Machine Learning Project

Erick Yegon

2022-07-18

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: websit (see the section on the Weight Lifting Exercise Dataset). ##Data

The training data for this project are available here: train_data

The test data are available here: test_data

The data for this project come from this source: source. If you use the document you create for this class for any purpose please cite them as they have been very generous in allowing their data to be used for this kind of assignment. ##What you should submit

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.

Your submission should consist of a link to a Github repo with your R markdown and compiled HTML file describing your analysis. Please constrain the text of the writeup to < 2000 words and the number of figures to be less than 5. It will make it easier for the graders if you submit a repo with a gh-pages branch so the HTML page can be viewed online (and you always want to make it easy on graders :-). You should also apply your machine learning algorithm to the 20 test cases available in the test data above. Please submit your predictions in appropriate format to the programming assignment for automated grading. See the programming assignment for additional details.

Below we carry out some preliminary Work

Reproduceability

An overall pseudo-random number generator seed was set at 1234 for all code. In order to reproduce the results below, the same seed should be used. Different packages were downloaded and installed, such as caret and randomForest. These should also be installed in order to reproduce the results below (please see code below for ways and syntax to do so). ##How the model was built

Our outcome variable is classe, a factor variable with 5 levels. For this data set, "participants were asked to perform one set of 10 repetitions of the Unilateral Dumbbell Biceps Curl in 5 different fashions:

exactly according to the specification (Class A)

```
* throwing the elbows to the front (Class B)

* lifting the dumbbell only halfway (Class C)

* lowering the dumbbell only halfway (Class D)

* throwing the hips to the front (Class E)?
```

Class A corresponds to the specified execution of the exercise, while the other 4 classes correspond to common mistakes." [1] Prediction evaluations will be based on maximizing the accuracy and minimizing the out-of-sample error. All other available variables after cleaning will be used for prediction. Two models will be tested using decision tree and random forest algorithms. The model with the highest accuracy will be chosen as our final model.

Cross-validation

We will perform cross-validation by subsampling our training data set randomly without replacement into 2 subsamples: subTraining data (75% of the original Training data set) and subTesting data (25%). Our models will be fitted on the subtraining data set, and tested on the subtesting data. Once the most accurate model is choosen, it will be tested on the original Testing data set.

Expected out-of-sample error

Loading required packages, libraries and setting seed

Installing packages, loading libraries, and setting the seed for reproduceability:

Set working directory

```
setwd("C:/Users/Erick Yegon/Dropbox/My PC (DESKTOP-1I4SCDT)/Desktop/Prediction")
```

Load required R packages and set a seed.

```
RequiredPackages <- c("caret","randomForest","rpart","rpart.plot","RColorBrewer","rattle","corrplot")
for (i in RequiredPackages) { #Installs packages if not yet installed
    if (!require(i, character.only = TRUE)) install.packages(i)
}
```

```
## Loading required package: caret

## Loading required package: ggplot2

## Loading required package: lattice

## Loading required package: randomForest
```

```
## randomForest 4.7-1.1

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

##
## Attaching package: 'randomForest'

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

## Loading required package: rpart

## Loading required package: rpart.plot

## Loading required package: RColorBrewer

## Loading required package: rattle

## Loading required package: tibble

## Loading required package: bitops

## Rattle: A free graphical interface for data science with R.
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

##
## Attaching package: 'rattle'

## The following object is masked from 'package:randomForest':
##
##     importance

## Loading required package: corrplot

## corrplot 0.92 loaded
```

```
library(caret)
library(randomForest)
library(rpart)
library(rpart.plot)
library(corrplot)
library(RColorBrewer)
library(rattle)
set.seed(2254)
```

```
url_train <-"http://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
url_quiz  <-"http://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
data_train <-read.csv(url(url_train), strip.white = TRUE,
                      na.strings = c("NA",""))
data_quiz  <-read.csv(url(url_quiz),
                      strip.white = TRUE, na.strings = c("NA",""))
```

Checking the dimensions of the data

```
dim(data_train)[1]
```

```
## [1] 19622
```

```
dim(data_quiz)[1]
```

```
## [1] 20
```

Create two partitions (75% and 25%) within the original training dataset.

```
in_train  <-createDataPartition(data_train$classe, p=0.75, list=FALSE)

train_set <-data_train[ in_train, ]
test_set  <-data_train[-in_train, ]
dim(train_set)
```

```
## [1] 14718   160
```

```
dim(test_set)
```

```
## [1] 4904  160
```

The two datasets (train_setandtest_set) have a large number ofNAvalues as well as near-zero-variance (NZV) variables. Both will be removed together with their ID variables.

```
nzv_var <-nearZeroVar(train_set)

train_set <-train_set[ , -nzv_var]

test_set  <-test_set [ , -nzv_var]

dim(train_set)
```

```
## [1] 14718   119
```

```
dim(test_set)
```

```
## [1] 4904  119
```

Remove variables that are mostly NA. A threshlod of 95 % is selected.

```
na_var <-sapply(train_set, function(x) mean(is.na(x))) > 0.95

train_set <-train_set[ , na_var == FALSE]

test_set  <-test_set [ , na_var == FALSE]

dim(train_set)
```

```
## [1] 14718    59
```

```
dim(test_set)
```

```
## [1] 4904   59
```

Since columns 1 to 5 are identification variables only, they will be removed as well.

```
train_set <-train_set[ , -(1:5)]

test_set  <-test_set [ , -(1:5)]

dim(train_set)
```

```
## [1] 14718    54
```

```
dim(test_set)
```

```
## [1] 4904   54
```

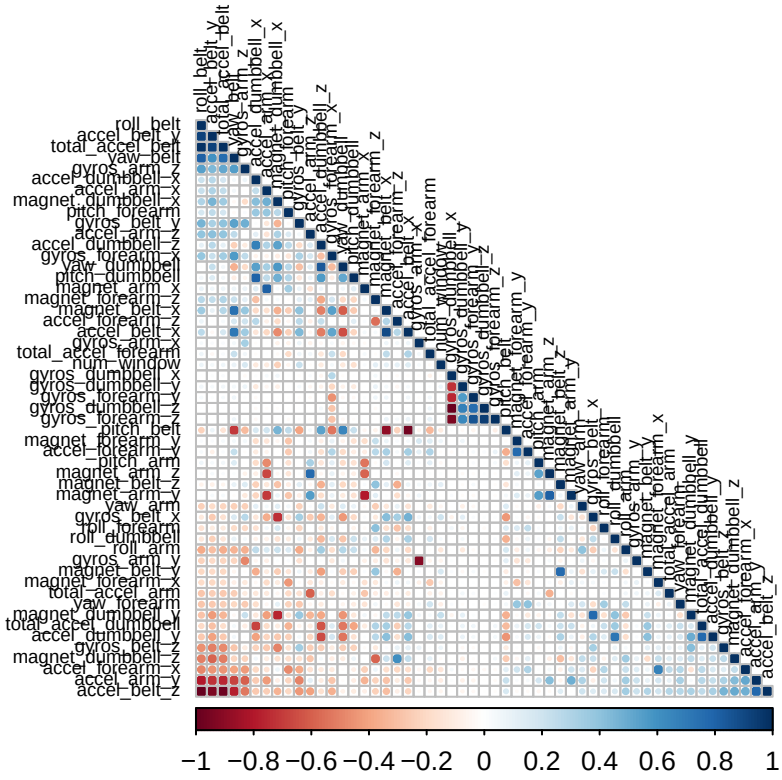
The number of variables for the analysis has been reduced from the original 160 down to 54.

