Predicting Curl Quality

Donald Hescht

11/29/2016

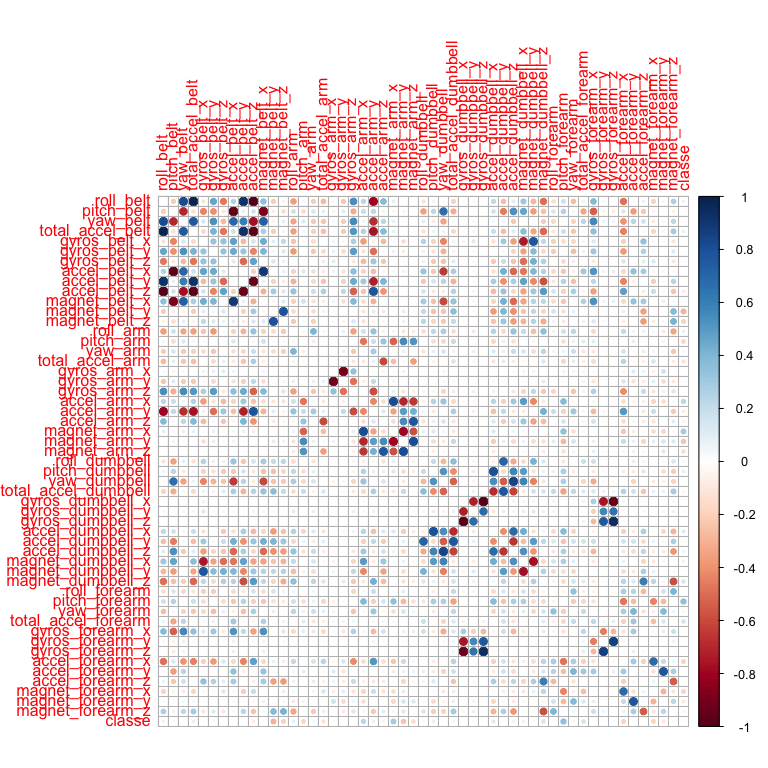
# Summary

This paper describes the author's approach to designing a 99% accurate model for predicting curl exercise type as defined in the paper "Qualitative Activity Recognition of Weight Lifting Exercises" [1]. This activity's data was captured by sampling 6 lifters with 4 attached sensors. These sensors were attached to the lifter's forearm, arm, belt and dumbbell. They were then monitored while performing 5 distinct curl exercises: one that was "correct" and four others with distinct errors. The aggregate of these sensor data and post calculations resulted in 19622 observations of 160 variables.

# Cleaning and Analysis

The lift data contains N/As and divide by zero errors. These values were post calculated as summary type data (max/mins and other stats) from the original sensors data. Therefore, actual sensor information was preserved after cleaning. After the cleaning two data sets were created: Training (14718, 54) and Testing (4904, 54).

The following Correlation Plot (with main diagonal removed) shows the correlation of these "cleaned" 53 features to the "classe"" output. Notice that there is no strong correlation with the "classe" out come. This indicates quite a few input variables will be required to create a 99% accurate model.



# Model Generation

The author used the Random Forest algorithm for two main reasons: 1) accuracy; 2) reduction in bias. (The reduction in bias is a result of it keeping out a portion of the training data to ensure good out of sample accuracy.)

The features for the Random Forest were found by ordering by their absolute correlation value. The author off-line picked the highest correlated features to reduce features from 53 to 37 while keeping 99% accuracy.

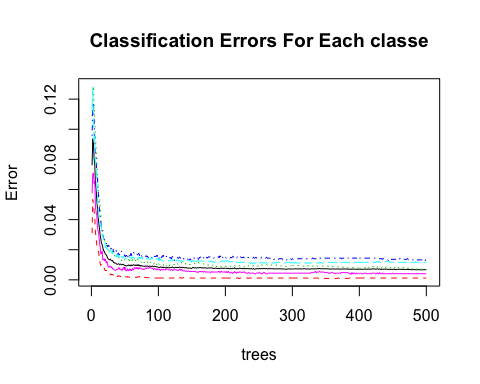
The final Random Forest **features =**

**{ pitch\_forearm, magnet\_belt\_y, magnet\_arm\_x, magnet\_arm\_y, accel\_arm\_x, magnet\_belt\_z, accel\_forearm\_x, magnet\_forearm\_x, pitch\_arm, total\_accel\_forearm, magnet\_dumbbell\_z, magnet\_arm\_z, total\_accel\_arm, accel\_dumbbell\_x, magnet\_forearm\_y, roll\_arm, accel\_belt\_z, accel\_arm\_y, total\_accel\_belt, pitch\_dumbbell, accel\_dumbbell\_z, magnet\_dumbbell\_x, roll\_belt, accel\_arm\_z, total\_accel\_dumbbell, yaw\_forearm, yaw\_arm, roll\_dumbbell, magnet\_forearm\_z, gyros\_dumbbell\_y, accel\_forearm\_y, roll\_forearm, magnet\_belt\_x, gyros\_arm\_y, gyros\_belt\_y, accel\_dumbbell\_y, yaw\_belt}**

