PredictionW4Project

Sheng Xu

5/16/2020

```r
# load libraries

library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```r
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
# download csv files and identify NA
downloadcsv <- function(url, nastrings) {
    temp <- tempfile()
    download.file(url, temp)
    data <- read.csv(temp, na.strings = nastrings)
    unlink(temp)
    return(data)
}

training_url <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
training <- downloadcsv(training_url, c("", "NA", "#DIV/0!"))

testing_url <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
testing <- downloadcsv(testing_url, c("", "NA", "#DIV/0!"))
```

```r
# Remove NA columns from the data set
comps <- complete.cases(t(training)) & complete.cases(t(testing))
training_clean <- training[,comps]
testing_clean <- testing[,comps]
```

```r
# Remove data that are not useful
training_clean <- training_clean[,-c(1,3,4,5,6,7)]
testing_clean <- testing_clean[,-c(1,3,4,5,6,7)]
```

```r
# splitting data into 2 sets for further validation
set.seed(123456)
```

```r
trainingset <- createDataPartition(training_clean$classe, p = 0.8, list = FALSE)
Training <- training_clean[trainingset, ]
Validation <- training_clean[-trainingset, ]

# Model Fitting and Results
modFit <- randomForest(classe~., data=Training)

TrainingResults <- predict(modFit, Training)
TrainingAccuracy <- sum(TrainingResults==Training$classe)/length(TrainingResults)
paste("Accuracy on training set =",TrainingAccuracy)
```

```
## [1] "Accuracy on training set = 1"
```

```r
ValidationResults <- predict(modFit, newdata=Validation)
ValidationAccuracy <- sum(ValidationResults==Validation$classe)/length(ValidationResults)
paste("Accuracy on validation set =",ValidationAccuracy)
```

```
## [1] "Accuracy on validation set = 0.99719602345144"
```

```r
TestingResults <- predict(modFit, newdata=testing_clean)
print("Classifications on the test set:"); TestingResults
```

```
## [1] "Classifications on the test set:"
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  B  A  B  A  A  E  D  B  A  A  B  C  B  A  E  E  A  B  B  B
## Levels: A B C D E
```