Machine Learning Exercise for Barbell Data Set

Robert Wiederstein

October 13, 2016

Executive Summary

Six participants were asked to wear accelerometers and perform barbell lifts correctly and incorrectly in five different ways. The variable “classe” was the outcome variable of interest. A random forest model fit the data the best with an accuracy of 99.5% and, when applied to the test set, predicted with 99.67% accuracy the likelihood of a lift being identified correctly. Interestingly, there was 100% accuracy in identifying when a lift was done correctly, option “A” in the classe variable.

Project Description

As background, the data originated from devices that were personally worn to measure activity. These devices such as Jawbone Up, Nike FuelBand, and Fitbit collect a large amount of data about personal activity inexpensively. This technology is part of the Human Activity Recognition (“HAR”) movement and could potentially improve care for the elderly, incentivize people to exercise, and monitor weight loss and other health indicators.

The Coursera website states that “[t]hese type of devices are part of the quantified self movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. One thing that people regularly do is quantify how much of a particular activity they do, but they rarely quantify how well they do it.”

This project’s data comes from accelerometers on the belt, forearm, arm, and dumbell of 6 participants using a 1.25 kg weight. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. The project is significant because previous research on accelerometers sought to identify the type of activity whereas in the current project, the goal is to identify whether the participant is doing it correctly. Class “A” identified the exercise being done correctly while “B,” “C,” “D,” and “E” identified an incorrect exercise. Thus, the project is a classification problem.

Loading and Cleaning the Data

The data were read into R and the testing and training sets were combined so that they could be uniformly processed. The original and combined set of data contained 19642 observations on 161 variables. Many of the variables were empty and, therefore, eliminated from the dataframe.

```
cols.na <- apply(bb, 2, function(x)(sum(is.na(x))))
bb <- bb[, which(cols.na == 0) ]
rows.na <- apply(bb, 1, function(x)(sum(is.na(x))))
bb <- bb[which(rows.na == 0), ]
```

Additionally, the first seven columns were used for purposes of identification. After reviewing the data, it was determined there was little potential for feature engineering and they were discarded.

```
cols.id <- 1:7
bb <- bb[, -cols.id]
```

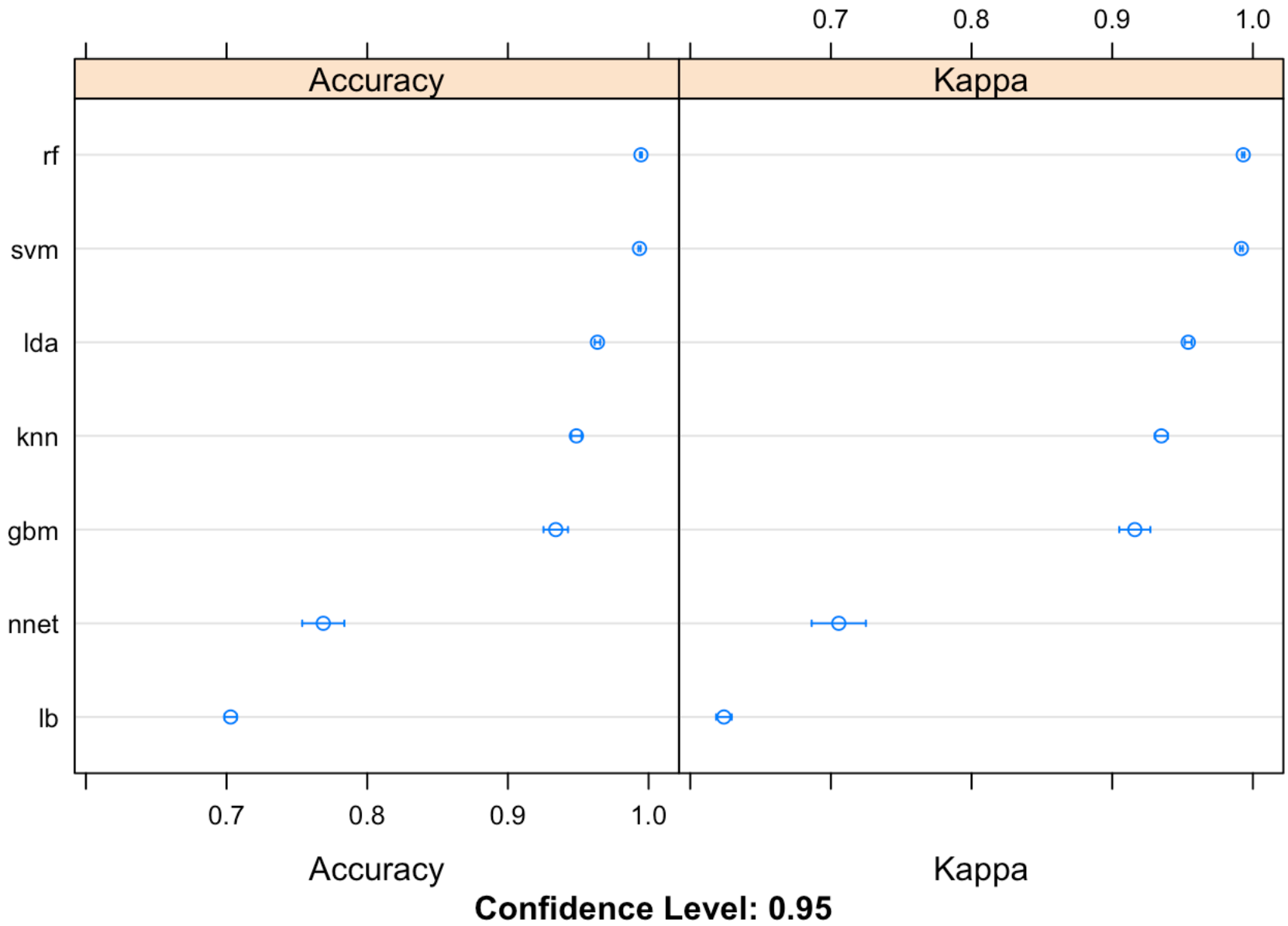
After eliminating the different variables, the final data set was comprised of 19642 rows and 54 variables. The observations were then returned to their original training and testing sets.

