Final Project - Course 8

John Theodore

July 30, 2019

Introduction

This final project seeks to discover the best machine learning algorithm that will predict the level of "correctness" in performing a specific weight lifting exercise. The sample data is collected using electronic measuring devices (e.g., accelerometers) strapped on a person during the exercise. Further explanation is detailed in the assignment notes:

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 types 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).

For this project, the specific goals are:

- **Goal 1** - Develop a ml model that predicts the manner (level of 'correctness"") in which participants did the specific dumbbell exercise (represented as the "classe" variable in the training set).
- **Goal 2** - Use this "best"" model to predict "classe" on 20 different test cases.

Executive Summary

Goal 1 - Best Model

Conducting a variety of different algorithms (both individual and ensemble models) on the training data, the model with the best "accuracy" was from an Extreme Gradient Boosting model that produced a mean accuracy on the testing data (hold-out subset from the "pmltraining" data) of .9937. The tuning parameters of this specific model are:

- Extreme Gradient Boost Model * - ** Accuracy: .9937 **
- nrounds = 600
- max_depth = 6
- eta = .1
- gamma = 0
- colsample_bytree = .4
- min_child_weight = 1
- subsample = 1

This model used 45 features (p=45) reduced from the original training data of 159 *potential* features.

Goal 2 - Quiz Assigment

Using the best model on the supplied 20 records ("pmltesting"), the following predictions were made:

[1] E B E E A E E E B E E E B E E B E B E B Levels: A B C D E ### Model Development Process

Model development was conducted in four steps:

1. Data import and preprocessing
2. Feature reduction
3. Individual model/algorithm testing
4. Ensemble model testing

Step 1 - Data import and preparation

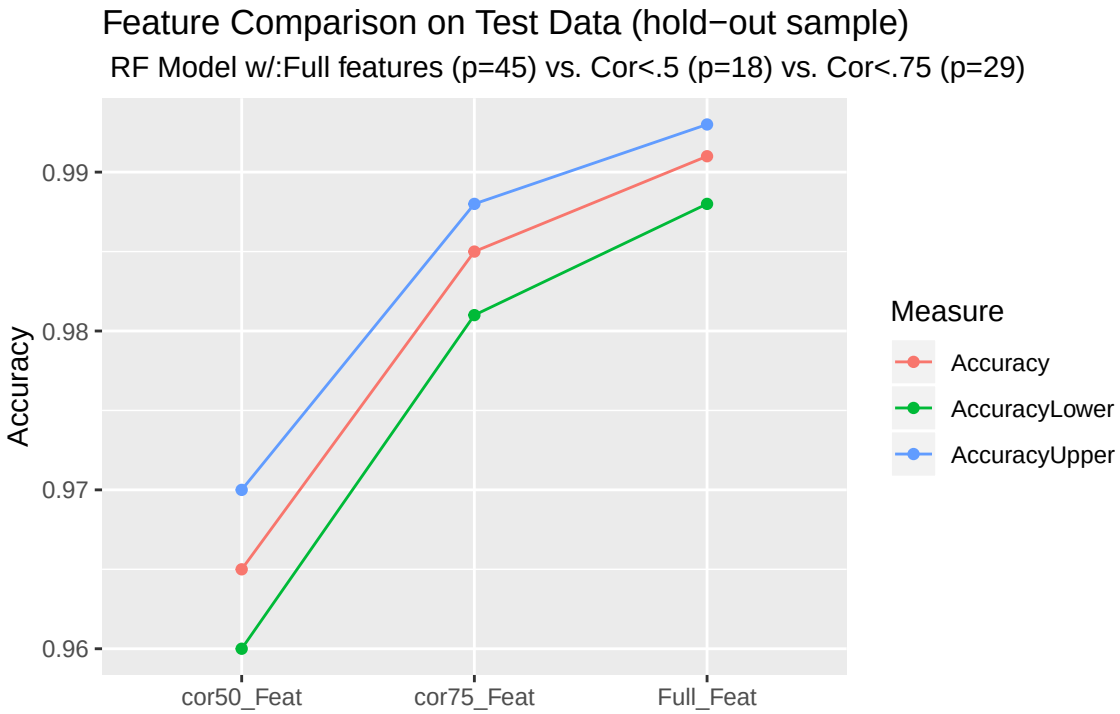
Data for this assignment was downloaded from: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

Data preprocessing included: - Removing non-feature variables and empty columns in the supplied "pmltesting" data, then matching features to the supplied "pmltraining" data to begin preprocessing. This reduced the data from 159 to 52 features. - Created training and testing sets with "pmltraining" with n sizes: Training (n=14,7180), Testing (n=4,904) - Ran preprocessing procedures of centering, scaling and elimination of near-zero features. This reduced the data set from 52 to 45 features.