Correlation Analysis

Correlation analysis between the variables before the modeling work itself is done. The "FPC" is used as the first principal component order

```
corr_matrix <-cor(train_set[ , -54])

corrplot(corr_matrix, order = "FPC", method = "circle", type = "lower",tl.cex = 0.6, tl.col = rgb(0, 0,
```



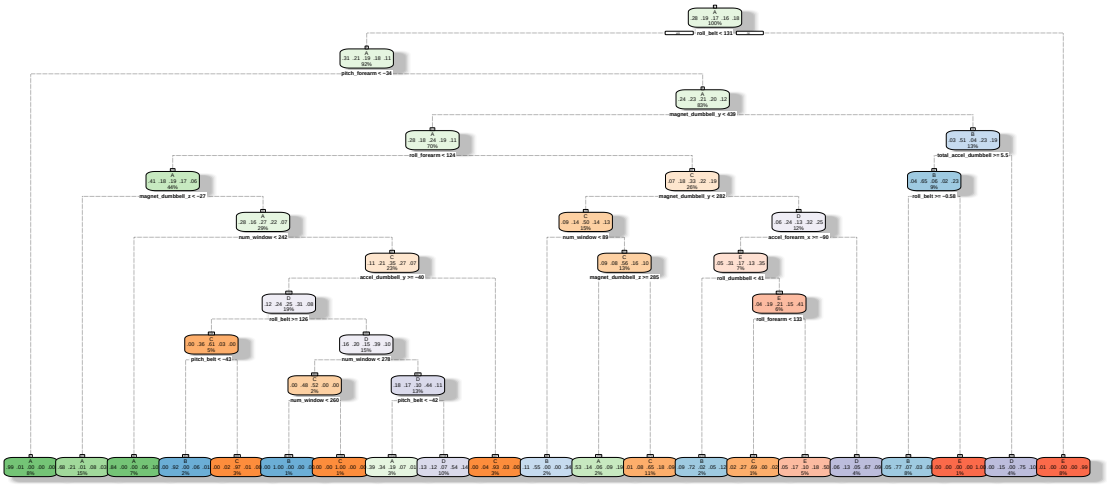
If two variables are highly correlated their colors are either dark blue (for a positive correlation) or dark red (for a negative correlations). Because there are only few strong correlations among the input variables, the Principal Components Analysis (PCA) will not be performed in this analysis. Instead, a few different prediction models will be built to have a better accuracy.

Prediction Models

Decision Tree Model

```
set.seed(2222)
fit_decision_tree <-rpart(classe ~ ., data = train_set, method="class")
fancyRpartPlot(fit_decision_tree)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



Rattle 2022−Jul−18 09:21:11 Erick Yegon

Predictions of the decision tree model ontest_set.predict_decision_tree

```
predict_decision_tree <-predict(fit_decision_tree, newdata = test_set, type="class")


conf_matrix_decision_tree <-confusionMatrix(predict_decision_tree, factor(test_set$classe))

conf_matrix_decision_tree
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1263  229   37  109   86
##          B   40  502   32   23   75
##          C   14   52  700  128   52
##          D   63  128   64  494  109
##          E   15   38   22   50  579
##
## Overall Statistics
##
##                Accuracy : 0.7215
##                  95% CI : (0.7087, 0.734)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
```

```
##                   Kappa : 0.6451
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9054   0.5290   0.8187   0.6144   0.6426
## Specificity            0.8686   0.9570   0.9392   0.9112   0.9688
## Pos Pred Value         0.7326   0.7470   0.7400   0.5758   0.8224
## Neg Pred Value         0.9585   0.8944   0.9608   0.9234   0.9233
## Prevalence             0.2845   0.1935   0.1743   0.1639   0.1837
## Detection Rate         0.2575   0.1024   0.1427   0.1007   0.1181
## Detection Prevalence   0.3515   0.1370   0.1929   0.1750   0.1436
## Balanced Accuracy      0.8870   0.7430   0.8790   0.7628   0.8057
```

The predictive accuracy of the decision tree model is relatively low at 75.2 %.Plot the predictive accuracy of the decision tree model.

```
print(summary(fit_decision_tree ))
```

```
## Call:
## rpart(formula = classe ~ ., data = train_set, method = "class")
##   n= 14718
##
##            CP nsplit rel error    xerror        xstd
## 1  0.11582645      0 1.0000000 1.0000000 0.005195739
## 2  0.06038166      1 0.8841735 0.8841735 0.005552214
## 3  0.03897275      4 0.7030286 0.7994873 0.005698653
## 4  0.03256432      6 0.6250831 0.5962214 0.005696697
## 5  0.03104529      7 0.5925188 0.5673597 0.005656323
## 6  0.02525396      8 0.5614735 0.5363144 0.005601302
## 7  0.02278553     10 0.5109655 0.4894142 0.005494585
## 8  0.02174119     11 0.4881800 0.4848571 0.005482653
## 9  0.02079180     12 0.4664388 0.4707111 0.005443794
## 10 0.01376626     13 0.4456470 0.4419444 0.005356080
## 11 0.01338650     14 0.4318808 0.3936201 0.005181038
## 12 0.01243710     15 0.4184943 0.3887781 0.005161478
## 13 0.01219975     16 0.4060572 0.3796639 0.005123613
## 14 0.01139277     18 0.3816576 0.3654230 0.005061652
## 15 0.01034843     19 0.3702649 0.3460553 0.004971700
## 16 0.01000000     20 0.3599165 0.3408336 0.004946287
##
## Variable importance
##            roll_belt           num_window        pitch_forearm
##                   13                    8                    6
##           pitch_belt         accel_belt_z    magnet_dumbbell_y
##                    6                    5                    5
##     accel_dumbbell_y total_accel_dumbbell        roll_dumbbell
##                    5                    5                    4
##     total_accel_belt    magnet_dumbbell_z             yaw_belt
##                    4                    4                    3
##      accel_forearm_x         roll_forearm         accel_belt_y
```

