Practical Machine Learning – Course Assignment Write Up

Sameer Parab

1) What is the question we are asking from the data?

Our aim is to build a model that predicts correct or incorrect way of lifting weights based on accelerometer measurement data. The training data is from the HAR paper referenced below. The response/outcome variable is a categorical variable with 5 classes (A= correct way, while B,C,D,E are incorrect ways of various types). The feature variables are a series of accelerometer measurements as described in detail on the website [1, 2].

2) Data inspection and pre-processing/munging and Feature Selection

Conceptually, the accelerometer data should measure any given person's biases in weight lifting. As such, the initial descriptor variables for a general model should be redundant since the same participants may not participate in the future. Hence, descriptor variables such as user_name, raw_timestamp_part_1, part 3, cvtd timestamp, new_window, num_window should be removed from the training data. In addition, there are avg_ , stdev_ and var_ readings for various measurements in the training dataset. These can be dropped as well.

A brief look at the raw data on a sample of measurements indicates that the data is quite noisy.

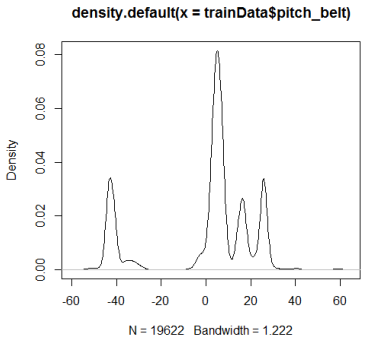
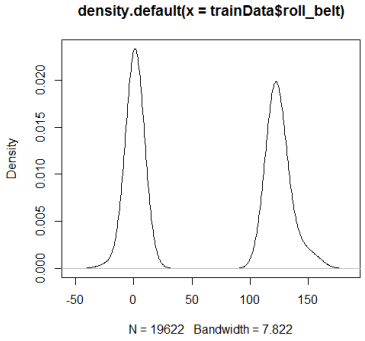


Figure 1: Density plots of raw feature variables

3) Model Selection and Fit Statistics:

Given the noisy nature of the accelerometer data, an ensemble method using random forest algorithm is recommended and tested. I have used the K=10 fold CV cross-validation in the caret package with method="rf".

The model fit statistics are as shown below:



a) Random Forest Algorithm:

```
> modelrf
Random Forest
```

```
14718 samples
   52 predictor
    5 classes: 'A', 'B', 'C', 'D', 'E'

No pre-processing
Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 13247, 13245, 13245, 13246, 13247, 13246, ...

Resampling results across tuning parameters:

  mtry  Accuracy  Kappa  Accuracy SD  Kappa SD
   2    0.993     0.991  0.00251      0.00318
  27    0.993     0.991  0.00241      0.00304
  52    0.986     0.982  0.00336      0.00426

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was mtry = 27.
```

Model Prediction Accuracy:

The comparison of the prediction accuracy is shown below. The confusion matrix and other model prediction parameters such as sensitivity, specifity on the training data set show that the random forest algorithm is fairly accurate.

Confusion matrix for Random forest algorithm:

```
  > confusionMatrix(predrf,testDatae$classe)
Confusion Matrix and Statistics

          Reference
Prediction    A    B    C    D    E
         A 1391    5    0    0    0
         B    1  942    2    1    0
         C    2    2  851    7    0
         D    0    0    2  795    3
         E    1    0    0    1  898

Overall Statistics

               Accuracy : 0.9945
                 95% CI : (0.992, 0.9964)
    No Information Rate : 0.2845
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.993
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: A Class: B Class: C Class: D Class: E
Sensitivity            0.9971   0.9926   0.9953   0.9888   0.9967
Specificity            0.9986   0.9990   0.9973   0.9988   0.9995
Pos Pred Value         0.9964   0.9958   0.9872   0.9938   0.9978
Neg Pred Value         0.9989   0.9982   0.9990   0.9978   0.9993
Prevalence             0.2845   0.1935   0.1743   0.1639   0.1837
Detection Rate         0.2836   0.1921   0.1735   0.1621   0.1831
Detection Prevalence   0.2847   0.1929   0.1758   0.1631   0.1835
Balanced Accuracy      0.9979   0.9958   0.9963   0.9938   0.9981
```

Figure 2 below shows the order of variable importance in the

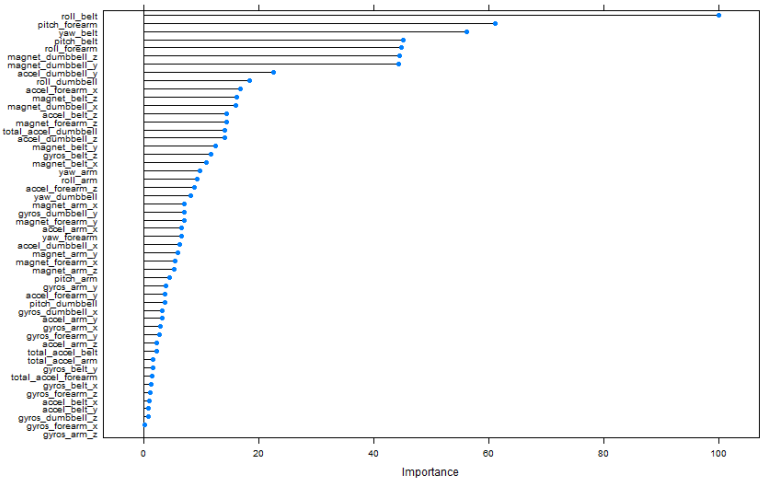


Figure 2: Variable importance for Random forest algorithm.

4) Expected out of sample error

The expected out of sample error for the random forest algorithm is (1-accuracy)% from the test data set confusion matrix = (1-0. 9945) = 0.55%

References:

1) http://groupware.les.inf.puc-rio.br/har

2) 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