Step 2 - Feature Reduction

From Step 1, the 45 features in the training data were further examined for potential reduction using three techniques: Boruta, Recursive Feature Elimination (RFE) and reduction based on high feature correlation. From these analysis, both Boruta and RFE results suggested that all features are important and should be included for training. Examinging feature correlations resulted in 18 features after eliminating features with correlations above .5, and 29 features after eliminating correlations above .75. Each of these three feature sets (all features: "Full_Feat", correlations above .5: "cor50_feat", correlations above .75: "cor75_feat") were modeled on the training data using a Random Forest algorithm. The results suggest that including all (n=45) features produces the highest accuracy (see Figure 1).

Figure 1. *Feature Reduction*



Step 3 - Individual Modeling/Algorithm Testing

Given a mulit-class prediction assignment, a variety of "classification" algorithms were considered, including: Random Forest, Adaboost, Gradient Boosting, Extreme Gradient Boosting, Support Vector Machines, C5.0 and Naive Bayesian models. Figure 2.b shows the "accuracy" comparisons on the created hold-out (testing) sample that was split from the supplied "pmltraining" data. The results suggest that the Extreme Gradient Boosting model ("xgbTree") has the greatest accuracy (.9937). This model also had the highest cross-validation accuracy of .9947 using repeated K-fold (n=10) with 3 repeats (Figure 2.a).

Figure 2.a *Cross-Validation Accuracy - Individual Models*

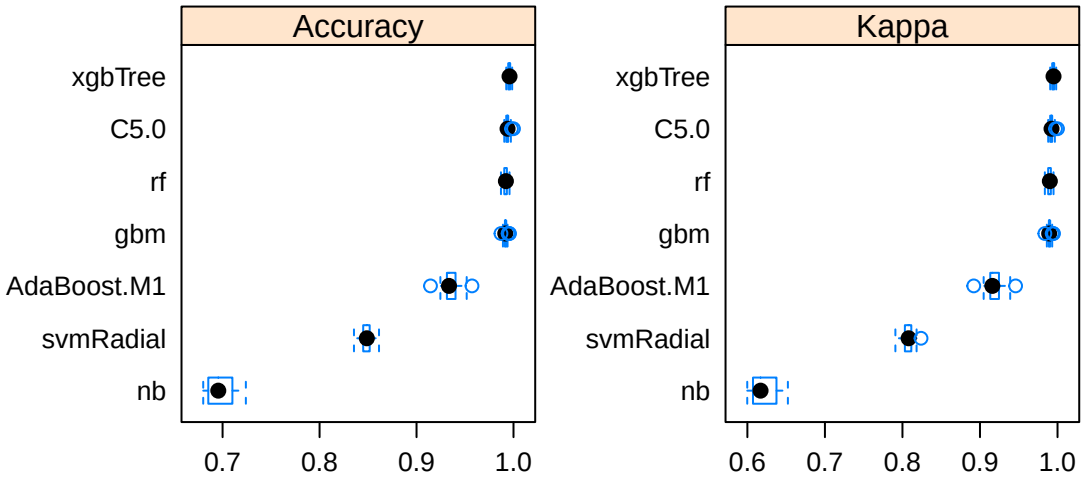
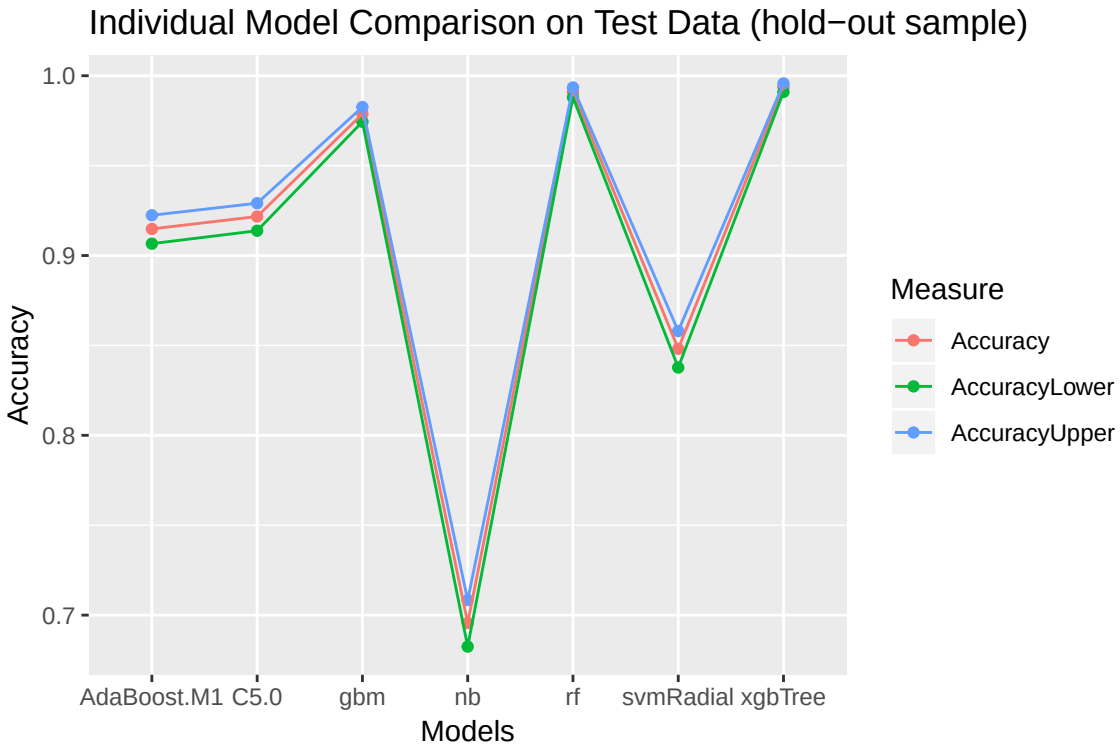


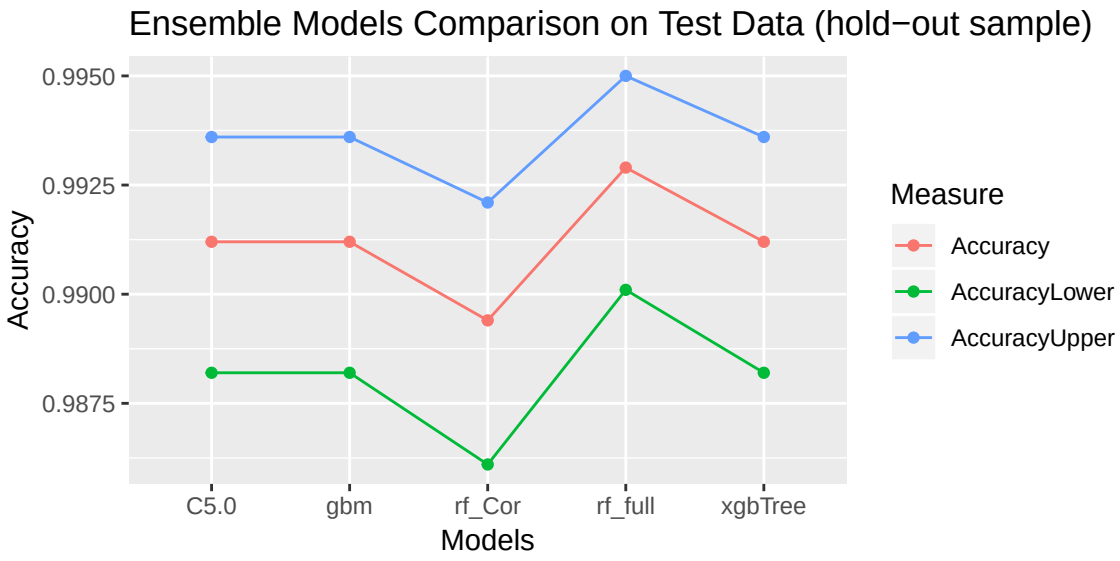
Figure 2.b *Testing Accuracy - Individual Models*



Step 4 - Ensemble Model Testing

Ensemble models were also created using the predictions of the differnet individual models. Ensemble algorithms included: Random Forest (both - full models and reduced models data), Gradient Boosting, Extreme Gradient Boosting and C5.0. All models were included as the training data for the ensemble modeling except for "rf_Cor", which used a reduced set of model predictions that excluded: Random Forest, C5.0 and Naive Bayes predictions (which showed relatively strong correlations with each other). Ensemble model results suggest that the Random Forest algorithm using predictions from all models ("rf_full") has the highest mean accuracy on the hold-out (testing) sample of .9929. This strony result, however, is actually below the individual Extreme Gradient Boost model with mean accuracy of .9937.

Figure 3. *Testing Accuracy - Ensemble Models*



Final Results

Comparing all model accuracies–both individual and ensemble models–shows that the individual Extreme Gradient Boost model is the best performing model, with accuracy of .9937 on the hold-out (testing) data.

Table 1. *Testing Accuracy - All Models*

	Accuracy	Kappa	AccuracyLower	AccuracyUpper
rf	0.9910	0.9887	0.9880	0.9935
ada	0.9148	0.8924	0.9066	0.9224
gbm	0.9788	0.9732	0.9744	0.9826
xgb	0.9937	0.9920	0.9910	0.9957
svm	0.8481	0.8071	0.8377	0.8580
c50	0.9217	0.9010	0.9138	0.9291
nb	0.6956	0.6093	0.6825	0.7084
ensemble_rf_full	0.9929	0.9910	0.9901	0.9950
ensemble_rf_Cor	0.9894	0.9866	0.9861	0.9921
ensemble_xgb	0.9912	0.9889	0.9882	0.9936
ensemble_gbm	0.9912	0.9889	0.9882	0.9936
ensemble_c50	0.9912	0.9889	0.9882	0.9936