Choosing the Model

Seven models were chosen from the 230 available on the caret model page (https://topepo.github.io/caret/available-models.html). Three of the models were specifically chosen because they were identified as being “widely-used” classifiers in An **Introduction to Statistical Learning**. The three models were logistic regression, linear discriminant analysis, and K-nearest neighbor. (James, 127). The models used in the analysis were the following:

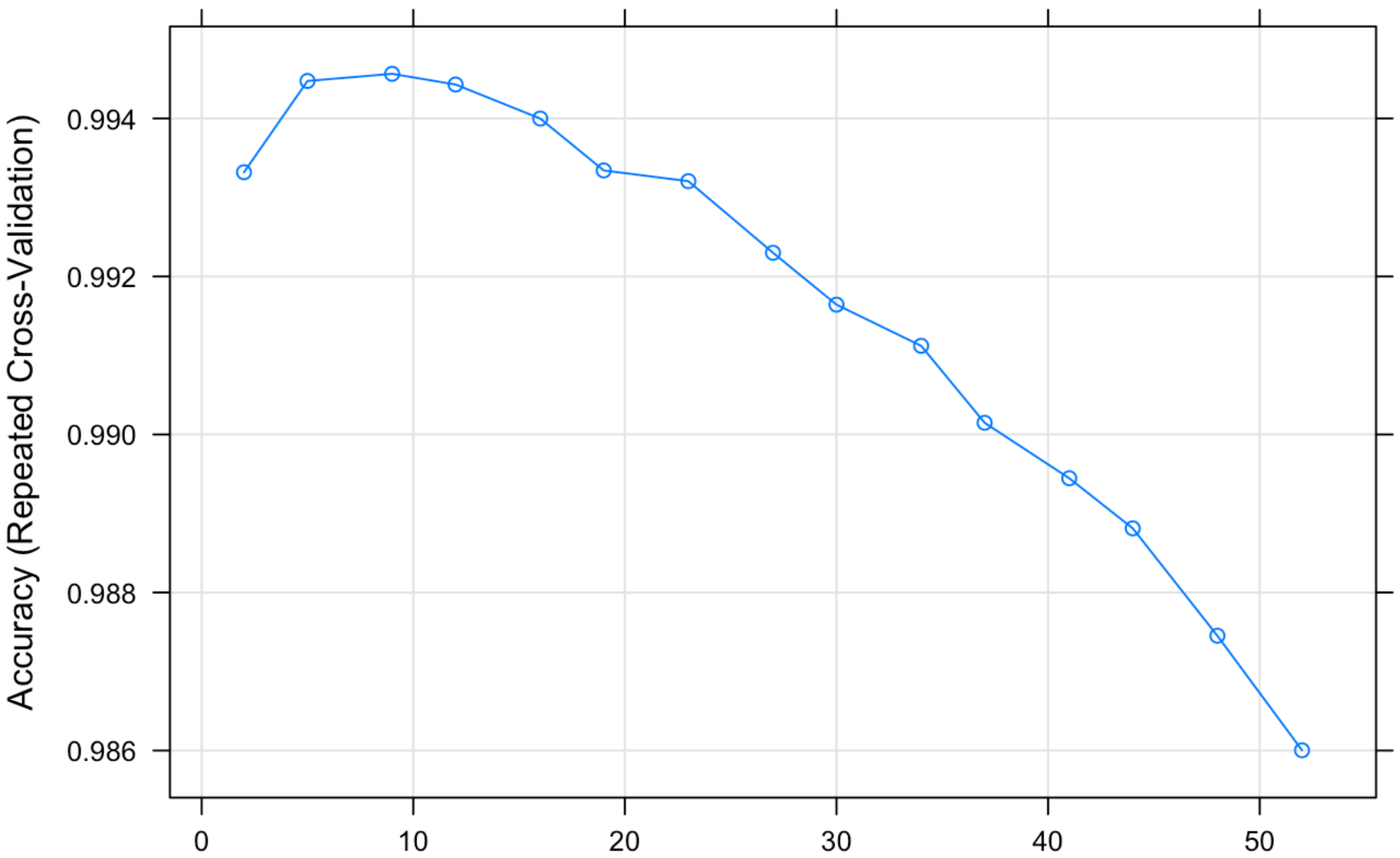
- stochastic gradient boosting (“gbm”)
- boosted logistic regression (“lb”)
- linear descriminate analysis (“lda”)
- k-nearest neighbor (“knn”)
- random forest (“rf”)
- svmRadialWeights (“svm”)
- neural net (“nnet”)

