Patricia Walsh
Coursera practical machine learning
Course project

Overview

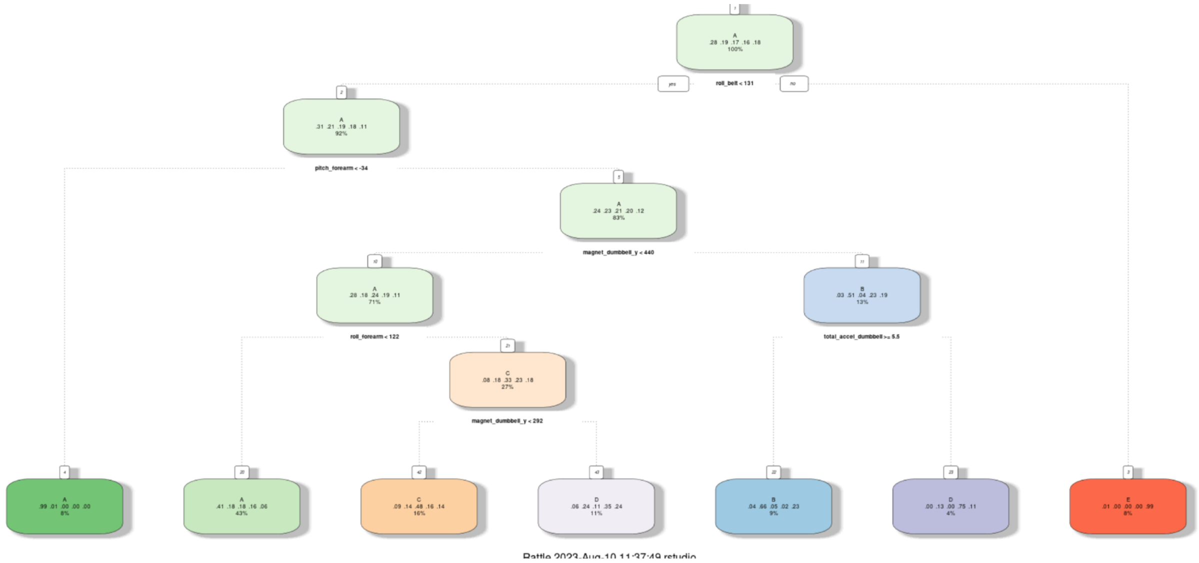
Using the test and training sets found here working to identify what model is best in terms of accuracy for the fitness data. The models to be compared are decision trees, random forest, and gradient boost.

Data

The training data for this project are available here:
https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv
The test data are available here:
https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

Step 1:
Loading the libraries including ggplot, rattle, kernlab, lattice, caret, coreplot.
Step 2
Set the seed to a value, I used 1234.
Step 3
Clean the data: I removed any irrelevant metadata and NA values.
Step 4
Creating the testing models.

```
mod_trees <- train(classe~., data=train, method="rpart", trControl = control, tuneLength = 5)
fancyRpartPlot(mod_trees$finalModel)
```



