Machine Learning Course Project - Prediction Assignment Writeup

Hiyong Byun

09/01/2019

Background

Using devices such as Jawbone Up, Nike FuelBand, and Fitbit it is now possible to collect a large amount of data about personal activity relatively inexpensively. These 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. In this project, your goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. More information is available from the website here: http://groupware.les.inf.puc-rio.br/har (http://groupware.les.inf.puc-rio.br/har) (see the section on the Weight Lifting Exercise Dataset).

The goal of your project is to predict the manner in which they did the exercise. This is the “classe” variable in the training set. You may use any of the other variables to predict with. You should create a report describing how you built your model, how you used cross validation, what you think the expected out of sample error is, and why you made the choices you did. You will also use your prediction model to predict 20 different test cases.

Exploratory Data Analyses

Performing str() revealed a large number of variables with missing values. These variables should not be included in the prediction algorithm, along with the variables that indicate identifiers and timestamps (X1, user_name, raw_timestamp_part_1, raw_timestamp_part_2, num_window, cvtd_timestamp, new_window). Since “classe” is a character variable, it was transformed into a factor variable.

Transforming the Dataset

```
varDelete <- NULL
for(i in 1:160) {
       if (mean(is.na(wleTrainingSource[,i])) > 0.9) {
               varDelete[i] <- names(wleTrainingSource[,i])
       }
}
varDelete <- na.omit(varDelete)
wleTrainingTidy <- select(wleTrainingSource, -varDelete)
wleTrainingTidy <- subset(wleTrainingTidy, select = -c(1:7))
wleTrainingTidy$classe <- as.factor(wleTrainingTidy$classe)
```

Re-examining the Dataset

After removing variables with NAs occuring greater than 90% and the first 7 variables, the dataset was re-examined using the summary() function and checked for near zero variance predictors before proceeding with development of a model.

```
nearZeroVar(wleTrainingTidy, saveMetrics=TRUE)
```

```
##                      freqRatio percentUnique zeroVar   nzv
## roll_belt             1.101904     6.7781062   FALSE FALSE
## pitch_belt            1.036082     9.3772296   FALSE FALSE
## yaw_belt              1.058480     9.9734991   FALSE FALSE
## total_accel_belt      1.063160     0.1477933   FALSE FALSE
## gyros_belt_x          1.058651     0.7134849   FALSE FALSE
## gyros_belt_y          1.144000     0.3516461   FALSE FALSE
## gyros_belt_z          1.066214     0.8612782   FALSE FALSE
## accel_belt_x          1.055412     0.8357966   FALSE FALSE
## accel_belt_y          1.113725     0.7287738   FALSE FALSE
## accel_belt_z          1.078767     1.5237998   FALSE FALSE
## magnet_belt_x         1.090141     1.6664968   FALSE FALSE
## magnet_belt_y         1.099688     1.5187035   FALSE FALSE
## magnet_belt_z         1.006369     2.3290184   FALSE FALSE
## roll_arm             52.338462    13.5256345   FALSE FALSE
## pitch_arm            87.256410    15.7323412   FALSE FALSE
## yaw_arm              33.029126    14.6570176   FALSE FALSE
## total_accel_arm       1.024526     0.3363572   FALSE FALSE
## gyros_arm_x           1.015504     3.2769341   FALSE FALSE
## gyros_arm_y           1.454369     1.9162165   FALSE FALSE
## gyros_arm_z           1.110687     1.2638875   FALSE FALSE
## accel_arm_x           1.017341     3.9598410   FALSE FALSE
## accel_arm_y           1.140187     2.7367241   FALSE FALSE
## accel_arm_z           1.128000     4.0362858   FALSE FALSE
## magnet_arm_x          1.000000     6.8239731   FALSE FALSE
## magnet_arm_y          1.056818     4.4439914   FALSE FALSE
## magnet_arm_z          1.036364     6.4468454   FALSE FALSE
## roll_dumbbell         1.022388    84.2065029   FALSE FALSE
## pitch_dumbbell        2.277372    81.7449801   FALSE FALSE
## yaw_dumbbell          1.132231    83.4828254   FALSE FALSE
## total_accel_dumbbell  1.072634     0.2191418   FALSE FALSE
## gyros_dumbbell_x      1.003268     1.2282132   FALSE FALSE
## gyros_dumbbell_y      1.264957     1.4167771   FALSE FALSE
## gyros_dumbbell_z      1.060100     1.0498420   FALSE FALSE
## accel_dumbbell_x      1.018018     2.1659362   FALSE FALSE
## accel_dumbbell_y      1.053061     2.3748853   FALSE FALSE
## accel_dumbbell_z      1.133333     2.0894914   FALSE FALSE
## magnet_dumbbell_x     1.098266     5.7486495   FALSE FALSE
## magnet_dumbbell_y     1.197740     4.3012945   FALSE FALSE
## magnet_dumbbell_z     1.020833     3.4451126   FALSE FALSE
## roll_forearm         11.589286    11.0895933   FALSE FALSE
## pitch_forearm        65.983051    14.8557741   FALSE FALSE
## yaw_forearm          15.322835    10.1467740   FALSE FALSE
## total_accel_forearm   1.128928     0.3567424   FALSE FALSE
## gyros_forearm_x       1.059273     1.5187035   FALSE FALSE
## gyros_forearm_y       1.036554     3.7763735   FALSE FALSE
## gyros_forearm_z       1.122917     1.5645704   FALSE FALSE
## accel_forearm_x       1.126437     4.0464784   FALSE FALSE
## accel_forearm_y       1.059406     5.1116094   FALSE FALSE
## accel_forearm_z       1.006250     2.9558659   FALSE FALSE
## magnet_forearm_x      1.012346     7.7667924   FALSE FALSE
## magnet_forearm_y      1.246914     9.5403119   FALSE FALSE
```

```
## magnet_forearm_z      1.000000     8.5771073   FALSE FALSE
## classe                1.469581     0.0254816   FALSE FALSE
```