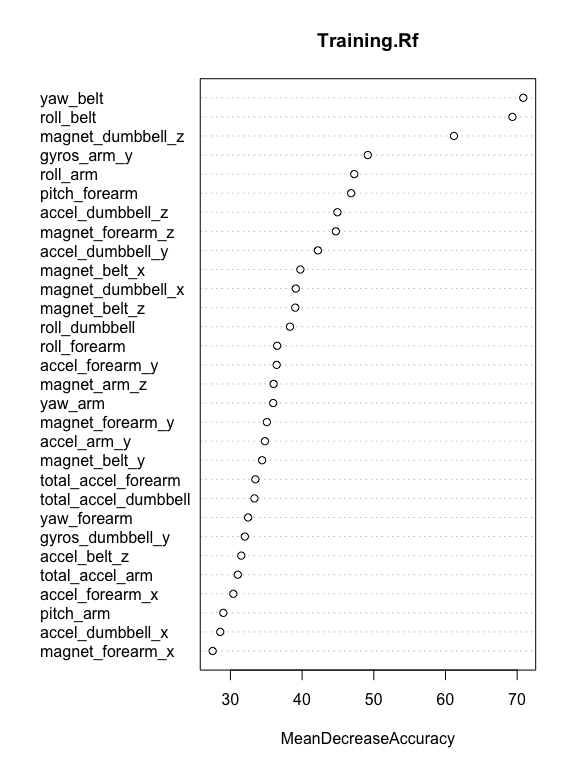
set.seed(19652)  
Training.Rf <- randomForest(Formula, data=Training, mtry=as.integer(sqrt(length(Features))), importance=TRUE)  
Training.Rf

##   
## Call:  
## randomForest(formula = Formula, data = Training, mtry = as.integer(sqrt(length(Features))), importance = TRUE)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 6  
##   
## OOB estimate of error rate: 0.67%  
## Confusion matrix:  
## A B C D E class.error  
## A 4180 4 1 0 0 0.001194743  
## B 13 2829 6 0 0 0.006671348  
## C 0 24 2532 11 0 0.013634593  
## D 1 0 24 2384 3 0.011608624  
## E 0 0 2 9 2695 0.004065041

The final model error is shown in this plot indicating that all classe values (A-E) are running around an 99% accuracy. A later project might consider the effect of reducing the number of trees.



The importance of features to accuracy is shown below. Removing even the least "important" features caused the model to approach or cross the required accuracy of 99%. Likewise, though there is a strong cross correlation between the Euler Angles and the raw values, removing them quickly caused the model accuracy to reduce below 99% accuracy. Therefore, the author kept the correlation ordered 37 features.



# Validation

# Function to report accuracy of model  
set.seed(19653)  
Training.Test.Pred <- predict(Training.Rf, Testing)  
ctable <- table(Training.Test.Pred, Testing$classe)   
Testing.Acc <- sum(diag(ctable)) / sum(ctable)  
Training.OOB.Acc <- sum(diag(Training.Rf$confusion[,1:5])) / sum(Training.Rf$confusion[,1:5])

The Random Forest model was validated by using a 25% hold out of the "pml-training.csv" training data. Using this hold out data gives an out of sample accuracy of **0.9918434**. This closely matches the accuracy given by the Random Forest Training Confusion Matrix **(0.9933415)** which is ~(1 - OOB).

# Test Data

Running the 20 test cases.

set.seed(19654)  
testData <- read.csv("pml-testing.csv", na.strings=c("#DIV/0", '', 'NA') ,stringsAsFactors = F)  
predict(Training.Rf, testData)

## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20   
## 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

# Citations

[1] Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) . Stuttgart, Germany: ACM SIGCHI, 2013. Read more: <http://groupware.les.inf.puc-rio.br/har#ixzz4QYU9sHrZ>

[2] <https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv>