Machine Learning Course Project

K. Divis

Executive Summary

This project utilizes the Human Activity Recognition (HAR) datset to predict the the quality of the weight lifting exercise being performed based on data from wearable accelerometers. A model utilizing random forest prediction with 52 measures from the wearable accelerometers performed well (99.9% accuracy from cross validation; 100% accuracy on the given test set of 20 items). Furthermore, this model optimized both accuracy and speed relative to other motels.

(This report was completed as part of (and under the guidelines of) the Johns Hopkins Data Science specialization Practical Machine Learning class on Coursera.)

Introduction to Data

The HAR data tests a unilateral dumbbell biceps curl. It includes measurements (e.g., pitch) from accelerometers and quality of exercise performed (exactly according to specification, throwing the elbows to the front, lifting the dumbbell only halfway, lowering the dumbbell only halfway, or throwing hips to the front). Additionally, the data includes aggregate variables across time for the accelerometer measures and bookmarking data such as participant name.

Cleaning/Exploration

Since the final testing data set does not take time into account (e.g., how movement changes over time), all of the variables that were aggregates over time were removed (e.g., kurtosis; please note by nature these were the columns that contained many NAs). Furthermore, any bookmarking variables (e.g., participant name or window) were removed. The ultimate goal is predict *future* performance (regardless of who performs it and without the built in convenicne of these bookmarking variables). After this clearning, I was left with 52 variables plus the variable to be predicted (classe).

In an effort to understand the structure of the remaining variables, I checked for linear combinations, near zero variance, and correlations. While linear combinations and near zero variance were not major players in this data, it looked like collinearity might be an issue and should be kept in mind when model building if issues arise during cross validation.

Model building

To prepare for cross validation, I first split the original training set into a new, smaller training set and a withheld testing set for cross validation purposes. The new training set was composed of 75% of the original training set; the withheld testing set was composed of the remaining 25% of the original training set.

Professor Leek frequently mentioned that Random Forest and Bootstrapping models often perform the best. I tried a simple linear regression model with the first three principal components to prove its inadequacy in this setting to myself. Given the distribution of some of the variables, I was not surprised that it performed poorly (it had about 50% accuracy). Satisfied, I moved onto a random forest method. The linear models took a very long time to run; my first goal was to create a random forest model that would (1) perform well and (2) run quickly. A random forest model using the out-of-bag (OOB) estimate method with all 52

predictor variables met both of these criteria, running in less than 10 minutes on my machine and reaching 99.9% accuracy when compared to the withheld testing set. Figure 1 highlights this model.

99.9% accuracy in less than 10 minutes was a nice speed-accuracy tradeoff; however, I was curious if I could reach an even higher level of accuracy with a modified random forest model. While the oob estimate works for random forest models, I also wanted to look at the more straightforward cross validation (cv) method, using 3 resamples and all 52 predictor variables. This model took far longer (45 minutes on my machine) and performed no better (accuracy=99.9%).

Cross validation and expected out of sample error

In order to use cross validation to calculate the expected out of sample accuracy and error rates, the original training data was split (75%-25%) into a training set and withheld testing set. I used these two sets to test the various models I tried and only applied the final model to the original test set of 20 trials. Because the random forest model using the oob method and all 52 predictors was the final model I chose, I will only go into the details of it. The overall accuracy was 99.9%. The out of sample error rate (estimated by comparing the predictions of the model fit on the withheld training set to the withheld testing set) was 0.1%.

Conclusions and Fit to Final Test Set

Since the random forest model with the oob method optimized accuracy and speed, I fit it to all of the original training data and then applied that fit to the final test data. Based on the prior cross validation accuracy, I expected good performance. As anticipated, all 20 test cases were correctly predicted by this model (100% on Prediction Assignment Submission aspect of this project). Looking at variable importance (see Figure 2), the 52 predictors could be eventually be trimmed down further. The roll and yaw of the belt appeared to be the most important predictors for quality of the exercise.

Code and References

Please note that all code can be found in the .Rmd file in the GitHub repository where this file is hosted; it was excluded from this document for readibility. Information on the HAR dataset can be found at http://groupware.les.inf.puc-rio.br/har

---

Figure 1: Model Fit

Fit of random forest model utilizing oob method

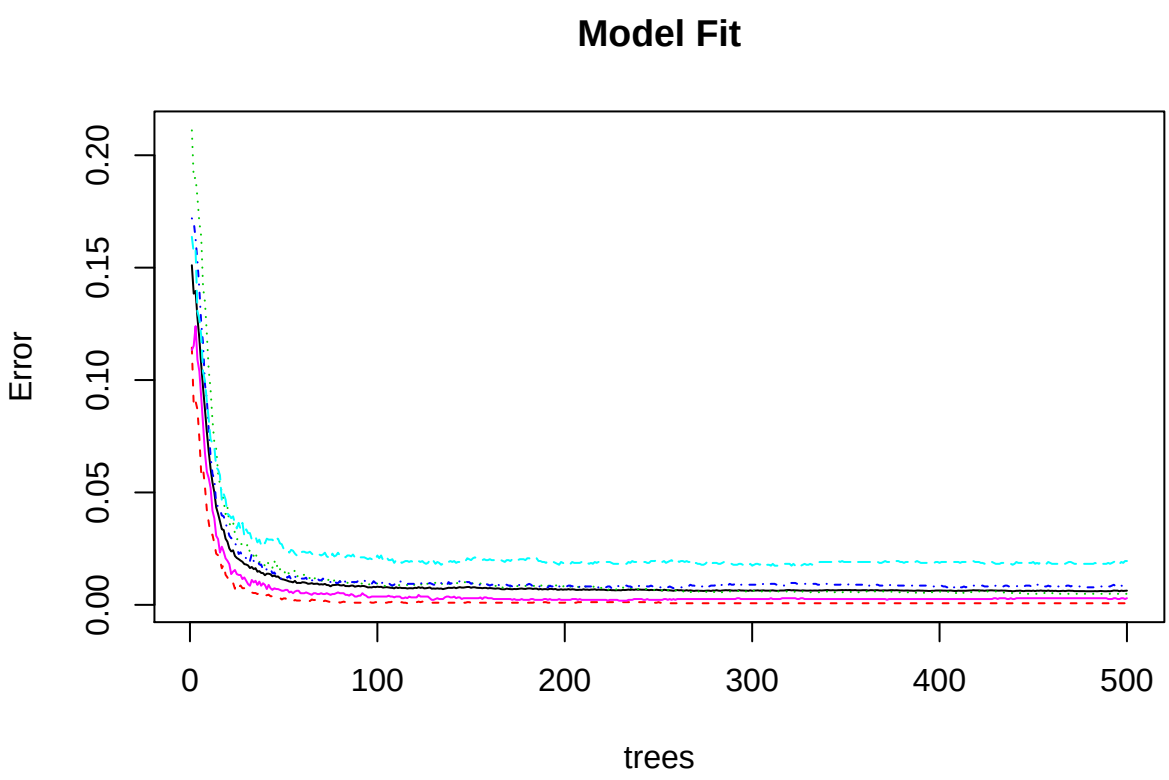
Model Fit
Error
0.00
0.05
0.10
0.15
0.20
0
100
200
300
400
500
trees

Figure 2: Variable Importance

Variable importance for the random forest model utilizing the oob method