Prediction Assignment Writeup

Anton Votinov

22/08/2014

Clearing Data.

The original training dataset is not tidy at all. First of all, there is a variable called “new_window”, which takes two values: “yes” and “no”. Since number of “yes” values is much lower, than number of “no” values (406 and 19216) and obervations with “yes” are a mistery, rows with “yes” values are deleted.

There are a lot of variables with too many NAs and with no values at all (“”). We delete such variables (columns).

Finally, we delete “new_window”,“cvtd_timestamp” and “X” variables (columns) since they don’t give us any useful information.

```
library(lattice)
library(ggplot2)
library(caret)
train <- read.csv("pml-training.csv")
test <- read.csv("pml-testing.csv")
train1 <- train[train$new_window == "no",]
train1 <- train1[,colSums(is.na(train1)) < nrow(train1)]
train1 <- train1[, colSums(  train1 != ""  ) != 0]
train1 <- subset(train1, select = - c(new_window,cvtd_timestamp,X))
```

Fitting model

Random forest was choosen as a prediction model. Cross-validation is used with 5 resampling iterations.

```
trControl <- trainControl(method = "cv", number = 5, allowParallel = TRUE)
model <- train(classe ~ ., data = train1, method = "rf", trControl = trControl)
```

```
## Loading required package: randomForest
## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.
```

```
model
```

```
## Random Forest
##
## 19216 samples
##    56 predictor
##     5 classes: 'A', 'B', 'C', 'D', 'E'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
##
## Summary of sample sizes: 15373, 15373, 15374, 15373, 15371
##
## Resampling results across tuning parameters:
##
##   mtry  Accuracy  Kappa  Accuracy SD  Kappa SD
##    2    1         1      1e-03        2e-03
##   31    1         1      5e-04        7e-04
##   60    1         1      1e-03        1e-03
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 31.
```

```
accuracy <- max(model$results[,2])
```

The estimate of the out of sample error is thought to be accuracy of the model. The accuracy is about 0.9991 percent.