Coursera Machine Learning – Final Project

Author:Srini Sundhar

Introduction

The purpose of this document is to describe the Program outline, Results, and Analysis including building of the predictor model for the "Coursera Machine Learning" Final Project.

The course project is described in the assignment as follows:

"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://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"

Model Building

The training set, as read from the 'training.csv' had about 19622 observed rows and 160 variables (columns). Since many of the columns had only 'NA' as values, those columns were removed from the data set. The processed data was then split into train and test data sets (70% train and 30% test data).

The training set was used to train five different predictive models using Bayesian Network, Recursive Partitioning (CART), Random Forest, Boosting, Bagged Tree methods. The predictive models were then tested using the test data set. A set of metrics (accuracy, kappa, time) were collected on both the training and test data set, and tabulated. The top 10 influential variables for the predictors (except Bayesian Network) were obtained and plotted.

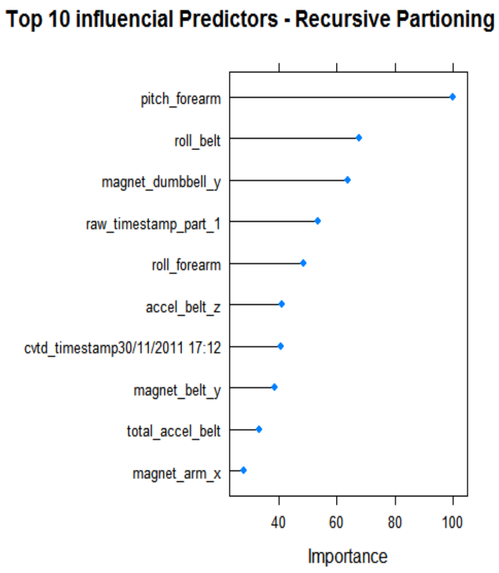
Finally, the 20 test cases were run through the five predictor models and the results were compared.

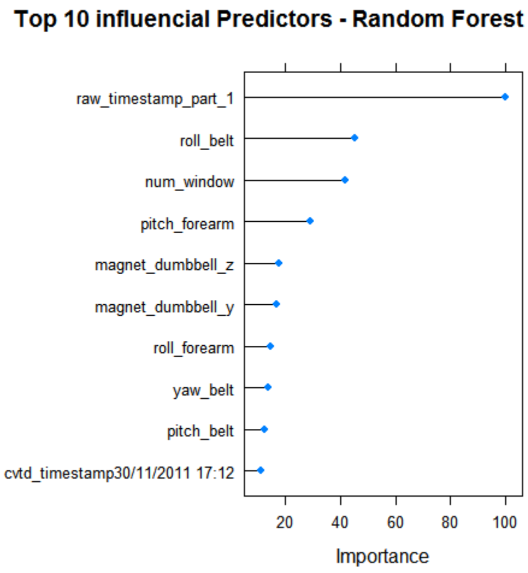
Program procedural outline

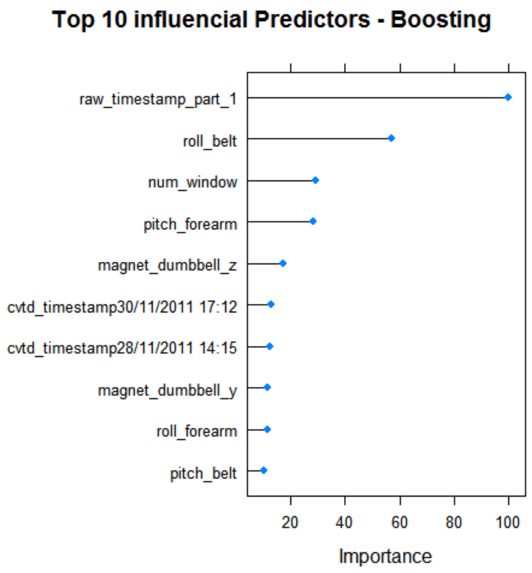
1. Read the training data and process it to remove the NA value columns
2. Split the training data set into train and test data
3. Train and predict using Bayesian Network
4. Train and predict using Recursive Partitioning
5. Train and predict using Random Forest
6. Train and predict using Boosting
7. Train and predict using Bagging Tree
8. Obtain the Confusion Matrix for the five predictive model outputs
9. Identify and plot the top influential predictors for the five predictive models
10. Plot the Decision Tree
11. Predict outcomes for the 20 Test Cases

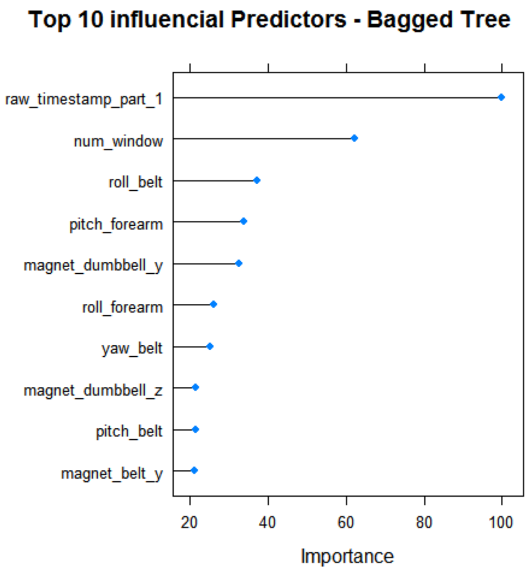
Predictor Results

Predictor Method	**Accuracy (from the Model on Training Dataset)**	**Kappa(from the Model on Training Dataset)**	**Accuracy (from Confusion Matrix on the Testing Dataset)**	**Kappa (from Confusion Matrix on the Testing Dataset)**	**Execution Time Taken for Training the Model**	**Predictor Output for 20 Test Cases (Validation)**
Bayesian Network	0.444454	0.2881898	0.427	0.2696	6.630559 mins	B B B A A B B B A A B B B A B B A B B B
Recursive Partitioning - CART	0.5479283	0.4168911	0.5056	0.3807	19.31393 secs	B A C A C C C C A A C C C A C C C B B C
Random Forest	0.9990972	0.9988581	0.9998	0.9997	3.472677 hours	B A B A A E D B A A B C B A E E A B B B
Boosting	0.9966660	0.9957833	0.9983	0.9979	2.451348 hours	B A B A A E D B A A B C B A E E A B B B
Bagged CART	0.9965955	0.9956941	0.9996	0.9995	8.426361 mins	B A B A A E D B A A B C B A E E A B B B


Top 10 influencial Predictors - Recursive Partioning


Top 10 influencial Predictors - Random Forest


Top 10 influencial Predictors - Boosting


Top 10 influencial Predictors - Bagged Tree

Analysis & Conclusion

An analysis of the results from the five predictors shows that:

1. Bayesian Network had the least accuracy
2. Random Partitioning was better than Bayesian Network but much worse than other three
3. Random Forest provided the best accuracy
4. Random Forest took the maximum execution time
5. Boosting’s accuracy was close to Random Forest but with better computational performance

6. Bagged Tree CART provided results that were as good as Random Forest and Boosting but with far better computational performance as those two methods.
7. Bagged Tree CART is recommended as the best performer among the five methods.
8. In terms of the top influential predictors, raw_timestamp_part1, num_window, row_belt, pitch_forearm, magnet_dumbbell_y, magnet_dumbbell_z were seen consistently in the top ten influential predictors.