Project

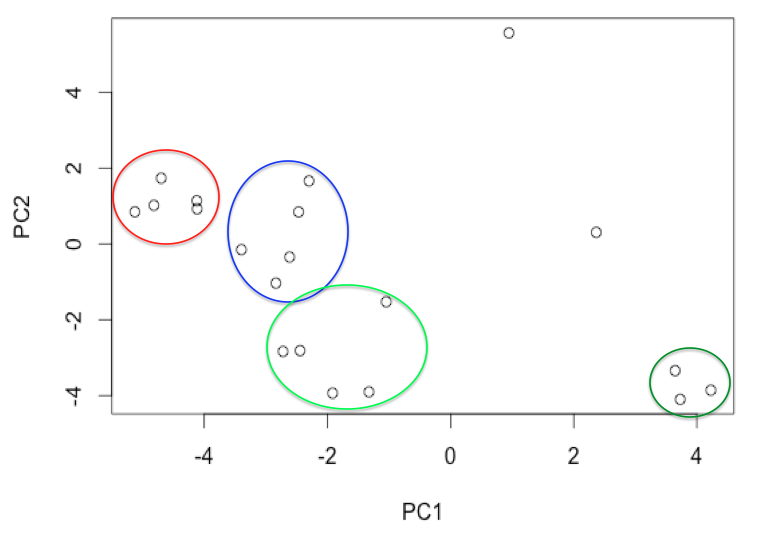
Preprocessing the data

The data contains 19,622 observations and 160 variables. Using the summary function I identified many variables with 19,216 observations (approx. 98%) that were missing and removed these variables since there is not enough valid information for data imputation. This then reduced the dimension down to 61 variables.

Other considerations were that the classe variable is evenly distributed so there is no chance of any of the bootstrapping techniques like random forests suffering from taking the majority class, thus no adjustments are needed with regards under/over sampling.

PCA

To get some idea of what is going on with the data I did a PCA analysis to see whether there were any distinct groupings that could identify the classe variable. Plotting the training data showed no clearly identifiable groups due to the confines of two-dimensional, though applying the model used with the training data on the test data produces the following plot:



There are four distinct groupings seen in the plot above with two points not associated with any other clusters.

Data partitioning

Given the size of the data and the inability of some of the models to run (timed out) with so many variables and observations, I partitioned the data based on the time variable so that I got a sample of ½ of the data from each day to give a balanced sample. The assignment of observations to the classe variable was also checked to make sure that there was a similar balance in the partitioned data and the training data.

CART model

To initially analyse the groupings the basic classification tree model was run. I ran a cross validation with k = 50 and the resulting model used 4 nodes to classify the training data. These 4 nodes were for the variables roll belt (< 130.5 and ≥ 130.5), pitch forearm (< -33.95 and ≥ -33.95), magnet dumbell (< 439.5 and ≥ 439.5) and roll forearm (<123.5 and ≥ 123.5). The cross validation used the accuracy for selecting the best model and the complexity parameter was 0.039 with the acccuracy being 52.7%. Classifying the test set using the model trained on the training set, 11 observations were classified as being in A and 9 were classified as being in C. This seems like there are several points that are being misclassified combined with a low accuracy that is just above that of random chance.

Linear Discriminant Analysis

The linear discriminant model with cross validation and k = 50 gave an accuracy of 75.0% accuracy and predicted the following spread of points: A = 9 observations, B = 6 observations, C = 2 observations, D = 3 observations and E = 0 observations. There seem to be consistencies with the the amount of observations in A.

Boosting trees

Running the boosting tree method with cross validation using k = 50 gave a model with 150 trees, an interaction depth of 3 and an accuracy of 97.8%. This model predicted the following classes: A = 7 observations, B = 8 observations, C = 1 observation, D = 1 observation and E = 3 observations. This matches closest with what was seen with the PCA classification.

Support Vector Machines

Running a linear model with cross validation and k = 50 gave an accuracy of 81.0%. This classified the test set with the following classes: A = 8 observations, B = 6 observations, C = 1 observation, D = 2 observation and E = 3 observations.

