemble.train[,1] <- tree.pred.train
ensemble.train[,2] <- lr.pred.train
ensemble.train[,3] <- lda.pred.train
ensemble.train[,4] <- qda.pred.train
ensemble.test[,1] <- tree.pred.test
ensemble.test[,2] <- lr.pred.test
ensemble.test[,3] <- lda.pred.test
ensemble.test[,4] <- qda.pred.test

#Ensemble model function
```

```
ensemble.model <- function(x)
{for (i in 1 : nrow(x))
  {
    good.ct <- 0
    bad.ct <- 0
    for (j in 1 : (ncol(x) - 1))
    {
      if (x[i, j] == 'Good')
      {
        good.ct <- good.ct + 1
      } else
      {
        bad.ct <- bad.ct + 1
      } }
    if (good.ct > bad.ct)
    {
      x[i, ncol(x)] = 'Good'
    } else
    {
      x[i, ncol(x)] = 'Bad'
    } }

  return(x[, ncol(x)])
}

ensemble.pred.train <- ensemble.model(ensemble.train)
head(ensemble.pred.train)
```

```
## [1] "Bad"  "Good" "Good" "Good" "Bad"  "Bad"
```

```
ensemble.pred.test <- ensemble.model(ensemble.test)
head(ensemble.pred.test)
```

```
## [1] "Good" "Good" "Good" "Good" "Good" "Good"
```

```
#Condusion Matirx
cm.ensemble.train <- table(ensemble.pred.train, Train[, 1])
prop.table(cm.ensemble.train)
```

```
##
## ensemble.pred.train       Bad      Good
##                Bad  0.1728571 0.1000000
##                Good 0.1071429 0.6200000
```

```
cm.ensemble.test <- table(ensemble.pred.test, Holdout[, 1])
prop.table(cm.ensemble.test)
```

```
##
## ensemble.pred.test       Bad      Good
##               Bad  0.1833333 0.1266667
##               Good 0.1633333 0.5266667
```

Interpretation: For the train dataset, 79% of data are correctly predicted, and the false positive rate is 10%. In the holdout part, 72% of the data are correcly predicted, the false positive rate is 13%.

Between these models, we can tell that for the ensemble model, the false positive rate decreases a lot. The QDA did not perform as well as LDA, however, its negative effect was largely reduced in the ensemble model. Also the tree model's effect is largely reduced in the ensemble model. The ensemble model achieves a better balance by combining these models. The ensemble model gives a better result by reducing the variances. Even though the improvement is not big, but this is the best model compared to the other individual 4 models.