Machine Learning - Exercise Prediction

Troy Schmid

August 7, 2018

Introduction

This assignment consists of using an activity database that is already broken up into training and testing sets to attempt to predict which exercise is being performed in the testing trials according to the patterns learned from the training dataset.

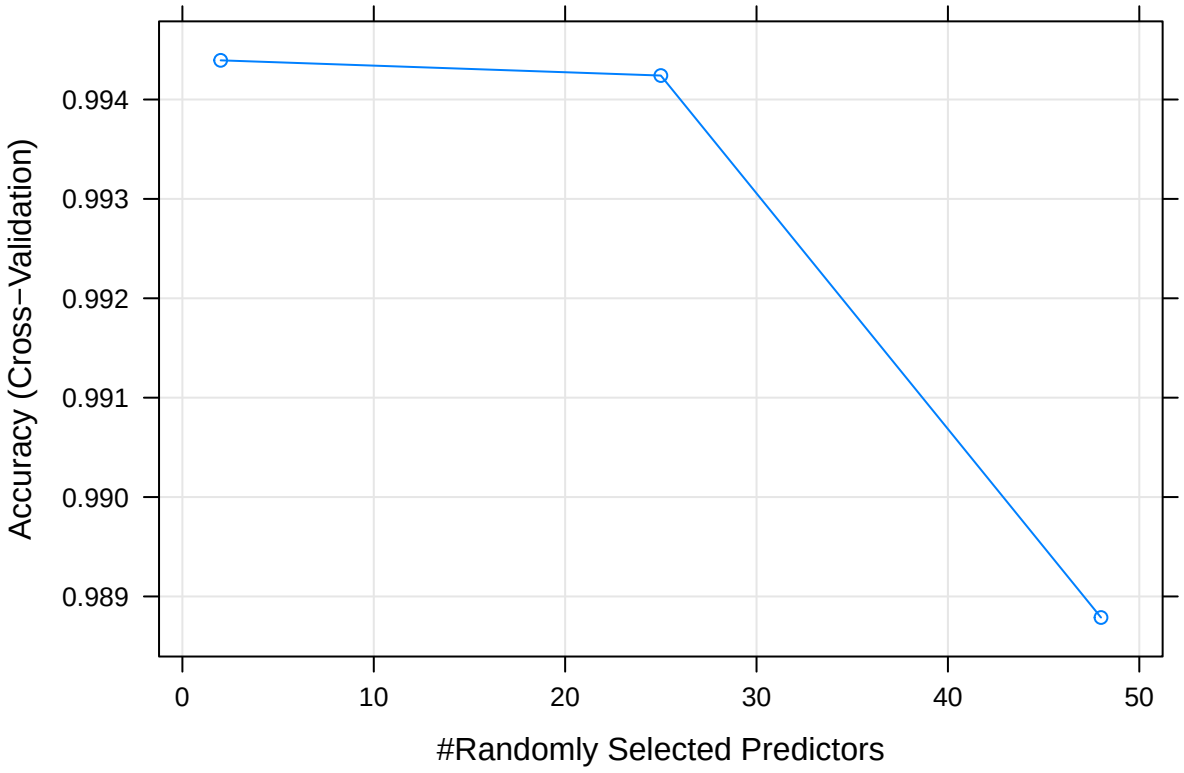
Exploratory Analysis

After having loaded both of the datasets into dataframes, I decided to remove many of the unnecessary and incomplete variables by removing those that were mostly NA values. Following that removal, I then removed the first 7 variables as I did not deem them necessary to predict future exercise data.

```
cluster <- makeCluster(detectCores() -1)
registerDoParallel(cluster)
training <- read.csv("pml-training.csv", header = TRUE)
testing <- read.csv("pml-testing.csv", header = TRUE)
training <- training[, which(colMeans(is.na(training)) == 0)] #Check which columns contain mostly empty
training <- training[, !grepl("kurtosis_|skewness_|avg_|var_|stddev_|max_|min_|amplitude_|total_", colna
training <- training[,-c(1:7)]
```

Next, as shown below, I test the predictors I've chosen and see how accurate my model will be. As read on the forums, a 99% out-of-sample estimated accuracy should enable me to safely predict and answer any of the test questions. I feel like the 5-fold cross validation model I have built is within the required accuracy and so I may proceed. I used the suggested cross validation and random forest techniques as was suggested in the forum. Culling some more variables might help the accuracy but it doesn't seem necessary at this point.

```
x <- training[,-ncol(training)]
y <- training[,ncol(training)]
set.seed(8675309)
fitControl <- trainControl(method="cv", number = 5, allowParallel = TRUE)
fit <- train(x,y, method="rf", data=training, trControl=fitControl, verbose=FALSE)
plot(fit)
```




```
stopCluster(cluster)
registerDoSEQ()
```

Conclusion

```
fit$finalModel
```

```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry, data = ..1, verbose = FALSE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 0.36%
## Confusion matrix:
##      A    B    C    D    E  class.error
## A 5578    2    0    0    0 0.0003584229
## B    9 3785    3    0    0 0.0031603898
## C    0   13 3406    3    0 0.0046756283
## D    0    0   33 3181    2 0.0108830846
## E    0    0    0    5 3602 0.0013861935
```

It appears that the training set is adequate enough to build an accurate model of the data using a 5-fold cross-validation. A 1% error rate being the highest among the activity classes makes me confident that the

model is satisfactory.