Naïve Bayes

The Naïve Nayes method with cross validation and k = 50 gave an accuracy of 74.4%. This classified the test set with the following classes: A = 13 observations, B = 3 observations, C = 1 observation, D = 1 observation and E = 2 observations.

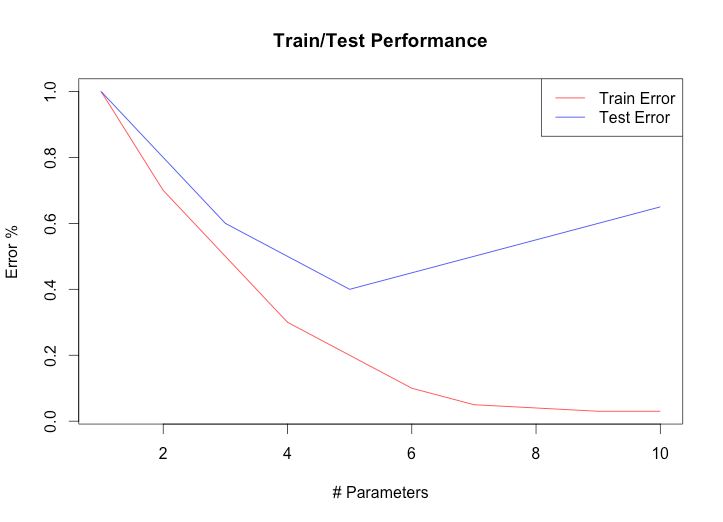
Random forest

The random forest model wouldn’t run, though this may be due to taking sampling with replacement for such a large dataset.

Summary

From the various models that were run I was able to get a measure of the accuracy of the respective models for the training data that they were run on, which is the error appropriate with cross validation.

The results were similar to what I expected where I expected the error appropriate with cross validation to be between 75% and 90% with the out of sample error being similarly within the same interval. This is better seen visually in the figure below:



Before checking the test data I would have used either the LDA or linear SVM to classify the data based on the lack of assumptions required for these methods (the Naïve Bayes requires independence between features). These two models used all the variables as I was unable to use variable importance associated with the random forest model or determine from the PCA what were the key variables in determining the first two components.

All the models run in this project were used to predict the respective classes for the test set and the details of their accuracy on the test set are given in the table below. The out of sample error and any overfitting between the training and test sets can be seen quite clearly. Initially I expected the boosting trees model that has an accuracy of 97.8% to be highly likely in overfitting the test data. However, although I submitted the SVM linear model with accuracies of 81% and 80% for the training and test datasets respectively, the boosting tree model had better accuracies with 98% and 100%. One can conclude from this that the boosting nature of the model was able to strengthen the weak classifiers and build on strong classifiers.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Observation | CART | LDA | Boosting trees | Naïve Bayes | SVM Linear | PCA | Test classification |
| 1 | C | C | B | A | C | Group 1 | B |
| 2 | A | A | A | A | A | Group 3 | A |
| 3 | C | B | B | A | B | Group 3 | B |
| 4 | A | A | A | A | A | Group 5 | A |
| 5 | A | A | A | A | A | Group 2 | A |
| 6 | C | C | E | E | E | Group 3 | E |
| 7 | C | D | D | D | D | Group 6 | D |
| 8 | A | D | B | C | D | Group 6 | B |
| 9 | A | A | A | A | A | Group 6 | A |
| 10 | A | A | A | A | A | Group 4 | A |
| 11 | C | D | B | A | B | Group 6 | B |
| 12 | C | A | C | A | A | Group 3 | C |
| 13 | C | B | B | B | B | Group 2 | B |
| 14 | A | A | A | A | A | Group 3 | A |
| 15 | C | B | E | E | E | Group 6 | E |
| 16 | A | A | E | A | E | Group 2 | E |
| 17 | A | A | A | A | A | Group 1 | A |
| 18 | A | B | B | B | B | Group 2 | B |
| 19 | A | B | B | A | B | Group 1 | B |
| 20 | C | B | B | B | B | Group 2 | B |
| Accuracy on the training data | 52.7% | 75.0% | 97.8% | 74.4% | 81.0% |  |  |
| Accuracy on the test data | 40% | 65% | 100% | 65% | 80% |  |  |