**Practical Machine Learning Week 4 Project Write-up**

The main goal of this project is to build a machine learning model that can predict exercise quality from data collected using personal devices such as Jawbone Up, Nike FuelBand, and Fitbit.

**Data Cleaning and Processing**

The first step is to read the given training and testing datasets and get a rough idea of the data. In the training dataset, data from accelerometers on the belt, forearm, arm, and dumbbell of 6 participants are given. The structure of the training data frame shows that it consists of 19622 observations of 160 variables. The “classe” variable in the training set represents the qualify of the exercise. Examining the structure of testing data frame, I found that it consists of 20 observations of 160 variables. But the “classe” variable in the testing set is NULL. This confirms that the “classe” variable is the prediction target.

After determining the target variable, a more detailed exploratory analysis and key feature selection on the other 159 variables were conducted. I noticed that many columns contain NA values or blank values for almost all observations. I decided to delete the columns where more than 90% of the observations are NA or blank because these data won’t produce effective information for the classification. A total of 100 variables were removed due to NA or blank. In addition, the first 7 columns contain irrelevant information such as names of participants, activity start time/end time etc. Hence, I also removed these variables. In summary, the final processed training/testing datasets only have 53 variables from accelerometers on the belt, forearm, arm, and dumbbell.

**Machine Learning**

Next step is to apply proper machine learning algorithms for this classification problem. Before applying machine learning algorithms, training dataset was divided into two categories - 80% training and 20% testing.

Two machine learning models (classification tress and random forest) were implemented and compared. Cross-validation technique was utilized to prevent overfitting. I used k-folds method, to be specific, 5 folds method for cross-validation. In the “train” function, trControl argument was specified as trControl = trainControl(method = “cv”, number = 5). The algorithm that gives the lower out of sample error will be chosen.

**Method 1 - Classification Tree**

*model\_ct <- train(classe ~., method ="rpart", data = train\_data,*

*trControl = trainControl(method="cv",number=n\_folds))*

*pd\_ct <- predict(model\_ct,test\_data)*

The final value used for the model was complexity parameter = 0.03676012, which gives the highest accuracy value of 0.5248 in the training data. Detailed classification tree can be found in Figure 1. When comparing the results of the predicted “classe” with “classe” information in testing data, the out of sample accuracy rate is approximately 0.4986. The accuracy rate in the training set is slightly higher than the out of sample accuracy rate, which makes sense. However, I would say this is not a desirable accuracy rate overall so this classification tree method is not ideal.