```
##                   3                   3                   3
##       magnet_belt_x       magnet_belt_z        accel_belt_x
##                   2                   2                   2
##    accel_dumbbell_x       magnet_belt_y    magnet_dumbbell_x
##                   2                   2                   2
##        yaw_dumbbell    accel_dumbbell_z     magnet_forearm_x
##                   2                   2                   1
##             yaw_arm     accel_forearm_z     gyros_dumbbell_x
##                   1                   1                   1
##            roll_arm
##                   1
##
## Node number 1: 14718 observations,    complexity param=0.1158265
##   predicted class=A  expected loss=0.7156543  P(node) =1
##     class counts:  4185  2848  2567  2412  2706
##    probabilities: 0.284 0.194 0.174 0.164 0.184
##   left son=2 (13474 obs) right son=3 (1244 obs)
##   Primary splits:
##       roll_belt        < 130.5     to the left,  improve=1115.1000, (0 missing)
##       pitch_forearm    < -33.65    to the left,  improve= 806.0055, (0 missing)
##       accel_belt_z     < -187.5    to the right, improve= 678.6299, (0 missing)
##       magnet_belt_y    < 555.5     to the right, improve= 626.7191, (0 missing)
##       total_accel_belt < 20.5      to the left,  improve= 562.3080, (0 missing)
##   Surrogate splits:
##       accel_belt_z     < -187.5    to the right, agree=0.966, adj=0.599, (0 split)
##       total_accel_belt < 20.5      to the left,  agree=0.958, adj=0.509, (0 split)
##       magnet_belt_y    < 553.5     to the right, agree=0.933, adj=0.212, (0 split)
##       magnet_belt_z    < -448.5    to the right, agree=0.933, adj=0.211, (0 split)
##       accel_belt_x     < 55.5      to the left,  agree=0.923, adj=0.088, (0 split)
##
## Node number 2: 13474 observations,    complexity param=0.06038166
##   predicted class=A  expected loss=0.6902924  P(node) =0.9154776
##     class counts:  4173  2848  2567  2412  1474
##    probabilities: 0.310 0.211 0.191 0.179 0.109
##   left son=4 (1190 obs) right son=5 (12284 obs)
##   Primary splits:
##       pitch_forearm     < -33.65    to the left,  improve=761.6414, (0 missing)
##       roll_forearm      < 123.5     to the left,  improve=454.9988, (0 missing)
##       magnet_dumbbell_y < 438.5     to the left,  improve=436.1623, (0 missing)
##       magnet_arm_x      < 67.5      to the left,  improve=414.6067, (0 missing)
##       accel_arm_x       < -270.5    to the left,  improve=405.4459, (0 missing)
##   Surrogate splits:
##       accel_forearm_x      < 220.5     to the right, agree=0.932, adj=0.230, (0 split)
##       accel_dumbbell_x     < -201      to the left,  agree=0.931, adj=0.218, (0 split)
##       total_accel_dumbbell < 36.5      to the right, agree=0.931, adj=0.215, (0 split)
##       yaw_arm              < -160.5    to the left,  agree=0.923, adj=0.132, (0 split)
##       accel_dumbbell_z     < -219.5    to the left,  agree=0.923, adj=0.127, (0 split)
##
## Node number 3: 1244 observations
##   predicted class=E  expected loss=0.009646302  P(node) =0.08452235
##     class counts:    12     0     0     0  1232
##    probabilities: 0.010 0.000 0.000 0.000 0.990
##
## Node number 4: 1190 observations
```

```
##   predicted class=A  expected loss=0.01008403  P(node) =0.08085338
##     class counts:  1178    12     0     0     0
##    probabilities: 0.990 0.010 0.000 0.000 0.000
##
## Node number 5: 12284 observations,    complexity param=0.06038166
##   predicted class=A  expected loss=0.7561869  P(node) =0.8346243
##     class counts:  2995  2836  2567  2412  1474
##    probabilities: 0.244 0.231 0.209 0.196 0.120
##   left son=10 (10366 obs) right son=11 (1918 obs)
##   Primary splits:
##       magnet_dumbbell_y < 438.5     to the left,  improve=351.9234, (0 missing)
##       num_window        < 45.5      to the right, improve=350.9075, (0 missing)
##       yaw_belt          < 169.5     to the right, improve=348.7021, (0 missing)
##       roll_forearm      < 123.5     to the left,  improve=336.0297, (0 missing)
##       roll_dumbbell     < 64.45452  to the left,  improve=294.1711, (0 missing)
##   Surrogate splits:
##       roll_dumbbell        < 81.66746  to the left,  agree=0.899, adj=0.351, (0 split)
##       accel_dumbbell_y     < 195.5     to the left,  agree=0.874, adj=0.196, (0 split)
##       total_accel_dumbbell < 31.5      to the left,  agree=0.855, adj=0.069, (0 split)
##       gyros_dumbbell_y     < 1.21      to the left,  agree=0.853, adj=0.058, (0 split)
##       accel_forearm_x      < -401.5    to the right, agree=0.850, adj=0.042, (0 split)
##
## Node number 10: 10366 observations,    complexity param=0.06038166
##   predicted class=A  expected loss=0.7168628  P(node) =0.7043077
##     class counts:  2935  1865  2485  1976  1105
##    probabilities: 0.283 0.180 0.240 0.191 0.107
##   left son=20 (6480 obs) right son=21 (3886 obs)
##   Primary splits:
##       roll_forearm      < 123.5     to the left,  improve=368.9514, (0 missing)
##       yaw_belt          < 169.5     to the right, improve=320.2713, (0 missing)
##       num_window        < 241.5     to the left,  improve=295.8974, (0 missing)
##       magnet_dumbbell_z < -24.5     to the left,  improve=292.8034, (0 missing)
##       accel_dumbbell_y  < -39.5     to the right, improve=243.8687, (0 missing)
##   Surrogate splits:
##       total_accel_dumbbell < 21.5      to the left,  agree=0.727, adj=0.271, (0 split)
##       accel_dumbbell_x     < -65.5     to the right, agree=0.724, adj=0.263, (0 split)
##       roll_belt            < 1.315     to the right, agree=0.722, adj=0.258, (0 split)
##       accel_forearm_z      < -87.5     to the right, agree=0.720, adj=0.254, (0 split)
##       accel_belt_z         < -2.5      to the left,  agree=0.720, adj=0.252, (0 split)
##
## Node number 11: 1918 observations,    complexity param=0.03104529
##   predicted class=B  expected loss=0.4937435  P(node) =0.1303166
##     class counts:    60   971    82   436   369
##    probabilities: 0.031 0.506 0.043 0.227 0.192
##   left son=22 (1378 obs) right son=23 (540 obs)
##   Primary splits:
##       total_accel_dumbbell < 5.5       to the right, improve=312.5889, (0 missing)
##       num_window           < 258.5     to the right, improve=259.2518, (0 missing)
##       accel_dumbbell_y     < 46.5      to the right, improve=256.2817, (0 missing)
##       roll_belt            < -0.58     to the right, improve=238.0849, (0 missing)
##       yaw_belt             < -2.825    to the left,  improve=229.2540, (0 missing)
##   Surrogate splits:
##       accel_dumbbell_y < 45.5      to the right, agree=0.965, adj=0.874, (0 split)
##       yaw_belt         < -3.305    to the left,  agree=0.844, adj=0.444, (0 split)
```