Additionally, each of the models resampled the data using the “repeatedcv” method, as opposed to the bootstrap method, and repeated for a total of three resampling iterations. Cross-validation is a method of randomly choosing samples from the training set and repeatedly fitting the model to the set. This method allows one to draw an inference as to the variability of the results as well. (James, 127).



In descending order of accuracy was random forest with 99.45%, svm radial weights with 99.35% and linear discriminate analysis with 96.37%.

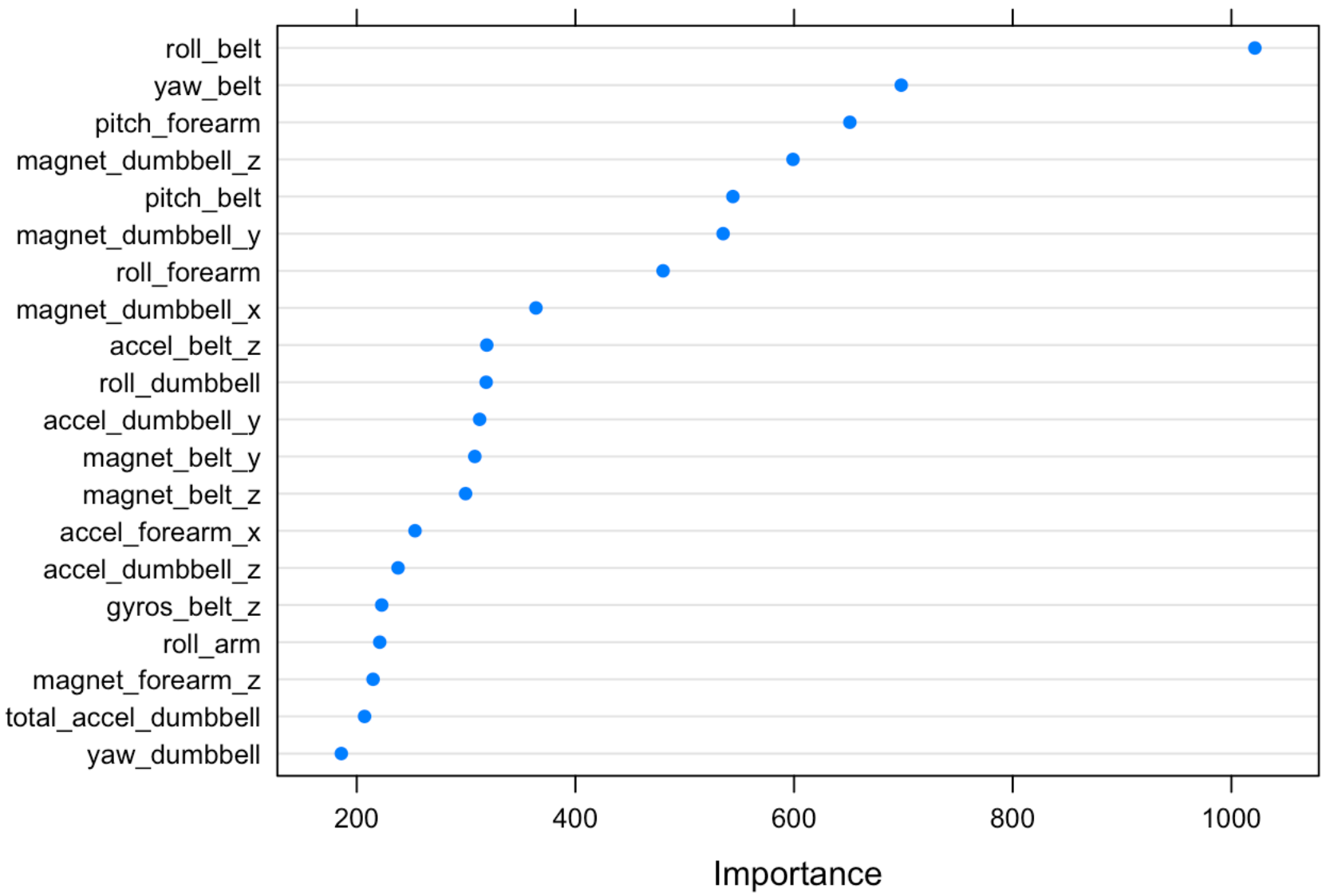
Modeling the Data




```
##   mtry  Accuracy     Kappa  AccuracySD     KappaSD
## 3    9 0.9945646 0.9931243 0.001845194 0.002334679
```

Using the caret package, the training control was set to a parameter of 15. The results show that after 9 randomly chosen variables that the accuracy of the model declines.

Feature Selection



Pursuant to the model, the variables most helpful in accurately classifying the barbell lift were “roll_belt,” “yaw_belt,” and “pitch_forearm.”

Predictions

Applying the model to the test set provided in the course, the following 20 predictions were made. After submission to coursera, all 20 predictions were correct.

```
##  [1] 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
```

Upon attempting to generate a confusion matrix for the Coursera test set, the following error message was generated.

```
confusionMatrix(data = rfClasses, testing$classe)
```

```
## Error in confusionMatrix.default(data = rfClasses, testing$classe): the data canno
t have more levels than the reference
```

