Prediction-Assignment

1/24/2023

1. Project goal

The goal of the project is to predict the manner in which 6 individuals exercised using data from accelerometers on the belt, forearm, arm, and dumbell, using machine learning algorithm.

They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. The main goal of the project is to predict the manner in which 6 participants performed those exercise. This is the "classe" variable in the training set, the one we aim to predict.

2. Data Loading and Cleaning

a. Data Source & Reproduceability

The training data for this project are available here: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv

More information on the experiment is available from the website here: http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset)

The following packages are needed to reproduce the results of this project : caret, rio.

b. Partition of the training set (for cross validation)

In order to get out-of-sample errors, we split the training data in training (75%) and testing (25%) data subsets.

```
set.seed(3011)
inTrain <- createDataPartition(y=training$classe, p=0.75, list=FALSE)
TrainSet <- training[inTrain, ]
TestSet  <- training[-inTrain, ]
```

c. Removing the near zero variables as well as the "mostly NAs" variables

Both created datasets have 160 variables.

```
#NZV
NZV <- nearZeroVar(TrainSet)
TrainSet <- TrainSet[, -NZV]
TestSet  <- TestSet[, -NZV]
# remove "mostly NAs" variables
```

```
AllNA    <- sapply(TrainSet, function(x) mean(is.na(x))) > 0.95
TrainSet <- TrainSet[, AllNA==FALSE]
TestSet  <- TestSet[, AllNA==FALSE]
```

```
## [1] 59
```

There are now 59 variables remaining, vs 160 initially.

d. A quick glance at the data in the classe variable (the one we aim to predict)

```
print(table(TrainSet$classe))
```

```
## 
##    A    B    C    D    E 
## 4185 2848 2567 2412 2706
```

Model building : random forest

We have uses K- fold Cross Validation for 3 iterations to create a number of partitions of sample observations, known as the validation sets, from the training data set. After fitting a model on to the training data, its performance is measured against each validation set and then averaged, gaining a better assessment of how the model will perform when asked to predict for new observations.

```
set.seed(301)
#for the K-fold
controlRF <- trainControl(method="cv", number=3, verboseIter=FALSE)
#model
modelRF <- train(classe ~ ., data=TrainSet, method="rf", trControl=controlRF)
```

```
prediction <- predict(modelRF, TestSet)
confusionMatrix(prediction, as.factor(TestSet$classe))
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    A    B    C    D    E
##          A 1395    0    0    0    0
##          B    0  949    0    0    0
##          C    0    0  855    0    0
##          D    0    0    0  804    0
##          E    0    0    0    0  901
## 
## Overall Statistics
## 
##                Accuracy : 1
##                  95% CI : (0.9992, 1)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
```

```
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            1.0000   1.0000   1.0000   1.0000   1.0000
## Specificity            1.0000   1.0000   1.0000   1.0000   1.0000
## Pos Pred Value         1.0000   1.0000   1.0000   1.0000   1.0000
## Neg Pred Value         1.0000   1.0000   1.0000   1.0000   1.0000
## Prevalence             0.2845   0.1935   0.1743   0.1639   0.1837
## Detection Rate         0.2845   0.1935   0.1743   0.1639   0.1837
## Detection Prevalence   0.2845   0.1935   0.1743   0.1639   0.1837
## Balanced Accuracy      1.0000   1.0000   1.0000   1.0000   1.0000
```

Conclusion

Based on this result, this model as a 100% accuracy.