```
##       accel_forearm_x  < -295.5    to the right, agree=0.828, adj=0.389, (0 split)
##       roll_belt        < 117.5     to the left,  agree=0.826, adj=0.383, (0 split)
##       pitch_belt       < 9.005     to the left,  agree=0.817, adj=0.350, (0 split)
##
## Node number 20: 6480 observations,    complexity param=0.03897275
##   predicted class=A  expected loss=0.5916667  P(node) =0.4402772
##     class counts:  2646  1149  1199  1103   383
##    probabilities: 0.408 0.177 0.185 0.170 0.059
##   left son=40 (2146 obs) right son=41 (4334 obs)
##   Primary splits:
##       magnet_dumbbell_z < -27.5     to the left,  improve=359.4466, (0 missing)
##       num_window        < 241.5     to the left,  improve=327.5623, (0 missing)
##       accel_dumbbell_y  < -40.5     to the right, improve=291.0313, (0 missing)
##       roll_forearm      < -131.5    to the right, improve=289.0846, (0 missing)
##       pitch_belt        < 15.45     to the right, improve=231.4264, (0 missing)
##   Surrogate splits:
##       pitch_belt        < 15.35     to the right, agree=0.886, adj=0.657, (0 split)
##       accel_belt_x      < -20.5     to the left,  agree=0.839, adj=0.513, (0 split)
##       magnet_belt_x     < 23.5      to the left,  agree=0.824, adj=0.470, (0 split)
##       accel_belt_y      < 53.5      to the right, agree=0.819, adj=0.454, (0 split)
##       magnet_dumbbell_y < 101       to the left,  agree=0.777, adj=0.326, (0 split)
##
## Node number 21: 3886 observations,    complexity param=0.03256432
##   predicted class=C  expected loss=0.6690685  P(node) =0.2640304
##     class counts:   289   716  1286   873   722
##    probabilities: 0.074 0.184 0.331 0.225 0.186
##   left son=42 (2159 obs) right son=43 (1727 obs)
##   Primary splits:
##       magnet_dumbbell_y < 281.5     to the left,  improve=186.6606, (0 missing)
##       accel_forearm_x   < -104.5    to the right, improve=177.2476, (0 missing)
##       magnet_arm_y      < 288.5     to the right, improve=153.5983, (0 missing)
##       magnet_dumbbell_z < 284.5     to the right, improve=151.6551, (0 missing)
##       magnet_forearm_z  < -251      to the left,  improve=151.4750, (0 missing)
##   Surrogate splits:
##       roll_dumbbell        < 20.87401  to the left,  agree=0.845, adj=0.652, (0 split)
##       accel_dumbbell_y     < 16.5      to the left,  agree=0.839, adj=0.638, (0 split)
##       total_accel_dumbbell < 14.5      to the left,  agree=0.754, adj=0.446, (0 split)
##       accel_dumbbell_z     < -110      to the right, agree=0.710, adj=0.349, (0 split)
##       magnet_belt_y        < 612.5     to the left,  agree=0.710, adj=0.347, (0 split)
##
## Node number 22: 1378 observations,    complexity param=0.0207918
##   predicted class=B  expected loss=0.3526851  P(node) =0.09362685
##     class counts:    60   892    81    30   315
##    probabilities: 0.044 0.647 0.059 0.022 0.229
##   left son=44 (1159 obs) right son=45 (219 obs)
##   Primary splits:
##       roll_belt         < -0.58     to the right, improve=265.5660, (0 missing)
##       num_window        < 258       to the right, improve=205.7285, (0 missing)
##       gyros_belt_z      < -0.255    to the right, improve=149.6044, (0 missing)
##       magnet_dumbbell_z < 17.5      to the left,  improve=107.2762, (0 missing)
##       magnet_belt_z     < -291.5    to the left,  improve=105.3596, (0 missing)
##   Surrogate splits:
##       magnet_belt_z < -289.5    to the left,  agree=0.902, adj=0.384, (0 split)
##       accel_belt_y  < -1.5      to the right, agree=0.893, adj=0.324, (0 split)
```

```
##       num_window    < 91.5      to the right, agree=0.875, adj=0.215, (0 split)
##       magnet_belt_x < -8.5      to the right, agree=0.864, adj=0.146, (0 split)
##       gyros_belt_y  < -0.04     to the right, agree=0.860, adj=0.119, (0 split)
##
## Node number 23: 540 observations
##   predicted class=D  expected loss=0.2481481  P(node) =0.03668977
##     class counts:     0    79     1   406    54
##    probabilities: 0.000 0.146 0.002 0.752 0.100
##
## Node number 40: 2146 observations
##   predicted class=A  expected loss=0.3247903  P(node) =0.1458079
##     class counts:  1449   448    26   162    61
##    probabilities: 0.675 0.209 0.012 0.075 0.028
##
## Node number 41: 4334 observations,    complexity param=0.03897275
##   predicted class=A  expected loss=0.7238117  P(node) =0.2944694
##     class counts:  1197   701  1173   941   322
##    probabilities: 0.276 0.162 0.271 0.217 0.074
##   left son=82 (1003 obs) right son=83 (3331 obs)
##   Primary splits:
##       num_window       < 241.5     to the left,  improve=582.1772, (0 missing)
##       yaw_belt         < 168.5     to the right, improve=301.3853, (0 missing)
##       accel_dumbbell_y < -40.5     to the right, improve=288.0807, (0 missing)
##       pitch_belt       < -42.45    to the left,  improve=236.6059, (0 missing)
##       roll_dumbbell    < -87.85003 to the right, improve=210.5195, (0 missing)
##   Surrogate splits:
##       magnet_forearm_x < 545.5     to the right, agree=0.816, adj=0.204, (0 split)
##       roll_belt        < 128.5     to the right, agree=0.811, adj=0.184, (0 split)
##       yaw_belt         < 172.5     to the right, agree=0.810, adj=0.180, (0 split)
##       yaw_arm          < -113.5    to the left,  agree=0.802, adj=0.143, (0 split)
##       magnet_belt_x    < 177.5     to the right, agree=0.801, adj=0.142, (0 split)
##
## Node number 42: 2159 observations,    complexity param=0.0133865
##   predicted class=C  expected loss=0.5048634  P(node) =0.1466911
##     class counts:   191   302  1069   313   284
##    probabilities: 0.088 0.140 0.495 0.145 0.132
##   left son=84 (256 obs) right son=85 (1903 obs)
##   Primary splits:
##       num_window        < 88.5      to the left,  improve=139.5089, (0 missing)
##       magnet_forearm_z  < -251      to the left,  improve=136.0914, (0 missing)
##       magnet_dumbbell_z < 287.5     to the right, improve=126.0386, (0 missing)
##       magnet_forearm_y  < 842       to the right, improve=115.4953, (0 missing)
##       pitch_belt        < 26.15     to the right, improve=100.0732, (0 missing)
##   Surrogate splits:
##       pitch_belt        < 26.15     to the right, agree=0.903, adj=0.184, (0 split)
##       magnet_dumbbell_z < -157.5    to the left,  agree=0.895, adj=0.117, (0 split)
##       magnet_belt_x     < -5.5      to the left,  agree=0.894, adj=0.109, (0 split)
##       pitch_forearm     < 62.25     to the right, agree=0.893, adj=0.102, (0 split)
##       magnet_arm_x      < -462      to the left,  agree=0.886, adj=0.039, (0 split)
##
## Node number 43: 1727 observations,    complexity param=0.02278553
##   predicted class=D  expected loss=0.6757383  P(node) =0.1173393
##     class counts:    98   414   217   560   438
##    probabilities: 0.057 0.240 0.126 0.324 0.254
```

```
##   left son=86 (1095 obs) right son=87 (632 obs)
##   Primary splits:
##       accel_forearm_x   < -90.5     to the right, improve=161.5773, (0 missing)
##       pitch_forearm     < 23.55     to the left,  improve=139.7361, (0 missing)
##       roll_dumbbell     < 42.85474  to the left,  improve=123.5807, (0 missing)
##       magnet_dumbbell_x < -89       to the right, improve=119.6957, (0 missing)
##       accel_dumbbell_x  < 95        to the right, improve=114.6502, (0 missing)
##   Surrogate splits:
##       pitch_forearm     < 23.25     to the left,  agree=0.875, adj=0.658, (0 split)
##       magnet_forearm_x  < -519.5    to the right, agree=0.836, adj=0.552, (0 split)
##       yaw_forearm       < 111.5     to the left,  agree=0.785, adj=0.413, (0 split)
##       accel_forearm_z   < -173.5    to the right, agree=0.684, adj=0.138, (0 split)
##       magnet_dumbbell_z < -9.5      to the right, agree=0.681, adj=0.128, (0 split)
##
## Node number 44: 1159 observations
##   predicted class=B  expected loss=0.230371  P(node) =0.07874711
##     class counts:    60   892    81    30    96
##    probabilities: 0.052 0.770 0.070 0.026 0.083
##
## Node number 45: 219 observations
##   predicted class=E  expected loss=0  P(node) =0.01487974
##     class counts:     0     0     0     0   219
##    probabilities: 0.000 0.000 0.000 0.000 1.000
##
## Node number 82: 1003 observations
##   predicted class=A  expected loss=0.1575274  P(node) =0.06814785
##     class counts:   845     1     0    57   100
##    probabilities: 0.842 0.001 0.000 0.057 0.100
##
## Node number 83: 3331 observations,    complexity param=0.02525396
##   predicted class=C  expected loss=0.6478535  P(node) =0.2263215
##     class counts:   352   700  1173   884   222
##    probabilities: 0.106 0.210 0.352 0.265 0.067
##   left son=166 (2840 obs) right son=167 (491 obs)
##   Primary splits:
##       accel_dumbbell_y < -40.5     to the right, improve=247.1387, (0 missing)
##       pitch_belt       < -42.55    to the left,  improve=197.4636, (0 missing)
##       roll_dumbbell    < -87.83843 to the left,  improve=175.0762, (0 missing)
##       pitch_dumbbell   < -25.831   to the left,  improve=153.9026, (0 missing)
##       accel_belt_y     < 56        to the left,  improve=145.7333, (0 missing)
##   Surrogate splits:
##       roll_dumbbell       < -91.12313 to the right, agree=0.957, adj=0.709, (0 split)
##       accel_belt_y        < 56        to the left,  agree=0.946, adj=0.635, (0 split)
##       pitch_belt          < 12.2      to the left,  agree=0.946, adj=0.633, (0 split)
##       magnet_dumbbell_y   < 399.5     to the left,  agree=0.920, adj=0.458, (0 split)
##       total_accel_forearm < 19.5      to the right, agree=0.881, adj=0.191, (0 split)
##
## Node number 84: 256 observations
##   predicted class=B  expected loss=0.4492188  P(node) =0.01739367
##     class counts:    27   141     0     0    88
##    probabilities: 0.105 0.551 0.000 0.000 0.344
##
## Node number 85: 1903 observations,    complexity param=0.0124371
##   predicted class=C  expected loss=0.4382554  P(node) =0.1292975
```

