AssignmentMachineLearning

WickyDonkey

Februari 1st 2019

Executive summary

This report contains the result of the Coursera peer-assignement of the Practical Machine Learning course. In this assignment the health exercise data is analysed to determine a model how well the persons investigated do their exercise.

The first section loads and cleans the training data of the reference exercices.

In the second section several models are created based on the training data to determine the quality of the excercises. The last model is a combined model with the result of the other models.

Based on the create test set the expected error level is estimated.

The third section estimates the expected quality of the "testing" set with 20 individual measurements, which are also part of the quiz.

Part 1 - Loading the excercise data and cleaning the data set

From the data description we get the following information about the "classe" (quality of the exercise):

(Class A) exactly according to the specification

(Class B) throwing the elbows to the front

(Class C) lifting the dumbbell only halfway

(Class D) lowering the dumbbell only halfway

(Class E) throwing the hips to the front

```
## [1] "X"                    "user_name"            "raw_timestamp_part_1"
## [4] "raw_timestamp_part_2" "cvtd_timestamp"       "new_window"
## [7] "num_window"
```

The summary shows that the first 7 columns are not relevant.

Further a lot of columns have a lot of NA values. These columns are removed from the training set.

Part 2 - Model the excercise quality

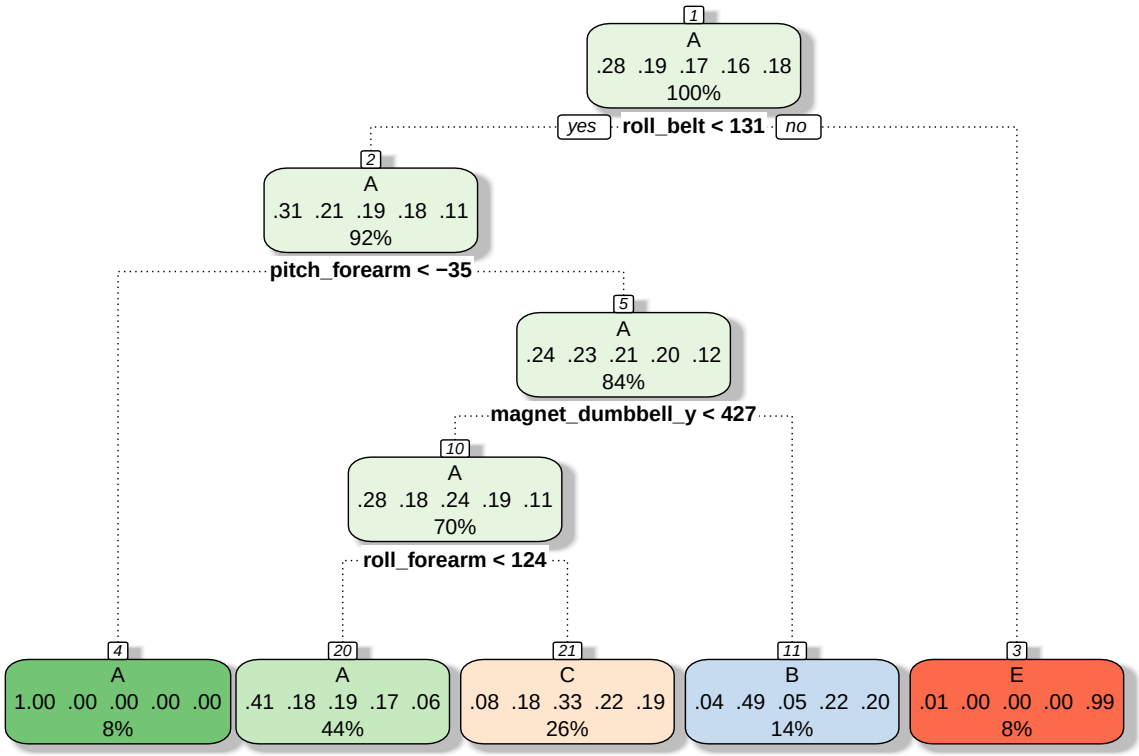
Several models are created based on the training data to determine the quality of the excercises.

To enable training and validation of the model the original training data is split into 75% training data and 25% validation data.

```
## [1] 14718    53
```

```
## [1] 4904   53
```

The first model is the standard regression tree (rpart)



Afterwards the other models are created: Bagging trees (bag), Random forest trees (rf), Boosting trees (gbm)

```
##
## Bagging classification trees with 25 bootstrap replications
```

```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 0.7%
## Confusion matrix:
##      A    B    C    D    E  class.error
## A 4182    3    0    0    0 0.0007168459
## B   15 2827    6    0    0 0.0073735955
## C    0   20 2544    3    0 0.0089598753
## D    0    0   46 2365    1 0.0194859038
## E    0    0    2    7 2697 0.0033259424
```

```
## A gradient boosted model with multinomial loss function.
## 150 iterations were performed.
## There were 52 predictors of which 52 had non-zero influence.
```

The last model is a combined model with the result of the other models. First the predicted values for each model are calculated, afterwards the predicted combined value is calculated.

Based on the create test set the expected error levels are estimated.

```
## [1] "rpart"
## [1] 0.5
## [1] "bag"
## [1] 0.9814437
## [1] "rf"
## [1] 0.99531
## [1] "gbm"
## [1] 0.9606444
## [1] "combined"
## [1] 0.9961256
```

Conclusion: the combined set gives the best result. This set is also used for the final quiz.

Part 3 - The quiz based on the "testing" set

A set of 20 excercises has been downloaded as part of this assignment. These excercises are run with the combined model

```
##  [1] 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
```

The results have been checked in the Coursera quiz and are correct