1. Executive summary

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.

2. Loading the data and libraries

Loading the data directly from the given URLs, for reproducibility.

```r
dTraining<-read.csv("https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv")
dTesting<-read.csv("https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv")
```

3. Selecting data columns

Based on exploring the dataset, the conclusion was made that many of the columns have large amount of missing data. Such columns were excluded.
Columns with date, time, user information was also excluded.

```r
set.seed(8524)

#Creating training dataset
inTraining = rbinom(length(dTraining), size = 1, prob = 0.5)
validation<-dTraining[inTraining==0,] ;   training<-dTraining[inTraining==1,]

#Deleting columns with high rate of missing data
training.missingData<-sapply(training,function(x) (sum((is.na(x) | x==""))/length(x)))
selectedCols<-training.missingData<0.8
#knitr::kable(training.missingData)

validation<-validation[,(selectedCols)] ; training<-training[,(selectedCols)]
validation<-validation[,-(1:7)] ;         training<-training[,-(1:7)]
```

3. Explorative analysis

Exploring the remaining parameters shows, that they have variance and unique values.

```r
training.summary<-
   data.frame(var=sapply(training,function(x) ifelse(is.factor(x),NA,as.character(var(x)))),
              unique=sapply(training,function(x) length(unique(x)))
              )
knitr::kable(training.summary)
```

5. ML Models

training based on multiple methods, which

```
model.1  <- svm(classe ~ . , data=training, trControl=trCntrl)
model.2 <- train(classe ~ . , data=training, method="treebag", trControl=trCntrl)
invisible(capture.output(model.3 <- train(classe ~ . , data=training, method="gbm", trControl=trCntrl))
model.4 <- train(classe ~ . , data=training, method="lda", trControl=trCntrl)
model.5 <- train(classe ~ . , data=training, method="rpart", trControl=trCntrl)
model.6 <- train(classe ~ . , data=training, method="rf", trace=F, trControl=trCntrl)
```

```
pred1<-predict(model.1,validation) ; pred1.i<-predict(model.1,training)
pred2<-predict(model.2,validation) ; pred2.i<-predict(model.2,training)
pred3<-predict(model.3,validation) ; pred3.i<-predict(model.3,training)
pred4<-predict(model.4,validation) ; pred4.i<-predict(model.4,training)
pred5<-predict(model.5,validation) ; pred5.i<-predict(model.5,training)
pred6<-predict(model.6,validation) ; pred6.i<-predict(model.6,training)
```

6. Cross validation

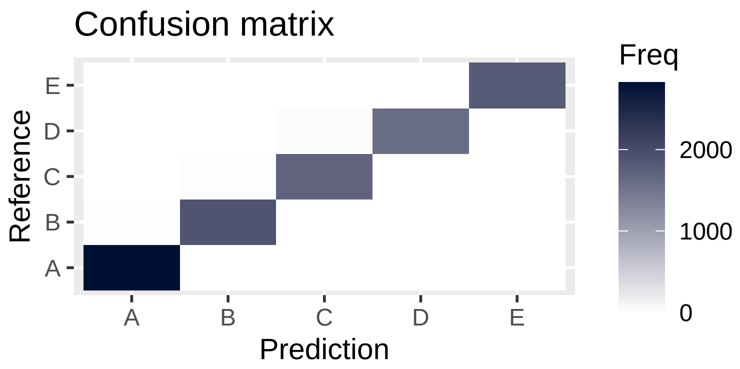
```
INaccuracy<- c(
      svm=confusionMatrix(pred1.i,training$classe)$overall['Accuracy'],
      treebag=confusionMatrix(pred2.i,training$classe)$overall['Accuracy'],
      gbm=confusionMatrix(pred3.i,training$classe)$overall['Accuracy'],
      lda=confusionMatrix(pred4.i,training$classe)$overall['Accuracy'],
      rpart=confusionMatrix(pred5.i,training$classe)$overall['Accuracy'],
      rf=confusionMatrix(pred6.i,training$classe)$overall['Accuracy']
)
OUTaccuracy<- c(
      svm=confusionMatrix(pred1,validation$classe)$overall['Accuracy'],
      treebag=confusionMatrix(pred2,validation$classe)$overall['Accuracy'],
      gbm=confusionMatrix(pred3,validation$classe)$overall['Accuracy'],
      lda=confusionMatrix(pred4,validation$classe)$overall['Accuracy'],
      rpart=confusionMatrix(pred5,validation$classe)$overall['Accuracy'],
      rf=confusionMatrix(pred6,validation$classe)$overall['Accuracy']
)
knitr::kable(data.frame(InSampleAccuracy=INaccuracy,OutOfSampleAccuracy=OUTaccuracy))

confmat.rf<-confusionMatrix(pred6,validation$classe)
```

