### COURSE PROJECT: PRACTICAL MACHINE LEARNING

#### Data Cleaning and Preparation

train\_in <- read.csv('./pml-training.csv', header=T)

validation <- read.csv('./pml-testing.csv', header=T)

#### Data Partitioning

set.seed(127)

training\_sample <- createDataPartition(y=train\_in$classe, p=0.7, list=FALSE)

training <- train\_in[training\_sample, ]

testing <- train\_in[-training\_sample, ]

#### Identification on Non-Zero Data

Remove NearZeroVariance variables

all\_zero\_colnames <- sapply(names(validation), function(x) all(is.na(validation[,x])==TRUE))

nznames <- names(all\_zero\_colnames)[all\_zero\_colnames==FALSE]

nznames <- nznames[-(1:7)]

nznames <- nznames[1:(length(nznames)-1)]

The models will be fit using the following data columns:

## [1] "accel\_arm\_x" "accel\_arm\_y" "accel\_arm\_z"

## [4] "accel\_belt\_x" "accel\_belt\_y" "accel\_belt\_z"

## [7] "accel\_dumbbell\_x" "accel\_dumbbell\_y" "accel\_dumbbell\_z"

## [10] "accel\_forearm\_x" "accel\_forearm\_y" "accel\_forearm\_z"

## [13] "gyros\_arm\_x" "gyros\_arm\_y" "gyros\_arm\_z"

## [16] "gyros\_belt\_x" "gyros\_belt\_y" "gyros\_belt\_z"

## [19] "gyros\_dumbbell\_x" "gyros\_dumbbell\_y" "gyros\_dumbbell\_z"

## [22] "gyros\_forearm\_x" "gyros\_forearm\_y" "gyros\_forearm\_z"

## [25] "magnet\_arm\_x" "magnet\_arm\_y" "magnet\_arm\_z"

## [28] "magnet\_belt\_x" "magnet\_belt\_y" "magnet\_belt\_z"

## [31] "magnet\_dumbbell\_x" "magnet\_dumbbell\_y" "magnet\_dumbbell\_z"

## [34] "magnet\_forearm\_x" "magnet\_forearm\_y" "magnet\_forearm\_z"

## [37] "pitch\_arm" "pitch\_belt" "pitch\_dumbbell"

## [40] "pitch\_forearm" "roll\_arm" "roll\_belt"

## [43] "roll\_dumbbell" "roll\_forearm" "total\_accel\_arm"

## [46] "total\_accel\_belt" "total\_accel\_dumbbell" "total\_accel\_forearm"

## [49] "yaw\_arm" "yaw\_belt" "yaw\_dumbbell"

## [52] "yaw\_forearm"

## Model building

The three model types used are:

1. Decision trees with CART (rpart)
2. Stochastic gradient boosting trees (gbm)
3. Random forest decision trees (rf)

model\_cart <- train(

classe ~ .,

data=training[, c('classe', nznames)],

trControl=fitControl,

method='rpart'

)

save(model\_cart, file='./ModelFitCART.RData')

model\_gbm <- train(

classe ~ .,

data=training[, c('classe', nznames)],

trControl=fitControl,

method='gbm'

)

save(model\_gbm, file='./ModelFitGBM.RData')

model\_rf <- train(

classe ~ .,

data=training[, c('classe', nznames)],

trControl=fitControl,

method='rf',

ntree=100

)

save(model\_rf, file='./ModelFitRF.RData')

### Cross validation

Cross validation is done for each model with K = 3.

fitControl <- trainControl(method='cv', number = 3)

### Model Assessment (Out of sample error)

predCART <- predict(model\_cart, newdata=testing)

cmCART <- confusionMatrix(predCART, testing$classe)

predGBM <- predict(model\_gbm, newdata=testing)

cmGBM <- confusionMatrix(predGBM, testing$classe)

predRF <- predict(model\_rf, newdata=testing)

cmRF <- confusionMatrix(predRF, testing$classe)

AccuracyResults <- data.frame(

Model = c('CART', 'GBM', 'RF'),

Accuracy = rbind(cmCART$overall[1], cmGBM$overall[1], cmRF$overall[1])

)

print(AccuracyResults)

## Model Accuracy

## 1 CART 0.4932880

## 2 GBM 0.9622770

## 3 RF 0.9926933

Based on an assessment of these 3 model fits and out-of-sample results, it looks like both gradient boosting and random forests outperform the CART model, with random forests being slightly more accurate. The confusion matrix for the random forest model is below.

## Reference

## Prediction A B C D E

## A 1671 9 0 0 0

## B 3 1126 4 4 2

## C 0 4 1020 6 1

## D 0 0 2 952 6

## E 0 0 0 2 1073