Predicting the manner in which an individual does an exercise

Rohit

April 16, 2017

Loading the data

The data was downloaded directly from the source and saved as training and testing data frames. The testing data frame was kept aside for applying the trained model prediction. The training dataset was investigated further to identify the relevant predictors and removing irrelvant data or fields with missing information.

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Warning: package 'caret' was built under R version 3.3.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'doParallel' was built under R version 3.3.3
```

```
## Loading required package: foreach
```

```
## Warning: package 'foreach' was built under R version 3.3.3
```

```
## Loading required package: iterators
```

```
## Warning: package 'iterators' was built under R version 3.3.3
```

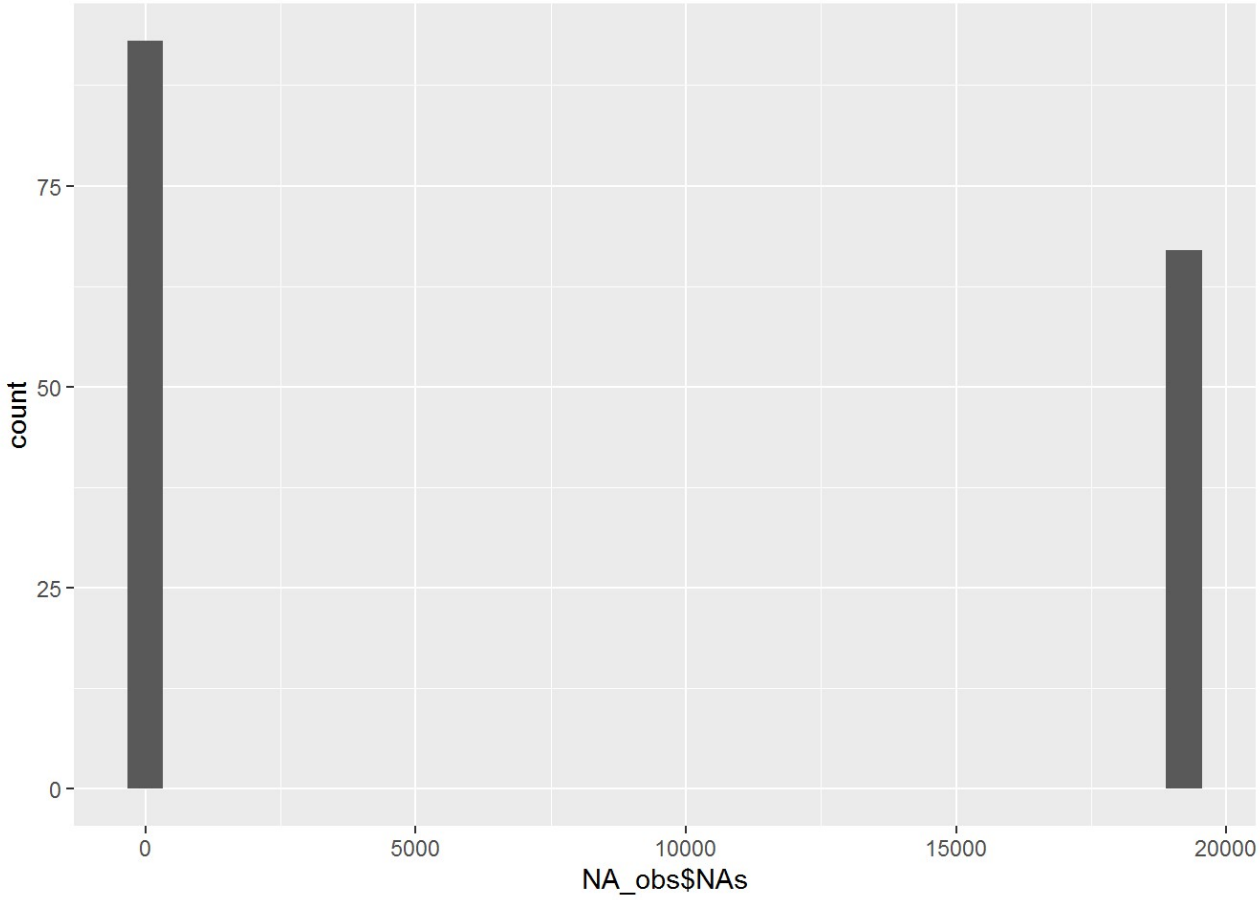
```
## Warning: package 'e1071' was built under R version 3.3.3
```

Exploratory Analysis

After loading the training dataset, the information contained was summarized using the summary function in R. The output provided relevant information regarding which predictors such as missing information, NAs and class.

The plot shown below identifes the number of predictors that contain NAs:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In the next step, the predictors having near zero variability are identified and removed from the data set. Several predictors with near zero variability are also those with NAs.