```
##     class counts:   164   161  1069   313   196
##    probabilities: 0.086 0.085 0.562 0.164 0.103
##   left son=170 (280 obs) right son=171 (1623 obs)
##   Primary splits:
##       magnet_dumbbell_z < 284.5     to the right, improve=151.62810, (0 missing)
##       magnet_forearm_z  < -248.5    to the left,  improve=151.45090, (0 missing)
##       magnet_forearm_y  < 842       to the right, improve=128.69310, (0 missing)
##       accel_forearm_z   < -167.5    to the right, improve= 85.85658, (0 missing)
##       magnet_belt_z     < -383.5    to the left,  improve= 83.22591, (0 missing)
##   Surrogate splits:
##       magnet_forearm_z < -248.5    to the left,  agree=0.912, adj=0.400, (0 split)
##       magnet_forearm_y < 842       to the right, agree=0.907, adj=0.368, (0 split)
##       accel_forearm_z  < 188.5     to the right, agree=0.894, adj=0.279, (0 split)
##       pitch_forearm    < -18.3     to the left,  agree=0.864, adj=0.079, (0 split)
##       magnet_arm_z     < 610.5     to the right, agree=0.860, adj=0.050, (0 split)
##
## Node number 86: 1095 observations,    complexity param=0.01376626
##   predicted class=E  expected loss=0.6547945  P(node) =0.0743987
##     class counts:    58   334   187   138   378
##    probabilities: 0.053 0.305 0.171 0.126 0.345
##   left son=172 (240 obs) right son=173 (855 obs)
##   Primary splits:
##       roll_dumbbell     < 41.01384  to the left,  improve=78.94146, (0 missing)
##       roll_forearm      < 132.5     to the left,  improve=77.20296, (0 missing)
##       accel_dumbbell_x  < 95.5      to the right, improve=76.37097, (0 missing)
##       magnet_dumbbell_x < -28.5     to the right, improve=76.27629, (0 missing)
##       magnet_arm_y      < 188.5     to the right, improve=73.78006, (0 missing)
##   Surrogate splits:
##       magnet_dumbbell_x < 39.5      to the right, agree=0.873, adj=0.421, (0 split)
##       pitch_dumbbell    < 77.13193  to the right, agree=0.872, adj=0.417, (0 split)
##       accel_dumbbell_x  < 105       to the right, agree=0.871, adj=0.412, (0 split)
##       yaw_dumbbell      < 106.5065  to the right, agree=0.859, adj=0.358, (0 split)
##       num_window        < 271.5     to the left,  agree=0.839, adj=0.267, (0 split)
##
## Node number 87: 632 observations
##   predicted class=D  expected loss=0.3322785  P(node) =0.04294062
##     class counts:    40    80    30   422    60
##    probabilities: 0.063 0.127 0.047 0.668 0.095
##
## Node number 166: 2840 observations,    complexity param=0.02525396
##   predicted class=D  expected loss=0.693662  P(node) =0.192961
##     class counts:   352   678   718   870   222
##    probabilities: 0.124 0.239 0.253 0.306 0.078
##   left son=332 (663 obs) right son=333 (2177 obs)
##   Primary splits:
##       roll_belt         < 125.5     to the right, improve=202.9925, (0 missing)
##       pitch_belt        < -42.55    to the left,  improve=171.1082, (0 missing)
##       magnet_dumbbell_y < 288.5     to the right, improve=143.2139, (0 missing)
##       yaw_dumbbell      < -93.58561 to the left,  improve=137.3968, (0 missing)
##       magnet_dumbbell_x < -551.5    to the left,  improve=119.8275, (0 missing)
##   Surrogate splits:
##       accel_belt_z     < -159.5    to the left,  agree=0.841, adj=0.320, (0 split)
##       total_accel_belt < 17.5      to the right, agree=0.828, adj=0.262, (0 split)
##       pitch_arm        < 40.75     to the right, agree=0.796, adj=0.127, (0 split)
```

