Predicting Correctness of Weight Lifting Exercises

Paolo Saracco

2024-06-08

Summary

We analyze the Weight Lifting Exercises Dataset, which gathers data from accelerometers on the belt, forearm, arm, and dumbell of 6 people performing barbell lifts correctly and incorrectly in 5 different ways. After fitting a predictive model for activity recognition based on these data, we use it to predict the correctness of the exercises performed by new users.

Packages

We will use the `dplyr` package for working with data frames, the `ggplot2` package for graphs, the `caret` package for training and predicting.

```
library(dplyr); library(ggplot2); library(caret); set.seed(0)
```

Background

Using devices such as Jawbone Up, Nike FuelBand, and Fitbit it is now possible to collect a large amount of data about personal activity relatively inexpensively. These type of devices are part of the quantified self movement - a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. One thing that people regularly do is quantify how much of a particular activity they do, but they rarely quantify how well they do it. In this project, your goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. More information is available from webpage of the project (see the section on the Weight Lifting Exercise Dataset).

Data

The data set originally comes from the Weight Lifting Exercises Dataset (see the section on the Weight Lifting Exercise Dataset on the webpage of the project - possibly see the web archive version of it at web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har). The training data for this project are also available at:

https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv

The test data are also available at:

https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

```
raw_training <- read.csv("data/train.csv")
raw_testing <- read.csv("data/test.csv")
```

Exploratory data analysis

We are dealing with a 19622 x 160 training data set and a 20 x 160 test data set. Let us inspect the training data and leave aside the test data set.

```
head(raw_training)
tail(raw_training)
str(raw_training)
summary(raw_training)
names(raw_training)
```

Data preprocessing

The first 7 columns contain data that are not useful for our purposes, hence we drop them. We also convert to numeric those columns that are erroneously interpreted as integers or characters. Then we drop those columns that are mostly (at least 90%) NAs. Finally, we check whether there are columns with almost zero variance, but there are not and we consider ourselves satisfied.

```
training <- raw_training %>%
        select(!c(X, user_name,
                  raw_timestamp_part_1,
                  raw_timestamp_part_2,
                  cvtd_timestamp,
```

```r
                  new_window,
                  num_window)) %>%
        mutate(across(!classe, as.numeric),
               classe = factor(classe)) %>%
        select(where(function(X) { mean(is.na(X)) < 0.9 }))
nzv <- nearZeroVar(training)
training <- select(training, !all_of(nzv))
dim(training)
```

```
## [1] 19622    53
```

Model selection

In order to select a model, we compare the accuracy of three models seen in the course: **decision tree**, **random forest** and **gradient-boosting trees** model.

We split our training data set into a data set for training the algorithms and a data set for validating the results and compare their accuracy.

```r
inTrain <- createDataPartition(y = training$classe,
                               p = 0.7,
                               list = FALSE)
trainSet <- training[inTrain,]
validationSet <- training[-inTrain,]
```

We also decide to perform **3-fold cross validation** in the training phase, to reduce the bias and improve the prediction capability of our models.

```r
control <- trainControl(method = "cv", number = 3)
```

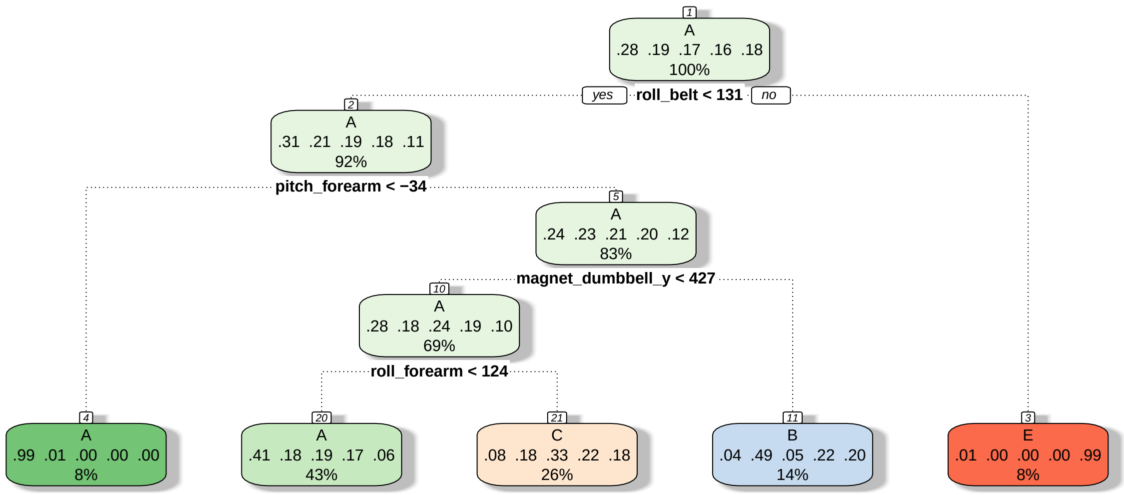
Decision tree

We start by training a single decision tree.