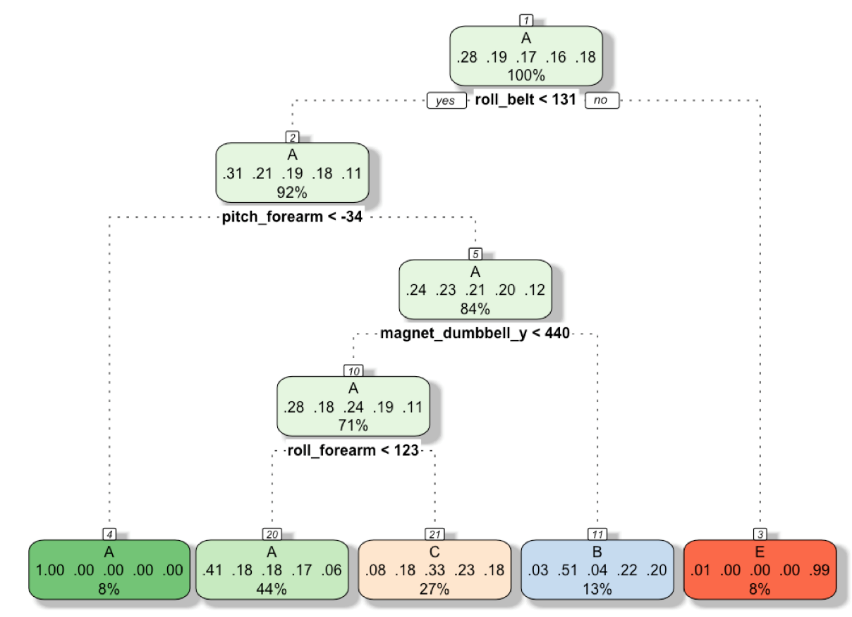


Figure 1. Classification Tree Model Prediction

**Method 2 - Random Forest**

*model\_rf <- train(classe ~., method ="rf", data = train\_data,verbose=FALSE,*

*trControl = trainControl(method="cv",number=n\_folds))*

*pd\_rf <- predict(model\_rf,test\_data)*

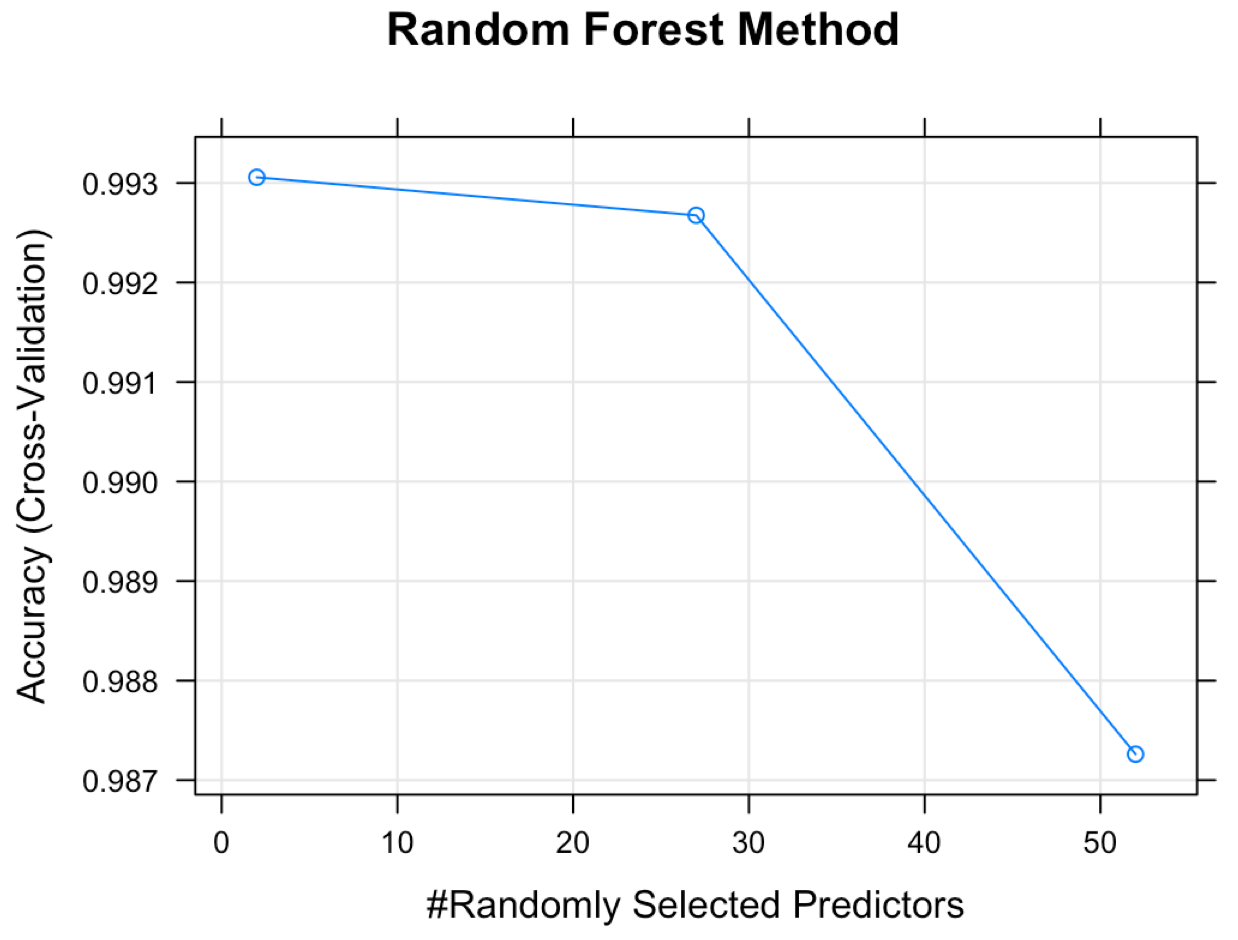
Model and prediction results from random forest method was also visualized and summarized. It is found that the accuracy decreases with increasing randomly selected predictors. 2 predictors gave the best accuracy rate of 0.993. 27 predictors will give slightly lower the accuracy rate to 0.9927. While 52 predictors produce the lowest accuracy rate of 0.9873. Therefore, the optimal number of predictors were chosen as 2. This result indicated that the 52 variables may not be totally independent. When comparing the results of the predicted “classe” with “classe” information in testing data, the out of sample accuracy rate is approximately 0.9946. This is pretty good.

Figure 2. Random Forest Method Accuracy

Based on the results, random forest model was finally selected as the classification system for predicting the activity qualify because the expected out of sample error is as low as 0.54%.

**Prediction Results**

At last, the proposed random forest classification model/system was utilized to predict 20 different test cases. The predicted quality of exercise is as follows:

B A B A A E D B A A B C B A E E A B B B