```
##                          freqRatio percentUnique zeroVar  nzv
## new_window                47.33005    0.01019264   FALSE TRUE
## kurtosis_roll_belt      1921.60000    2.02323922   FALSE TRUE
## kurtosis_picth_belt      600.50000    1.61553358   FALSE TRUE
## kurtosis_yaw_belt         47.33005    0.01019264   FALSE TRUE
## skewness_roll_belt      2135.11111    2.01304658   FALSE TRUE
## skewness_roll_belt.1     600.50000    1.72255631   FALSE TRUE
## skewness_yaw_belt         47.33005    0.01019264   FALSE TRUE
## max_yaw_belt              640.53333    0.34654979   FALSE TRUE
## min_yaw_belt              640.53333    0.34654979   FALSE TRUE
## amplitude_yaw_belt         50.04167    0.02038528   FALSE TRUE
## avg_roll_arm               77.00000    1.68178575   FALSE TRUE
## stddev_roll_arm            77.00000    1.68178575   FALSE TRUE
## var_roll_arm               77.00000    1.68178575   FALSE TRUE
## avg_pitch_arm              77.00000    1.68178575   FALSE TRUE
## stddev_pitch_arm           77.00000    1.68178575   FALSE TRUE
## var_pitch_arm              77.00000    1.68178575   FALSE TRUE
## avg_yaw_arm                77.00000    1.68178575   FALSE TRUE
## stddev_yaw_arm             80.00000    1.66649679   FALSE TRUE
## var_yaw_arm                80.00000    1.66649679   FALSE TRUE
## kurtosis_roll_arm         246.35897    1.68178575   FALSE TRUE
## kurtosis_picth_arm        240.20000    1.67159311   FALSE TRUE
## kurtosis_yaw_arm         1746.90909    2.01304658   FALSE TRUE
## skewness_roll_arm         249.55844    1.68688207   FALSE TRUE
## skewness_pitch_arm        240.20000    1.67159311   FALSE TRUE
## skewness_yaw_arm         1746.90909    2.01304658   FALSE TRUE
## max_roll_arm               25.66667    1.47793293   FALSE TRUE
## min_roll_arm               19.25000    1.41677709   FALSE TRUE
## min_pitch_arm              19.25000    1.47793293   FALSE TRUE
## amplitude_roll_arm         25.66667    1.55947406   FALSE TRUE
## amplitude_pitch_arm        20.00000    1.49831821   FALSE TRUE
## kurtosis_roll_dumbbell   3843.20000    2.02833554   FALSE TRUE
## kurtosis_picth_dumbbell  9608.00000    2.04362450   FALSE TRUE
## kurtosis_yaw_dumbbell      47.33005    0.01019264   FALSE TRUE
## skewness_roll_dumbbell   4804.00000    2.04362450   FALSE TRUE
## skewness_pitch_dumbbell  9608.00000    2.04872082   FALSE TRUE
## skewness_yaw_dumbbell      47.33005    0.01019264   FALSE TRUE
## max_yaw_dumbbell           960.80000    0.37203139   FALSE TRUE
## min_yaw_dumbbell           960.80000    0.37203139   FALSE TRUE
## amplitude_yaw_dumbbell      47.92020    0.01528896   FALSE TRUE
## kurtosis_roll_forearm      228.76190    1.64101519   FALSE TRUE
## kurtosis_picth_forearm     226.07059    1.64611151   FALSE TRUE
## kurtosis_yaw_forearm        47.33005    0.01019264   FALSE TRUE
## skewness_roll_forearm      231.51807    1.64611151   FALSE TRUE
## skewness_pitch_forearm     226.07059    1.62572623   FALSE TRUE
## skewness_yaw_forearm        47.33005    0.01019264   FALSE TRUE
## max_roll_forearm            27.66667    1.38110284   FALSE TRUE
## max_yaw_forearm            228.76190    0.22933442   FALSE TRUE
```