```
##       accel_belt_x     < 49.5      to the right, agree=0.783, adj=0.069, (0 split)
##       yaw_arm          < 140       to the right, agree=0.776, adj=0.041, (0 split)
##
## Node number 167: 491 observations
##   predicted class=C  expected loss=0.07331976  P(node) =0.03336051
##     class counts:     0    22   455    14     0
##    probabilities: 0.000 0.045 0.927 0.029 0.000
##
## Node number 170: 280 observations
##   predicted class=A  expected loss=0.4714286  P(node) =0.01902432
##     class counts:   148    38    17    25    52
##    probabilities: 0.529 0.136 0.061 0.089 0.186
##
## Node number 171: 1623 observations
##   predicted class=C  expected loss=0.3518176  P(node) =0.1102731
##     class counts:    16   123  1052   288   144
##    probabilities: 0.010 0.076 0.648 0.177 0.089
##
## Node number 172: 240 observations
##   predicted class=B  expected loss=0.275  P(node) =0.01630656
##     class counts:    21   174     5    11    29
##    probabilities: 0.088 0.725 0.021 0.046 0.121
##
## Node number 173: 855 observations,    complexity param=0.01034843
##   predicted class=E  expected loss=0.5918129  P(node) =0.05809213
##     class counts:    37   160   182   127   349
##    probabilities: 0.043 0.187 0.213 0.149 0.408
##   left son=346 (162 obs) right son=347 (693 obs)
##   Primary splits:
##       roll_forearm  < 132.5     to the left,  improve=82.06835, (0 missing)
##       magnet_belt_z < -326.5    to the right, improve=65.40667, (0 missing)
##       accel_belt_z  < 33.5      to the right, improve=60.92641, (0 missing)
##       pitch_forearm < -7.42     to the left,  improve=60.66279, (0 missing)
##       magnet_arm_y  < 188.5     to the right, improve=59.12904, (0 missing)
##   Surrogate splits:
##       pitch_forearm        < -13.1     to the left,  agree=0.897, adj=0.457, (0 split)
##       total_accel_dumbbell < 31.5      to the right, agree=0.814, adj=0.019, (0 split)
##       accel_dumbbell_z     < -219.5    to the left,  agree=0.814, adj=0.019, (0 split)
##       gyros_dumbbell_z     < -2.125    to the left,  agree=0.813, adj=0.012, (0 split)
##       gyros_forearm_x      < -1.945    to the left,  agree=0.813, adj=0.012, (0 split)
##
## Node number 332: 663 observations,    complexity param=0.02174119
##   predicted class=C  expected loss=0.3936652  P(node) =0.04504688
##     class counts:     0   237   402    22     2
##    probabilities: 0.000 0.357 0.606 0.033 0.003
##   left son=664 (249 obs) right son=665 (414 obs)
##   Primary splits:
##       pitch_belt        < -42.6     to the left,  improve=272.90940, (0 missing)
##       num_window        < 539.5     to the left,  improve=236.89240, (0 missing)
##       yaw_belt          < 162.5     to the right, improve=178.74820, (0 missing)
##       magnet_dumbbell_z < 77.5      to the right, improve=104.77980, (0 missing)
##       yaw_dumbbell      < -92.37101 to the right, improve= 99.72175, (0 missing)
##   Surrogate splits:
##       num_window        < 539.5     to the left,  agree=0.917, adj=0.779, (0 split)
```

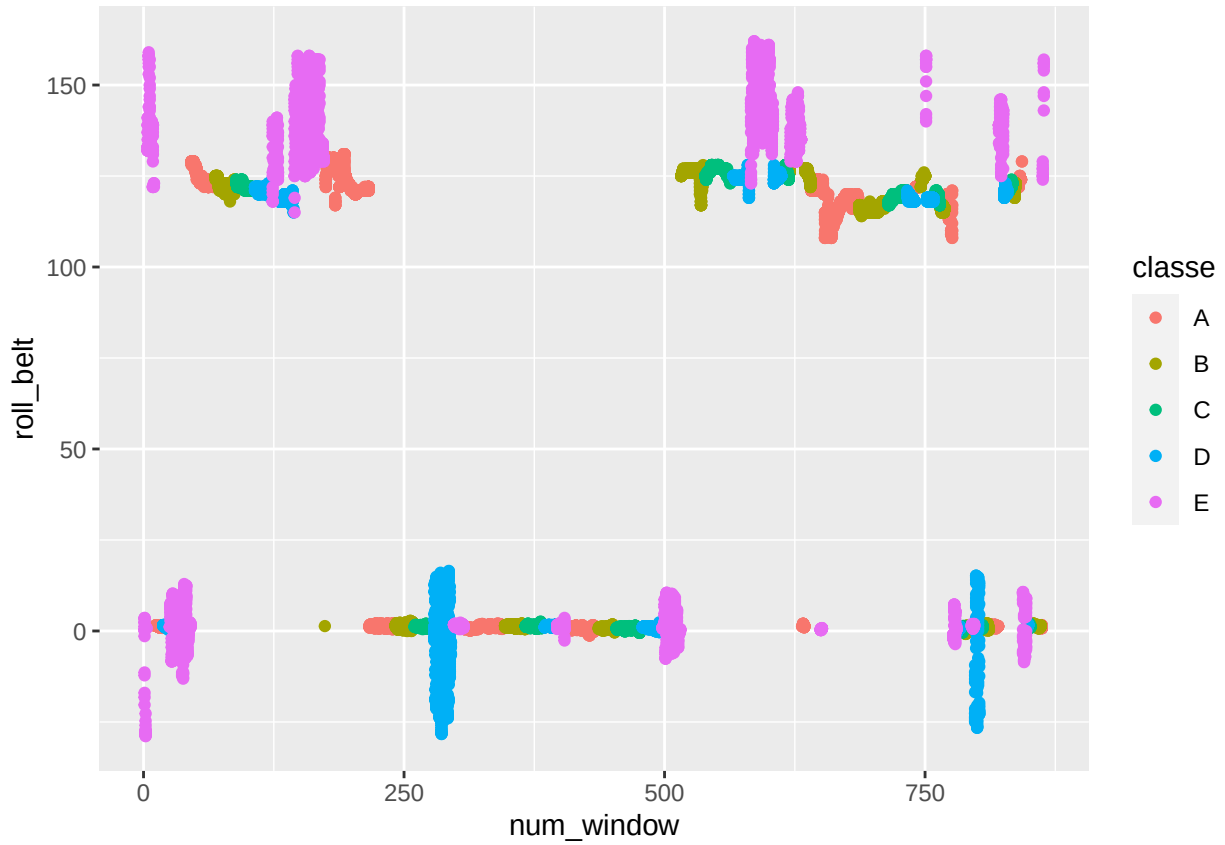
```
##       yaw_belt          < 162.5     to the right, agree=0.861, adj=0.631, (0 split)
##       yaw_dumbbell      < -92.37101 to the right, agree=0.822, adj=0.526, (0 split)
##       roll_dumbbell     < 55.38446  to the right, agree=0.821, adj=0.522, (0 split)
##       magnet_dumbbell_x < -546.5    to the right, agree=0.807, adj=0.486, (0 split)
##
## Node number 333: 2177 observations,    complexity param=0.01219975
##   predicted class=D  expected loss=0.6104731  P(node) =0.1479141
##     class counts:   352   441   316   848   220
##    probabilities: 0.162 0.203 0.145 0.390 0.101
##   left son=666 (248 obs) right son=667 (1929 obs)
##   Primary splits:
##       num_window       < 278       to the left,  improve=113.30790, (0 missing)
##       accel_dumbbell_z < 25.5      to the left,  improve=103.59170, (0 missing)
##       yaw_belt         < -87.65    to the left,  improve=102.85360, (0 missing)
##       roll_arm         < -3.16     to the right, improve= 98.82714, (0 missing)
##       roll_forearm     < 42.9      to the right, improve= 93.35556, (0 missing)
##   Surrogate splits:
##       gyros_dumbbell_x  < -1.13     to the left,  agree=0.904, adj=0.157, (0 split)
##       magnet_dumbbell_x < -61       to the right, agree=0.897, adj=0.093, (0 split)
##       accel_dumbbell_x  < 105       to the right, agree=0.893, adj=0.065, (0 split)
##       magnet_arm_z      < -534.5    to the left,  agree=0.892, adj=0.048, (0 split)
##       pitch_dumbbell    < 87.46726  to the right, agree=0.892, adj=0.048, (0 split)
##
## Node number 346: 162 observations
##   predicted class=C  expected loss=0.308642  P(node) =0.01100693
##     class counts:     3    44   112     0     3
##    probabilities: 0.019 0.272 0.691 0.000 0.019
##
## Node number 347: 693 observations
##   predicted class=E  expected loss=0.5007215  P(node) =0.0470852
##     class counts:    34   116    70   127   346
##    probabilities: 0.049 0.167 0.101 0.183 0.499
##
## Node number 664: 249 observations
##   predicted class=B  expected loss=0.07630522  P(node) =0.01691806
##     class counts:     0   230     1    16     2
##    probabilities: 0.000 0.924 0.004 0.064 0.008
##
## Node number 665: 414 observations
##   predicted class=C  expected loss=0.03140097  P(node) =0.02812882
##     class counts:     0     7   401     6     0
##    probabilities: 0.000 0.017 0.969 0.014 0.000
##
## Node number 666: 248 observations,    complexity param=0.01139277
##   predicted class=C  expected loss=0.483871  P(node) =0.01685012
##     class counts:     0   120   128     0     0
##    probabilities: 0.000 0.484 0.516 0.000 0.000
##   left son=1332 (120 obs) right son=1333 (128 obs)
##   Primary splits:
##       num_window        < 260       to the left,  improve=123.87100, (0 missing)
##       gyros_dumbbell_x  < -0.185    to the left,  improve= 85.42885, (0 missing)
##       magnet_dumbbell_x < -449      to the right, improve= 72.67097, (0 missing)
##       accel_dumbbell_z  < 10        to the right, improve= 68.90164, (0 missing)
##       yaw_dumbbell      < 33.00029  to the right, improve= 66.41188, (0 missing)
```

