Regression Models Course Project

Peer Review Portion

Random Forest

```
model1 <- randomForest(classe ~. , data=TrainTrainingSet, method="class")
prediction1 <- predict(model1, TestTrainingSet, type = "class")
confusionMatrix(prediction1, TestTrainingSet$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1395    4    0    0    0
##          B    0  944    6    0    0
##          C    0    1  848    8    0
##          D    0    0    1  795    5
##          E    0    0    0    1  896
##
## Overall Statistics
##
##                Accuracy : 0.9947
##                  95% CI : (0.9922, 0.9965)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9933
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            1.0000   0.9947   0.9918   0.9888   0.9945
## Specificity            0.9989   0.9985   0.9978   0.9985   0.9998
## Pos Pred Value         0.9971   0.9937   0.9895   0.9925   0.9989
## Neg Pred Value         1.0000   0.9987   0.9983   0.9978   0.9988
## Prevalence             0.2845   0.1935   0.1743   0.1639   0.1837
## Detection Rate         0.2845   0.1925   0.1729   0.1621   0.1827
## Detection Prevalence   0.2853   0.1937   0.1748   0.1633   0.1829
## Balanced Accuracy      0.9994   0.9966   0.9948   0.9937   0.9971
```
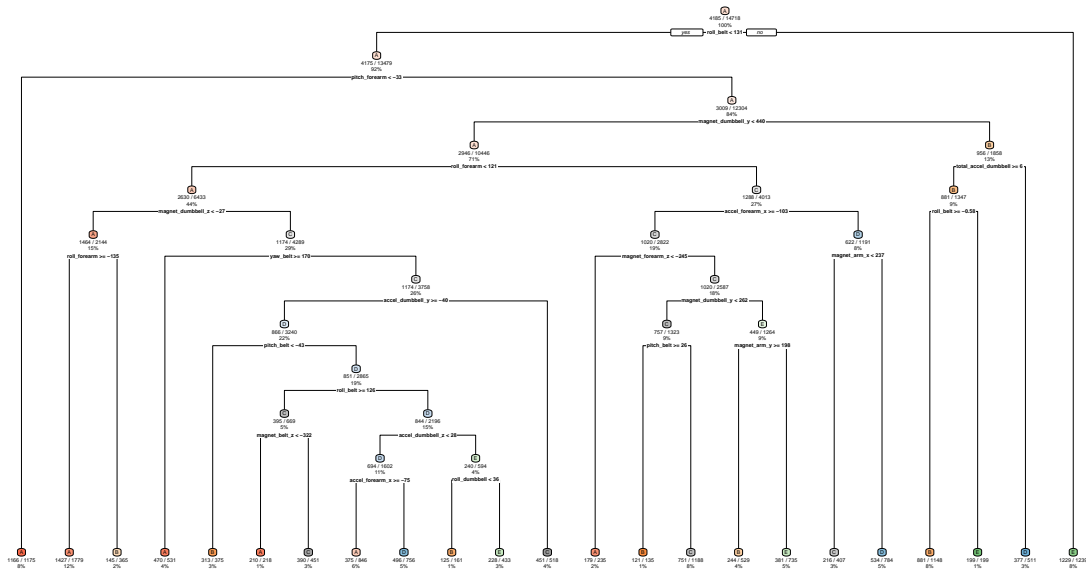
Decision Tree

```
model2<- rpart(classe ~ ., data=TrainTrainingSet, method="class")
prediction2 <- predict(model2, TestTrainingSet, type = "class")
rpart.plot(model2, main="Classification Tree", extra=102, under=TRUE, faclen=0)
```

Classification Tree

The Test results on our TestTrainingSet data set is as follows:

```
confusionMatrix(prediction2, TestTrainingSet$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1273  154   44   74   16
##          B   47  608   95   75   80
##          C   30   72  622   64   75
##          D   35   78   66  510   62
##          E   10   37   28   81  668
##
## Overall Statistics
##
##                Accuracy : 0.7506
##                  95% CI : (0.7383, 0.7627)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6831
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                     Class: A Class: B Class: C Class: D Class: E
## Sensitivity           0.9125   0.6407   0.7275   0.6343   0.7414
## Specificity           0.9179   0.9249   0.9405   0.9412   0.9610
```

```
## Pos Pred Value          0.8155   0.6718   0.7207   0.6791   0.8107
## Neg Pred Value          0.9635   0.9147   0.9423   0.9292   0.9429
## Prevalence              0.2845   0.1935   0.1743   0.1639   0.1837
## Detection Rate          0.2596   0.1240   0.1268   0.1040   0.1362
## Detection Prevalence    0.3183   0.1845   0.1760   0.1531   0.1680
## Balanced Accuracy       0.9152   0.7828   0.8340   0.7878   0.8512
```

Model selection

We could notice that Random Forest is better than Decision Trees. Accuracy for Random Forest model was 0.995 (95% CI: (0.993, 0.997)) compared to Decision Tree model with 0.739 (95% CI: (0.727, 0.752)). The Random Forests model is choosen. The expected out-of-sample error is estimated at 0.005, or 0.5%.

```
predictfinal <- predict(model1, testingset, type="class")
predictfinal
```

```
##  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
```