Practical Machine Learning Project

How I built the model and why I made these choices

For this project I wanted to use a simple random forest model because they are typically quite powerful. Given the computational cost of random forest, I needed to decrease the number of predictors as much as possible (without sacrificing accuracy). First, I removed predictors that included mostly NAs. Second, I removed the first seven predictors (which included information that was not predictive, like subject names). Finally, I removed the all factor predictors, many of which had lots of missing data and looked quite strange, except the outcome variable (classe).

How you used cross-validation

For this project I used a simple cross-validation strategy of splitting the training file in two parts, with 75% allocated to a training set and 25% allocated to a testing set. I held out the entire testing file as a validation set.

What you think the expected out of sample error is

The accuracy of this model was 0.9939 and it predicted all 20 items in the validation set correctly. Therefore, I believe the out of sample error to be very low.

```
library(caret)

# Load the data
original = read.csv("./pml-training.csv")
validation  = read.csv("./pml-testing.csv")

# Drop predictors with mostly NAs
sm <- original[,colSums(is.na(original)) == 0]

# Drop the predictors that are meaningless for this task
sm <- sm[ -c(1:7) ]

# Drop weird and inconsistent factor predictors
sm$classe <- as.character(sm$classe)
sm <- sm[, !sapply(sm, is.factor)]
sm$classe <- as.factor(sm$classe)

# Partition the testing data into a training and testing set
inTrain = createDataPartition(sm$classe, p = 3/4, list=FALSE)
training = sm[ inTrain,]
testing = sm[-inTrain,]

# Fit the random forest model
# This takes a VERY LONG time
fit1 <- train(classe ~ ., method="rf", data = training)
pred <- predict(fit1, testing)

# Check out the accuracy of the model
confusionMatrix(pred, testing$classe)
# Accuracy: 0.9939

# Use the model to predict the classification of
# the 20 new items in the validation set
predNew <- predict(fit1, validation)

# Output the predictions
predNew
# 20 of 20 predictions are correct
```