```
##   Surrogate splits:
##       gyros_dumbbell_x  < -0.185    to the left,  agree=0.911, adj=0.817, (0 split)
##       magnet_dumbbell_x < -456.5    to the right, agree=0.871, adj=0.733, (0 split)
##       accel_dumbbell_z  < 8.5       to the right, agree=0.859, adj=0.708, (0 split)
##       yaw_dumbbell      < -12.29108 to the right, agree=0.855, adj=0.700, (0 split)
##       roll_arm          < 82.35     to the right, agree=0.843, adj=0.675, (0 split)
##
## Node number 667: 1929 observations,    complexity param=0.01219975
##   predicted class=D  expected loss=0.560394  P(node) =0.131064
##     class counts:   352   321   188   848   220
##    probabilities: 0.182 0.166 0.097 0.440 0.114
##   left son=1334 (402 obs) right son=1335 (1527 obs)
##   Primary splits:
##       pitch_belt       < -42.45    to the left,  improve=116.20960, (0 missing)
##       yaw_belt         < 166.5     to the right, improve=102.94170, (0 missing)
##       num_window       < 297.5     to the right, improve=101.76610, (0 missing)
##       roll_forearm     < 42.9      to the right, improve= 97.88308, (0 missing)
##       accel_dumbbell_z < 32.5      to the left,  improve= 93.30422, (0 missing)
##   Surrogate splits:
##       yaw_belt        < 166.5     to the right, agree=0.965, adj=0.831, (0 split)
##       magnet_belt_x   < 166.5     to the right, agree=0.895, adj=0.495, (0 split)
##       pitch_arm       < 42.45     to the right, agree=0.850, adj=0.279, (0 split)
##       accel_forearm_x < -340.5    to the left,  agree=0.817, adj=0.122, (0 split)
##       accel_belt_x    < 50.5      to the right, agree=0.815, adj=0.112, (0 split)
##
## Node number 1332: 120 observations
##   predicted class=B  expected loss=0  P(node) =0.008153282
##     class counts:     0   120     0     0     0
##    probabilities: 0.000 1.000 0.000 0.000 0.000
##
## Node number 1333: 128 observations
##   predicted class=C  expected loss=0  P(node) =0.008696834
##     class counts:     0     0   128     0     0
##    probabilities: 0.000 0.000 1.000 0.000 0.000
##
## Node number 1334: 402 observations
##   predicted class=A  expected loss=0.6094527  P(node) =0.02731349
##     class counts:   157   137    75    28     5
##    probabilities: 0.391 0.341 0.187 0.070 0.012
##
## Node number 1335: 1527 observations
##   predicted class=D  expected loss=0.4629993  P(node) =0.1037505
##     class counts:   195   184   113   820   215
##    probabilities: 0.128 0.120 0.074 0.537 0.141
##
## n= 14718
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##    1) root 14718 10533 A (0.28 0.19 0.17 0.16 0.18)
##      2) roll_belt< 130.5 13474  9301 A (0.31 0.21 0.19 0.18 0.11)
##        4) pitch_forearm< -33.65 1190    12 A (0.99 0.01 0 0 0) *
##        5) pitch_forearm>=-33.65 12284  9289 A (0.24 0.23 0.21 0.2 0.12)
```

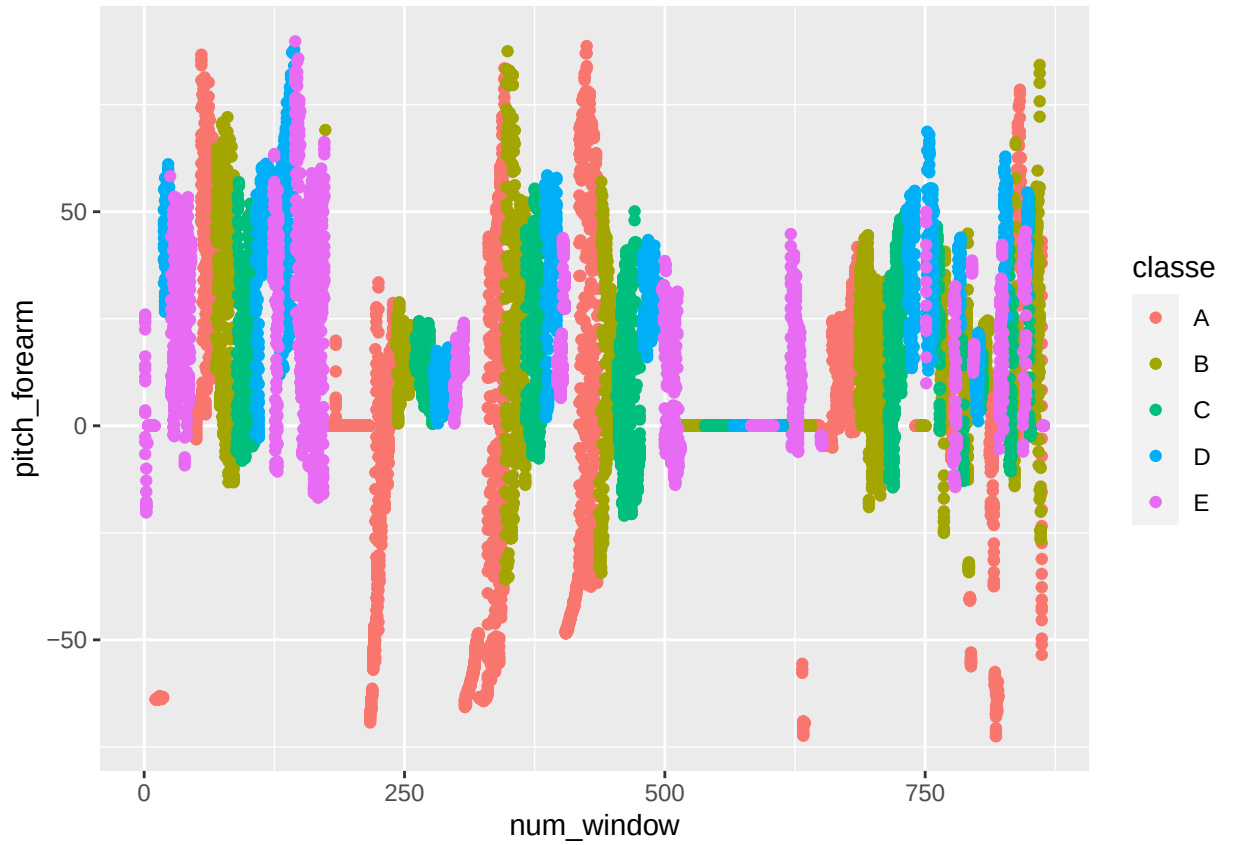
```