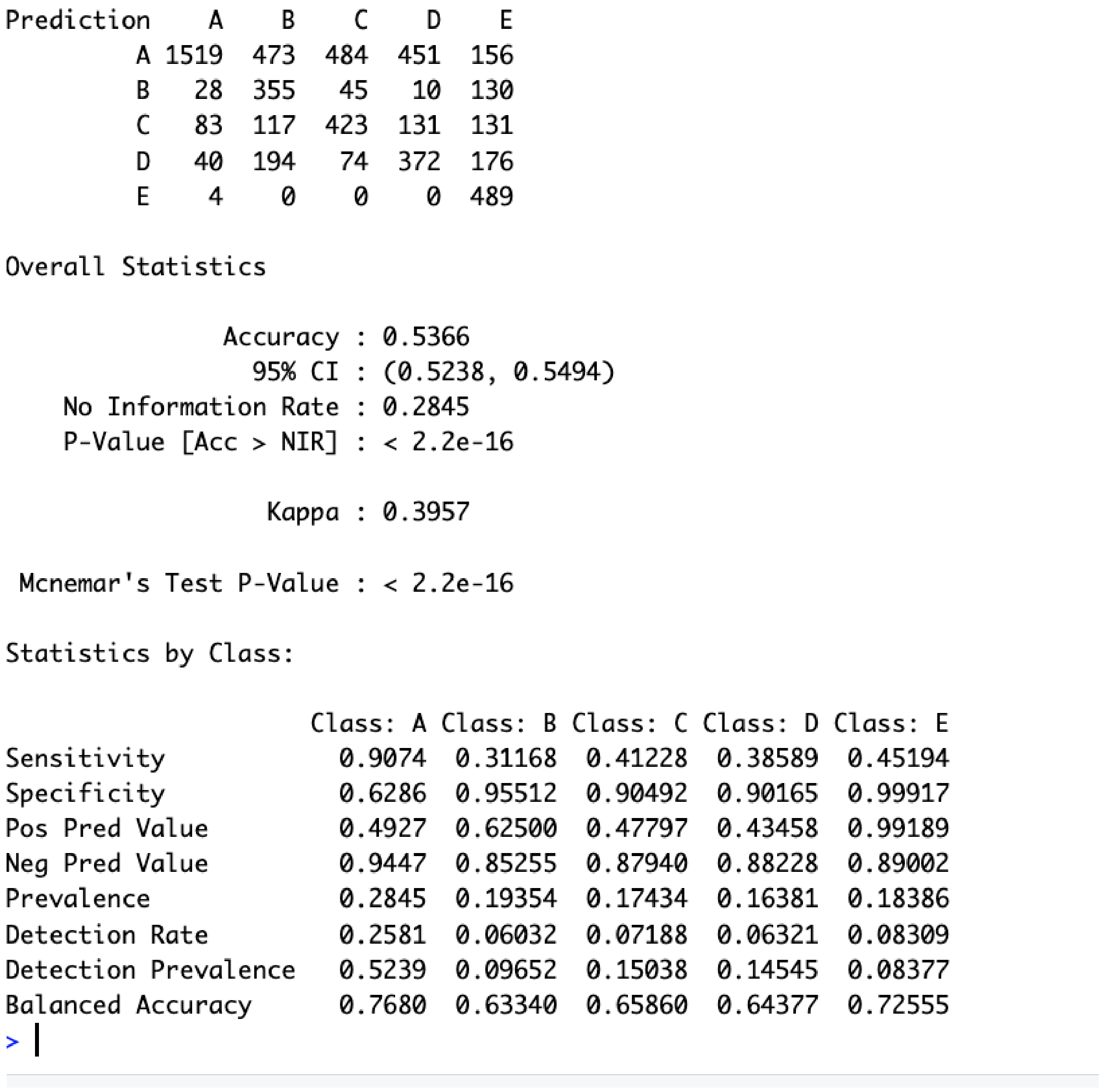
Random forest

```
mod_rf <- train(classe~., data=train, method="rf", trControl = control, tuneLength = 5)
```

```
mod_gbm <- train(classe~., data=train, method="gbm", trControl = control,
tuneLength = 5, verbose = F)
```

Comparing the confusion matrix for all three

For example



```
Prediction    A    B    C    D    E
         A 1519  473  484  451  156
         B   28  355   45   10  130
         C   83  117  423  131  131
         D   40  194   74  372  176
         E    4    0    0    0  489

Overall Statistics

               Accuracy : 0.5366
                 95% CI : (0.5238, 0.5494)
    No Information Rate : 0.2845
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3957

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: A Class: B Class: C Class: D Class: E
Sensitivity            0.9074  0.31168  0.41228  0.38589  0.45194
Specificity            0.6286  0.95512  0.90492  0.90165  0.99917
Pos Pred Value         0.4927  0.62500  0.47797  0.43458  0.99189
Neg Pred Value         0.9447  0.85255  0.87940  0.88228  0.89002
Prevalence             0.2845  0.19354  0.17434  0.16381  0.18386
Detection Rate         0.2581  0.06032  0.07188  0.06321  0.08309
Detection Prevalence   0.5239  0.09652  0.15038  0.14545  0.08377
Balanced Accuracy      0.7680  0.63340  0.65860  0.64377  0.72555
>
```

The result was that the best model for prediction of the fitness data is actually the random forest.