Partitioning into Training and Testing Datasets

Dataset was partitioned into 70% training and 30% testing.

```
set.seed(5336)
inTrain <- createDataPartition(wleTrainingTidy$classe, p=0.70, list=FALSE)
wleTraining <- wleTrainingTidy[inTrain,]
wleTesting <- wleTrainingTidy[-inTrain,]
```

Developing a Model

Random forest was chosen to develop a prediction model for classifying 5 classes of exercise. All 52 variables were included in the model. randomForest() function was used for faster processing speed.

```
modelRF <- randomForest(classe~., data=wleTraining)
```

Cross Validation and Accuracy

```
confusionMatrix(wleTesting$classe,predict(modelRF,wleTesting))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1673    0    0    0    1
##          B    3 1135    1    0    0
##          C    0    6 1019    1    0
##          D    0    0   15  949    0
##          E    0    0    1    3 1078
##
## Overall Statistics
##
##                Accuracy : 0.9947
##                  95% CI : (0.9925, 0.9964)
##     No Information Rate : 0.2848
##     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            0.9982   0.9947   0.9836   0.9958   0.9991
## Specificity            0.9998   0.9992   0.9986   0.9970   0.9992
## Pos Pred Value         0.9994   0.9965   0.9932   0.9844   0.9963
## Neg Pred Value         0.9993   0.9987   0.9965   0.9992   0.9998
## Prevalence             0.2848   0.1939   0.1760   0.1619   0.1833
## Detection Rate         0.2843   0.1929   0.1732   0.1613   0.1832
## Detection Prevalence   0.2845   0.1935   0.1743   0.1638   0.1839
## Balanced Accuracy      0.9990   0.9969   0.9911   0.9964   0.9991
```

The model using the random forest algorithm was able to predict the testing dataset with 99.47% accuracy and near zero p-value. Out of sample error was calculated to be 0.53% (31 incorrect predictions out of 5885 observations).

Reference

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.