Practical Machine Learning Course Project: Prediction Assignment

Emmanuel Abbey

August 5, 2018

Synopsis

The goal of this project is to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants to predict the manner in which they did the exercise. Subject were asked to perform barbell lifts correctly and incorrectly in 5 different ways. More information is available from the website here: http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset).

Data Cleaning Preparation

Download data from source.

```
## Set varibles for raw data download fron source

url_TrainingData <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
fileDest_TrainingData <- "pml-training.csv"

url_TestingData <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
fileDest_TestingData <- "pml-testing.csv"

## Download training data
download.file(url=url_TrainingData, destfile=fileDest_TrainingData, method="curl")

## Download testing data
download.file(url=url_TestingData, destfile=fileDest_TestingData, method="curl")
```

Load Training data and split into Training and Testing

```
TrainData <- read.csv('pml-training.csv', header=T)

## set seed
set.seed(1569)
## partition data 70% / 30%
```

```
TrainData_Part <- createDataPartition(y=TrainData$classe, p=0.7, list=FALSE)
## set training and testing data from partition
Training <- TrainData[TrainData_Part, ]
Testing <- TrainData[-TrainData_Part, ]

dim(Training)

## [1] 13737   160

dim(Testing)

## [1] 5885  160
```

Cleaning data

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

## Identify non-zero features
ZeroCol_names <- sapply(names(validation),
                        function(x) all(is.na(validation[,x]) == TRUE))

NonZeroCol_names <- names(ZeroCol_names)[ZeroCol_names == FALSE]
NonZeroCol_names <- NonZeroCol_names[-(1:7)]
NonZeroCol_names <- NonZeroCol_names[1:(length(NonZeroCol_names)-1)]
```

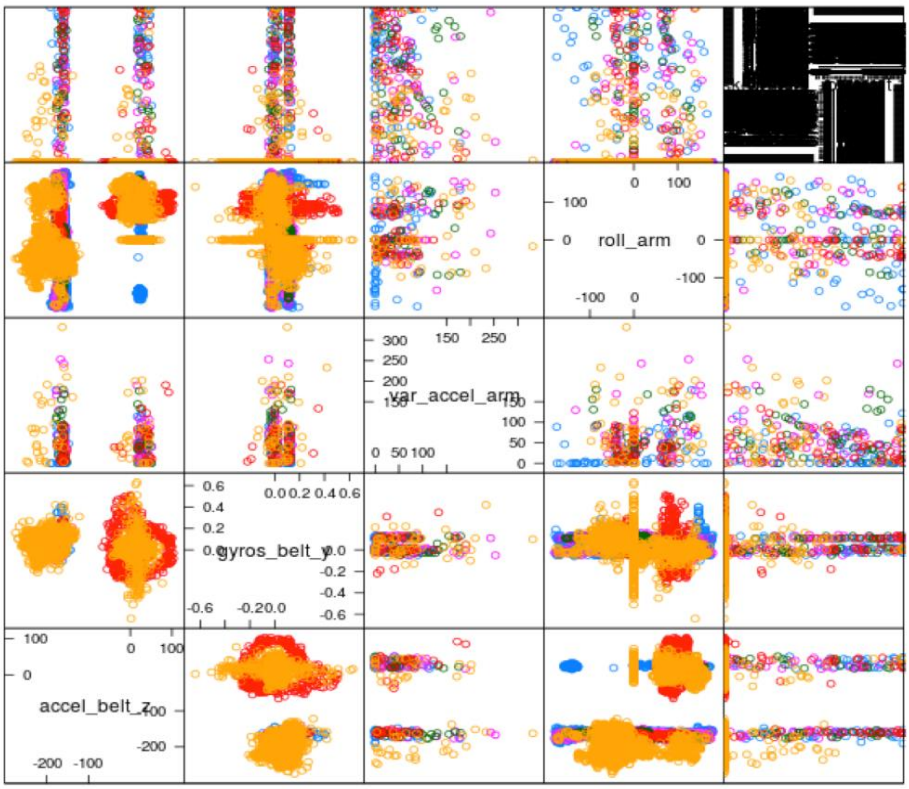
Build Model

Will try 3 models: Stochastic gradient boosting trees, Random forest decision trees and Decision trees with CART to see which provides the best acuracy.

```
## Cross validate

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

## Stochastic gradient boosting trees
Model_gbm <- caret::train(Training$classe ~ ., data=Training[, c('classe',
NonZeroCol_names)], trControl=fitControl, method='gbm')
save(Model_gbm, file='ModelGBM.RData')

## Random forest decision trees
Model_rf <- caret::train(Training$classe ~ ., data=Training[, c('classe',
NonZeroCol_names)], trControl=fitControl, ntree=110, method='rf')
save(Model_rf, file='ModelRF.RData')


## Decision trees with CART
Model_cart <- caret::train(Training$classe ~ ., data=Training[, c('classe',
NonZeroCol_names)], trControl=fitControl, method='rpart')
save(model_cart, file='ModelCART.RData')
```

Evaluation

```
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)
```



Scatter Plot Matrix

Prediction

Prediction using the Testing data downloaded from soure.

```
Predict_Validation <- predict(champion_model, newdata=validation)
ValidationPredictions <- data.frame(problem_id = validation$problem_id,
predicted = Predict_Validation)

print(ValidationPredictions)
```

```
##    problem_id predicted
## 1           1         B
## 2           2         A
## 3           3         B
## 4           4         A
## 5           5         A
## 6           6         E
## 7           7         D
## 8           8         B
## 9           9         A
## 10         10         A
## 11         11         B
## 12         12         C
## 13         13         B
## 14         14         A
## 15         15         E
## 16         16         E
## 17         17         A
## 18         18         B
## 19         19         B
## 20         20         B
```

Conclusion

I am able to fit a reasonably sound model with a high degree of accuracy in predicting out of sample observations. The random forest model with cross-validation produces a surprisingly accurate model that is sufficient for predictive analytics.