Prediction Assignment Writeup for JHU/Coursera Practical Machine Learning Class

dandybits

Synopsis: This project involves analyzing of the Weight Lifting Exercises (WLE) Dataset. This document describes the approach for building classification model that allows distinguishing properly conducted weight lifting exercise movements from those conducted with common mistakes.

This research was conducted as a test assignment for the Data Science Certification on Coursera.. The code for this assignment is available on Github

For more information about the collection and the original analysis of the WLE dataset see research article Qualitative Activity Recognition of Weight Lifting Exercises by Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H.

Exploring the WLE dataset The first step in data analysis is loading the data. The dataset provided by COursera is a subset of the dataset made publicly available by the authors of the original article.

```
## load libraries
library(lattice)
library(ggplot2)
library(caret)
library(rpart)

## load data
wle.data <- read.csv("data\\pml-training.csv")
wle.grade <- read.csv("data\\pml-testing.csv")
```

Observation notes on WLE dataset As described in the paper referenced above as well as revealed in exploring the data, the dataset contains records with various levels of granularity. There are 'timestamp'-level records that contain a set of sensor measurements as well as 'summary' records that contained averaged measurements for time windows of several sizes from 0.5 sec to 2.5 sec. This makes the task somewhat ambiguous. We are trying to predict if a record belongs to a properly executed movement while any meaningful classification only applies at the level of the entire set of records for a particular movement.

Moreover, since the surrogate identifier for the movement, num_window attribute, is present in the test data set, it is possible to predict based on the num_window attribute alone.

While this may seem trivial, similar 'over_inclusive' datasets occasionally caused unintended results even in high-profile ML competitions.

```
## splitting data for model validation
set.seed(130265)
inTrain <- createDataPartition(wle.data$classe, p = 0.7, list = FALSE)
wle.train <- wle.data[inTrain,]
wle.test <- wle.data[-inTrain,]
fit.winonly.rpart <- rpart(classe ~ num_window, data=wle.train, method = "class", cp = 0.0025)
```

Predicting based on window_num only The aboove approach gave 100% accurate results on the prediction quiz.

```
predict(fit.winonly.rpart, wle.grade, type = "class")
```

```
##  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
```

The approach also works fairly well on the allocated test set

```
wle.test.pred.winonly.rpart <- predict(fit.winonly.rpart, wle.test, type = "class")
conf.winonly <- confusionMatrix(wle.test$classe, wle.test.pred.winonly.rpart)
conf.winonly$table
```

```
##           Reference
## Prediction    A    B    C    D    E
##          A 1664   10    0    0    0
##          B    0 1129   10    0    0
##          C    0    0 1026    0    0
##          D    0    0    0  964    0
##          E   13    3    0    8 1058
```

generating good prediction with overall accuracy 0.9925234.

Reducing cp argument above (from its default value of 0.01) leads to better accuracy traded for the larger tree size and the increased processing time.

Predicting based on sensor measurements It is expected that we can get an accurate prediction for the classe variable if we properly select a set of predictors and fit a model using one of the popular methods. However, the nature of the dataset makes me doubt that it can by ifself provide a sufficient material for 'learning' data patterns. The main deficiency is that the dataset is extremely limited and it prominently carries characterisitcs of participants. While it is not an intended goal of the assignment, it can be shown that almost any small set of variables with almost any model can reliably predict a participant.

This is illustarted below by fitting a random forest model on a set of sensor measurements selected without any particular startegy. Nevertheless, the accuracy of the participant identification, regardless of the quality of the movement, is very high.

```
fit.rf.user.wle.train <- train(user_name ~ magnet_arm_x +
                                           magnet_arm_y +
                                           magnet_arm_z +
                                           magnet_belt_x +
                                           magnet_belt_y +
                                           magnet_belt_z +
                                           magnet_dumbbell_x +
                                           magnet_dumbbell_y +
                                           magnet_dumbbell_z +
                                           magnet_forearm_x +
                                           magnet_forearm_y +
                                           magnet_forearm_z,
                                data = wle.train, method = "rf", ntree = 20)
```

```
## Loading required package: randomForest
## randomForest 4.6-12
## Type rfNews() to see new features/changes/bug fixes.
```

```
pred.rf.user.wle.test <- predict(fit.rf.user.wle.train, wle.test)
conf.rf.user <- confusionMatrix(pred.rf.user.wle.test, wle.test$user_name)
conf.rf.user$table
```

```
##           Reference
## Prediction adelmo carlitos charles eurico jeremy pedro
##   adelmo     1120        0       0      0      0     0
##   carlitos      0      949       0      0      0     0
##   charles       0        1    1073      0      0     0
##   eurico        0        0       0    929      0     0
##   jeremy        0        0       0      0   1019     0
##   pedro         0        0       0      0      0   794
```

```
conf.rf.user$overall[[1]]
```

```
## [1] 0.9998301
```

Research alternatives Authors of the original article bring up some arguments favoring usage of Kinect technology over ML approach. However, qualitative analysis of the weight lifting and other fitness training exercises using machine learning can be made more effective with some changes to the approach taken by the authors of the article.

First, the granularity of the data collection can be changed to include aggregate characteristics of one or several repetitions or corresponding time series, which are likely to be more representative than single-timestamp measurements.

Second, the data collection can be applied at a much larger scale. Given that one of the objectives of the research was a prevention of the sport injuries, it is important to collect data on exercises that led to injuries and use that data for model training. Obviously, it will be unethical to conduct a study where participants are injured intentionally, so the data collection can only be conducted in an observational study.

Related to the scale of data collection is the need to generalize the data set and the model to a wider variety of body types and athletic abilities. This may in turn require a non-trivial approach to normalizing the data.

I believe the above measures can positively affect the efficiency of the qualitative analysis of fitness training exercises using machine learning methods.