```
## min_roll_forearm          27.66667    1.37091020   FALSE TRUE
## min_yaw_forearm          228.76190    0.22933442   FALSE TRUE
## amplitude_roll_forearm    20.75000    1.49322189   FALSE TRUE
## amplitude_yaw_forearm     59.67702    0.01528896   FALSE TRUE
## avg_roll_forearm          27.66667    1.64101519   FALSE TRUE
## stddev_roll_forearm       87.00000    1.63082255   FALSE TRUE
## var_roll_forearm          87.00000    1.63082255   FALSE TRUE
## avg_pitch_forearm         83.00000    1.65120783   FALSE TRUE
## stddev_pitch_forearm      41.50000    1.64611151   FALSE TRUE
## var_pitch_forearm         83.00000    1.65120783   FALSE TRUE
## avg_yaw_forearm           83.00000    1.65120783   FALSE TRUE
## stddev_yaw_forearm        85.00000    1.64101519   FALSE TRUE
## var_yaw_forearm           85.00000    1.64101519   FALSE TRUE
```

Model Creation

After removing the irrelevant predictors, the new training data set is split into two parts: (1) a data frame of the predictors; and (2) a data frame comprising of the output variable. Rando Forest was used to train the data. The train function in caret package was used to model the data. To speed up the processing, additional clusters were allocated using the functions in parallel and doParallel packages. The inbuilt train control function was used to create 10-folds in the dataset. This was directly included as an argument in the train function.

```
## Loading required package: randomForest
```

```
## Warning: package 'randomForest' was built under R version 3.3.3
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

Results

Using the model along with the 10-fold crossvalidation approach, the results indicate a out of sample error of 0.43%

```
##     Accuracy     Kappa   Resample
## 1  0.9939734 0.9923770 Resample09
## 2  0.9927988 0.9908893 Resample05
## 3  0.9920591 0.9899352 Resample01
## 4  0.9945025 0.9930402 Resample10
## 5  0.9921292 0.9900558 Resample06
## 6  0.9941885 0.9926554 Resample02
## 7  0.9935819 0.9918763 Resample11
## 8  0.9936376 0.9919424 Resample07
## 9  0.9914626 0.9892144 Resample03
## 10 0.9943440 0.9928498 Resample12
## 11 0.9921129 0.9900105 Resample08
## 12 0.9920910 0.9899772 Resample04
## 13 0.9919878 0.9898738 Resample13
## 14 0.9918339 0.9896675 Resample22
## 15 0.9914459 0.9891822 Resample18
## 16 0.9947974 0.9934077 Resample14
## 17 0.9918396 0.9896665 Resample23
## 18 0.9927697 0.9908499 Resample19
## 19 0.9925042 0.9905047 Resample15
## 20 0.9914753 0.9892250 Resample24
## 21 0.9927435 0.9908304 Resample20
## 22 0.9901676 0.9875516 Resample16
## 23 0.9926358 0.9906928 Resample25
## 24 0.9937265 0.9920483 Resample21
## 25 0.9921238 0.9900358 Resample17
```

```
## Bootstrapped (25 reps) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction    A    B    C    D    E
##          A 28.5  0.1  0.0  0.0  0.0
##          B  0.0 19.2  0.2  0.0  0.0
##          C  0.0  0.0 17.3  0.3  0.0
##          D  0.0  0.0  0.0 16.0  0.0
##          E  0.0  0.0  0.0  0.0 18.4
##
##  Accuracy (average) : 0.9927
```

The model was then used to predict the outcome using the testing dataset. A similar preprocessing was also performed on the testing dataset i.e., removing predictors with NAs, near zero variance and that are irrelevant.