Because a confusion matrix could not be generated, the validation set was used consisting of 4904 observations. The results for the confusion matrix were as follows:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1395    7    0    0    0
##          B    0  939    0    0    0
##          C    0    3  855    6    0
##          D    0    0    0  798    0
##          E    0    0    0    0  901
##
## Overall Statistics
##
##                Accuracy : 0.9967
##                  95% CI : (0.9947, 0.9981)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9959
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            1.0000   0.9895   1.0000   0.9925   1.0000
## Specificity            0.9980   1.0000   0.9978   1.0000   1.0000
## Pos Pred Value         0.9950   1.0000   0.9896   1.0000   1.0000
## Neg Pred Value         1.0000   0.9975   1.0000   0.9985   1.0000
## Prevalence             0.2845   0.1935   0.1743   0.1639   0.1837
## Detection Rate         0.2845   0.1915   0.1743   0.1627   0.1837
## Detection Prevalence   0.2859   0.1915   0.1762   0.1627   0.1837
## Balanced Accuracy      0.9990   0.9947   0.9989   0.9963   1.0000
```

Potential Strengths and Weaknesses

The stochastic gradient boosting model computing requirements exceeded the tolerance of my available computer. In order to complete the assignment, the training data set was subset to a random 1000 observations. This likely led to a larger confidence interval surrounding the estimate and may have impacted the overall accuracy measure too.

Additionally, the accuracy and kappa metrics revealed that the random forest model is an excellent strategy in predicting the classification of the barbell lifting. However, a more complete analysis would include ensembling methods to see if those metrics might be improved futher, potentially reaching 100% accuracy.