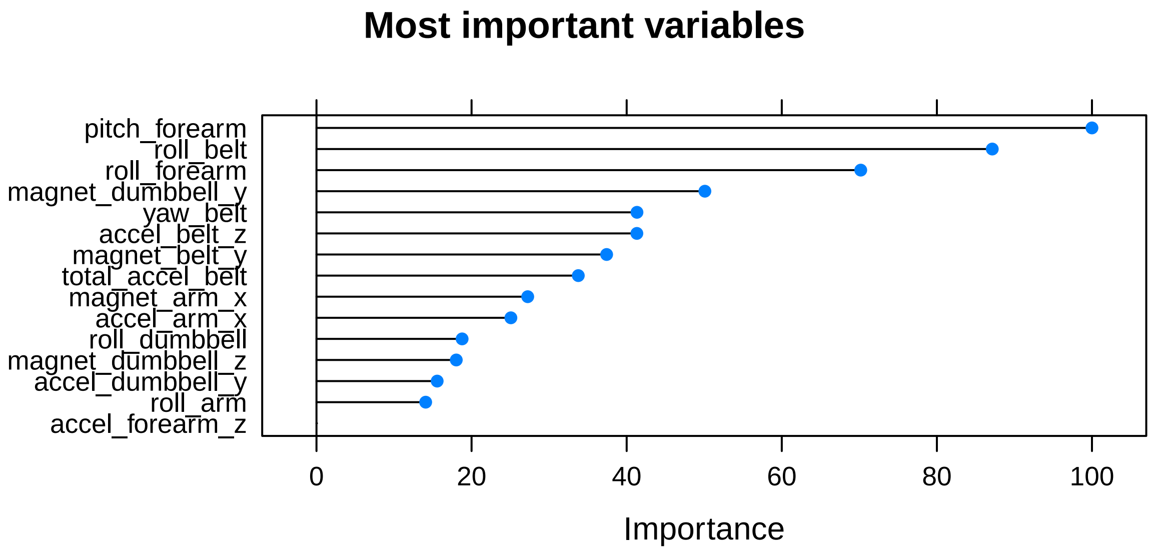
	InSampleAccuracy	OutOfSampleAccuracy
svm.Accuracy	0.9427304	0.9384755
treebag.Accuracy	1.0000000	0.9860034
gbm.Accuracy	0.9749252	0.9610311
lda.Accuracy	0.7021979	0.7022455
rpart.Accuracy	0.4953049	0.4971302
rf.Accuracy	1.0000000	0.9916423

7. Analysis of the most accurate model (random forest)

```r
library(ggplot2)
ggplot(as.data.frame(confmat.rf$table),aes(x=Prediction,y=Reference, fill=Freq)) + geom_tile() + scale_
```



```r
plot(varImp(model.5), top=15, main="Most important variables")
```



7. Prediction on the test data

```r
solution<- data.frame(ID=dTesting$X, Prediction=predict(model.6,dTesting))
knitr::kable(solution)
```

ID	Prediction
1	B
2	A
3	B

ID	Prediction
4	A
5	A
6	E
7	D
8	B
9	A
10	A
11	B
12	C
13	B
14	A
15	E
16	E
17	A
18	B
19	B
20	B

00. Assignment info

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.

Review Criteria

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.

Peer Review Portion Your submission for the Peer Review portion 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 :-).

Course Project Prediction Quiz Portion Apply your machine learning algorithm to the 20 test cases available in the test data above and submit your predictions in appropriate format to the Course Project Prediction Quiz for automated grading.

Reproducibility

Due to security concerns with the exchange of R code, your code will not be run during the evaluation by your classmates. Please be sure that if they download the repo, they will be able to view the compiled

HTML version of your analysis.

Prediction Assignment Writeup

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://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset).

Data

The training data for this project are available here:

https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv

The test data are available here:

https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

The data for this project come from this source: http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har. 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.