##         10) magnet_dumbbell_y< 438.5 10366  7431 A (0.28 0.18 0.24 0.19 0.11)
##           20) roll_forearm< 123.5 6480  3834 A (0.41 0.18 0.19 0.17 0.059)
##             40) magnet_dumbbell_z< -27.5 2146   697 A (0.68 0.21 0.012 0.075 0.028) *
##             41) magnet_dumbbell_z>=-27.5 4334  3137 A (0.28 0.16 0.27 0.22 0.074)
##               82) num_window< 241.5 1003   158 A (0.84 0.001 0 0.057 0.1) *
##               83) num_window>=241.5 3331  2158 C (0.11 0.21 0.35 0.27 0.067)
##                 166) accel_dumbbell_y>=-40.5 2840  1970 D (0.12 0.24 0.25 0.31 0.078)
##                   332) roll_belt>=125.5 663   261 C (0 0.36 0.61 0.033 0.003)
##                     664) pitch_belt< -42.6 249    19 B (0 0.92 0.004 0.064 0.008) *
##                     665) pitch_belt>=-42.6 414    13 C (0 0.017 0.97 0.014 0) *
##                   333) roll_belt< 125.5 2177  1329 D (0.16 0.2 0.15 0.39 0.1)
##                     666) num_window< 278 248   120 C (0 0.48 0.52 0 0)
##                       1332) num_window< 260 120     0 B (0 1 0 0 0) *
##                       1333) num_window>=260 128     0 C (0 0 1 0 0) *
##                     667) num_window>=278 1929  1081 D (0.18 0.17 0.097 0.44 0.11)
##                       1334) pitch_belt< -42.45 402   245 A (0.39 0.34 0.19 0.07 0.012) *
##                       1335) pitch_belt>=-42.45 1527   707 D (0.13 0.12 0.074 0.54 0.14) *
##                 167) accel_dumbbell_y< -40.5 491    36 C (0 0.045 0.93 0.029 0) *
##           21) roll_forearm>=123.5 3886  2600 C (0.074 0.18 0.33 0.22 0.19)
##             42) magnet_dumbbell_y< 281.5 2159  1090 C (0.088 0.14 0.5 0.14 0.13)
##               84) num_window< 88.5 256   115 B (0.11 0.55 0 0 0.34) *
##               85) num_window>=88.5 1903   834 C (0.086 0.085 0.56 0.16 0.1)
##                 170) magnet_dumbbell_z>=284.5 280   132 A (0.53 0.14 0.061 0.089 0.19) *
##                 171) magnet_dumbbell_z< 284.5 1623   571 C (0.0099 0.076 0.65 0.18 0.089) *
##             43) magnet_dumbbell_y>=281.5 1727  1167 D (0.057 0.24 0.13 0.32 0.25)
##               86) accel_forearm_x>=-90.5 1095   717 E (0.053 0.31 0.17 0.13 0.35)
##                 172) roll_dumbbell< 41.01384 240    66 B (0.088 0.72 0.021 0.046 0.12) *
##                 173) roll_dumbbell>=41.01384 855   506 E (0.043 0.19 0.21 0.15 0.41)
##                   346) roll_forearm< 132.5 162    50 C (0.019 0.27 0.69 0 0.019) *
##                   347) roll_forearm>=132.5 693   347 E (0.049 0.17 0.1 0.18 0.5) *
##               87) accel_forearm_x< -90.5 632   210 D (0.063 0.13 0.047 0.67 0.095) *
##         11) magnet_dumbbell_y>=438.5 1918   947 B (0.031 0.51 0.043 0.23 0.19)
##           22) total_accel_dumbbell>=5.5 1378   486 B (0.044 0.65 0.059 0.022 0.23)
##             44) roll_belt>=-0.58 1159   267 B (0.052 0.77 0.07 0.026 0.083) *
##             45) roll_belt< -0.58 219     0 E (0 0 0 0 1) *
##           23) total_accel_dumbbell< 5.5 540   134 D (0 0.15 0.0019 0.75 0.1) *
##      3) roll_belt>=130.5 1244    12 E (0.0096 0 0 0 0.99) *
```

The above list shows the ranking of variables in our GBM. We see that num_window, roll_belt, and pitch_forearm are the most performant ones. We can checkout a few plots demonstrating their power:

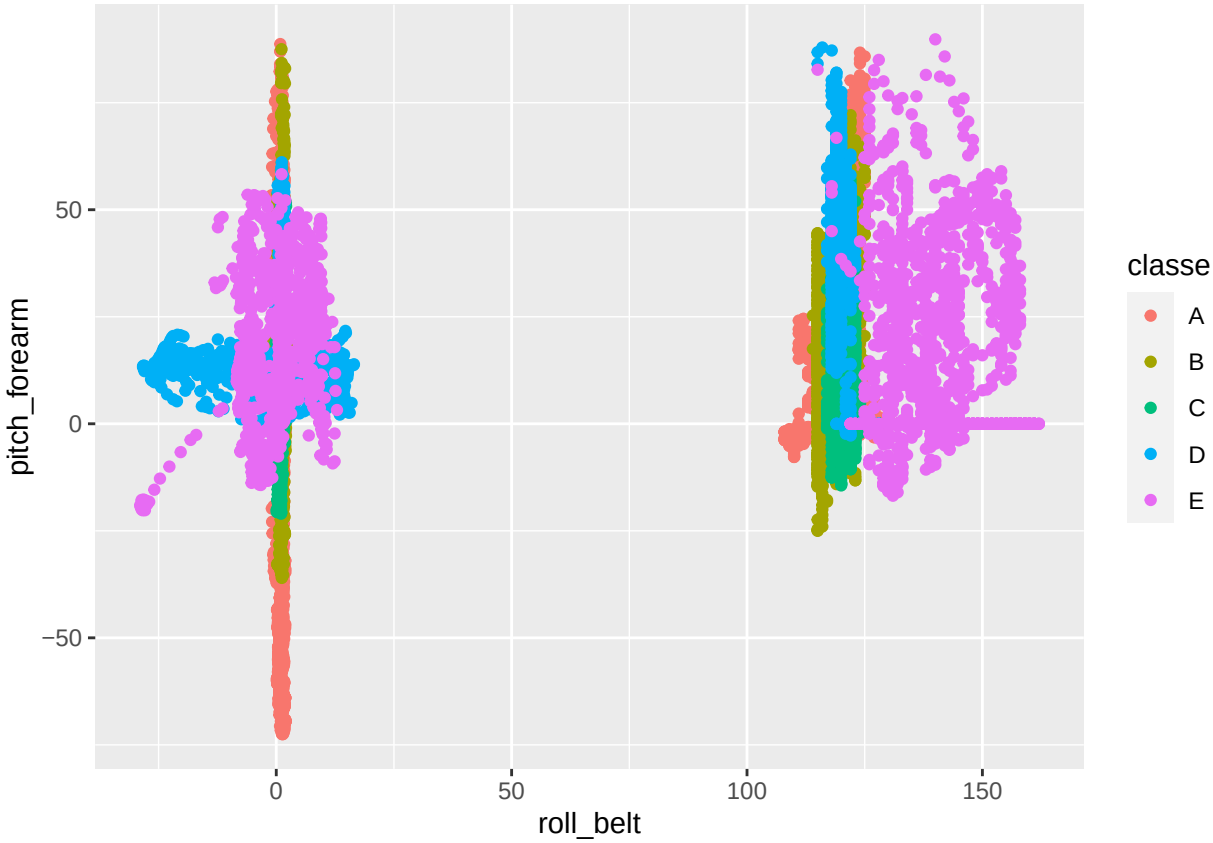
```
qplot(num_window, roll_belt    , data = train_set, col = classe)
```




```
qplot(num_window, pitch_forearm, data = train_set, col = classe)
```




```
qplot(roll_belt , pitch_forearm, data = train_set, col = classe)
```


pitch_forearm
50
0
−50
0
50
100
150
roll_belt
classe
A
B
C
D
E