Creating the model

```r
Tree <- train(classe ~ ., data = trainSet,
              method = "rpart", trControl = control)

rattle::fancyRpartPlot(Tree$finalModel)
```



Rattle 2024−Jun−08 21:35:36 paolo

Check performance on validation data

```r
predTree <- predict(Tree, validationSet)
CnfMtxTree <- confusionMatrix(predTree, validationSet$classe)
CnfMtxTree
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1519  484  475  422  171
##          B   42  394   32  184  147
##          C  109  261  519  358  278
##          D    0    0    0    0    0
##          E    4    0    0    0  486
##
## Overall Statistics
##
##                Accuracy : 0.4958
##                  95% CI : (0.483, 0.5087)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3408
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9074  0.34592  0.50585   0.0000  0.44917
## Specificity            0.6314  0.91466  0.79296   1.0000  0.99917
## Pos Pred Value         0.4946  0.49312  0.34033      NaN  0.99184
## Neg Pred Value         0.9449  0.85352  0.88372   0.8362  0.88953
```

```
## Prevalence            0.2845  0.19354  0.17434   0.1638  0.18386
## Detection Rate        0.2581  0.06695  0.08819   0.0000  0.08258
## Detection Prevalence  0.5218  0.13577  0.25913   0.0000  0.08326
## Balanced Accuracy     0.7694  0.63029  0.64940   0.5000  0.72417
```

We conclude that this single decision tree has an accuracy of 0.4958369

Random forest

As second option, we train a random forest.

Creating the model

```
Forest <- train(classe ~ ., data = trainSet,
               method = "rf", trControl = control)
```

Check performance on validation data

```
predRF <- predict(Forest, validationSet)
CnfMtxRF <- confusionMatrix(predRF, validationSet$classe)
CnfMtxRF
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1671    4    0    0    2
##          B    2 1130    8    0    0
##          C    1    5 1012   15    0
##          D    0    0    6  948    4
##          E    0    0    0    1 1076
##
## Overall Statistics
##
##                Accuracy : 0.9918
##                  95% CI : (0.9892, 0.994)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9897
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9982   0.9921   0.9864   0.9834   0.9945
## Specificity            0.9986   0.9979   0.9957   0.9980   0.9998
## Pos Pred Value         0.9964   0.9912   0.9797   0.9896   0.9991
## Neg Pred Value         0.9993   0.9981   0.9971   0.9968   0.9988
## Prevalence             0.2845   0.1935   0.1743   0.1638   0.1839
## Detection Rate         0.2839   0.1920   0.1720   0.1611   0.1828
## Detection Prevalence   0.2850   0.1937   0.1755   0.1628   0.1830
## Balanced Accuracy      0.9984   0.9950   0.9910   0.9907   0.9971
```

We conclude that this random forest model has an accuracy of 0.9918437

Gradient-boosted Trees Model

Finally, we perform gradient boosting with decision trees as base learners.

Creating the model

```
GradBoostTrees <- train(classe ~ ., data = trainSet, method = "gbm",
                        verbose = FALSE, trControl = control)
```

Check performance on validation data

```
predGBM <- predict(GradBoostTrees, validationSet)
CnfMtxGBM <- confusionMatrix(predGBM, validationSet$classe)
CnfMtxGBM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1645   37    0    0    4
##          B   19 1070   35    4   14
##          C    8   30  980   31    7
##          D    2    2    8  919   14
##          E    0    0    3   10 1043
##
## Overall Statistics
##
##                Accuracy : 0.9613
##                  95% CI : (0.956, 0.966)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.951
##
##  Mcnemar's Test P-Value : 2.011e-07
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9827   0.9394   0.9552   0.9533   0.9640
## Specificity            0.9903   0.9848   0.9844   0.9947   0.9973
## Pos Pred Value         0.9757   0.9370   0.9280   0.9725   0.9877
## Neg Pred Value         0.9931   0.9855   0.9905   0.9909   0.9919
## Prevalence             0.2845   0.1935   0.1743   0.1638   0.1839
## Detection Rate         0.2795   0.1818   0.1665   0.1562   0.1772
## Detection Prevalence   0.2865   0.1941   0.1794   0.1606   0.1794
## Balanced Accuracy      0.9865   0.9621   0.9698   0.9740   0.9806
```

We conclude that this gradient-boosted trees model has an accuracy of 0.9612574

Sum up

Model	In sample accuracy	In sample error	Out of sample accuracy	Out of sample error
Decision Tree	0.4973	0.5027	0.4958	0.5042
Random Forest	1	0	0.9918	0.0082
Gradient-boosted Trees	0.9743	0.0257	0.9613	0.0387

