HAR Analysis of Weight Lifting Exercises: a Random Forests Predictive Model

Justin Papreck

June 16, 2018

Introduction

Human Activity Recognition, HAR, has provided an abundance of information available to data scientists interested in questions related data from simple to complex human activity patterns and behaviors. In this study, the data were collected from individuals doing 10 sets of dumbell exercises in 5 different ways. The first class of data was the appropriate method for lifting the dumbell, but each other class of data follows a specific pattern to which the dumbell was lifted inappropriately. By accounting the different parameters from the data collected, we will attempt to establish a model to predict how the user is lifting the weight, which could have an impact in delivering immediate results to let a potential weight lifter know if they are appropriately lifting, or how they need to correct their movement.

Pre-Processing

Upon the exploratory analysis of the data, it was evident that much of the collected data were incomplete. Furthermore, the class of data in each of the fields was different: a mix of integer, numerical, and factor variables. To clean the data, all of the columns without collected data were removed, as well as were the data such as timestamps for collected data, the names of the users, and the numerical values for the collection. After removing these columns, the data were split into two groups for training and validation of the predictive model.

Validation

The data was broken into two subsets, one for training, and one for validation. To establish a validation model, 30% of the training data were separated by 'Classe' to create a Validation Test data set. These data were put aside until the testing of the predictive model.

Training Model

In this analysis, random forests were used to train and create a predictive model. Prior to the use of random forests, recursive partitioning was also used to establish a model with cross-validation. After the initial runs through with recursive partitioning, the validation had a low accuracy, and therefore opting for the more computationally demanding random forest was a better option. The use random forests typically yields the best output in machine learning models. To assess the accuracy of the predictive model, the training set was compared with the validation data and passed through a confusion matrix.

```
##
## Call:
##  randomForest(formula = classe ~ ., data = TrainingData, ntree = 250,      allowParallel = TRUE)
##                Type of random forest: classification
##                      Number of trees: 250
## No. of variables tried at each split: 7
##
##         OOB estimate of  error rate: 0.29%
## Confusion matrix:
##      A    B    C    D    E  class.error
## A 3905    1    0    0    0 0.0002560164
## B    6 2651    1    0    0 0.0026335591
## C    0   11 2383    2    0 0.0054257095
## D    0    0   15 2236    1 0.0071047957
## E    0    0    0    3 2522 0.0011881188
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1674    3    0    0    0
##          B    0 1136    1    0    0
##          C    0    0 1025    5    0
##          D    0    0    0  959    0
##          E    0    0    0    0 1082
##
## Overall Statistics
##
##                Accuracy : 0.9985
##                  95% CI : (0.9971, 0.9993)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9981
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            1.0000   0.9974   0.9990   0.9948   1.0000
## Specificity            0.9993   0.9998   0.9990   1.0000   1.0000
## Pos Pred Value         0.9982   0.9991   0.9951   1.0000   1.0000
## Neg Pred Value         1.0000   0.9994   0.9998   0.9990   1.0000
## Prevalence             0.2845   0.1935   0.1743   0.1638   0.1839
## Detection Rate         0.2845   0.1930   0.1742   0.1630   0.1839
## Detection Prevalence   0.2850   0.1932   0.1750   0.1630   0.1839
## Balanced Accuracy      0.9996   0.9986   0.9990   0.9974   1.0000
```

```
##  Accuracy
## 0.9984707
```

Results

The estimated rate of error in the training model was 0.29%. When run through the validation testing, the accuracy of the confusion matrix was 99.85% accurate with a 95% confidence interval of .9971 to .9993. The p-value for this analysis was on the order of 10^-16, and thus a significant finding. This just means that the predictive ability of this model is very accurate according to the validation data provided, but this does not comment on any potential over-fitting.

```
##  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
```

Prediction

Finally, the model was used to make predictions from the set of test data from Uh, et al.

Acknowledgements

This model would not have been possible without the generous allowance for Coursera's Data Science Specialization through Johns Hopkins University by Velloso, et al. and Groupware@LES.

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.