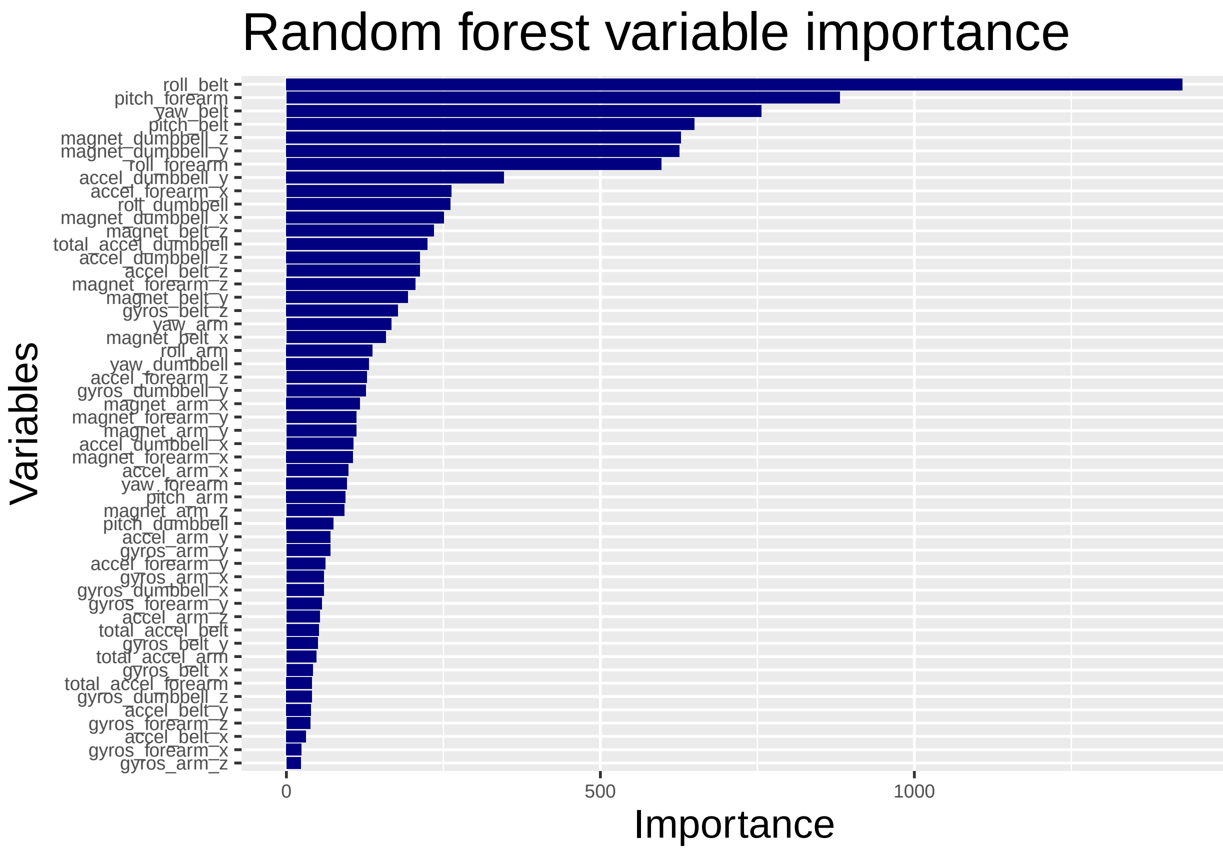
Therefore we decide to select the random forest model, since it performed pretty well on the testing data having a 99.18% out-of-sample accuracy rate.

Check variable importance

Since we opted for the random forest model, we can extract the variable importance from it. This tells us how the variables helped the model in predicting the class of the data. Higher importance means that the variable is useful to the model.

```
var_imp <- varImp(Forest, scale = F)$importance
var_imp <- data.frame(variables = row.names(var_imp),
                      importance = var_imp$Overall)
ggplot(aes(x = reorder(variables, importance), y = importance),
       data = var_imp) + geom_bar(stat = 'identity', fill = 'navy') +
        coord_flip() + labs(x = 'Variables',
                            y = 'Importance',
                            title='Random forest variable importance') +
```

```r
        theme(axis.text = element_text(size = 7),
              axis.title = element_text(size = 15),
              plot.title = element_text(size = 20))
```



Prediction on test set

We may finally apply our model on testing data.

```r
testing <- raw_testing %>%
        select(names(training[,-53]))
dim(testing)
```

```
## [1] 20 52
```

```r
TestRF <- predict(Forest, testing)
TestRF
```

```
##  [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
```