**Practical Machine Learning - Course Project**

**Mehrshad Esfahani**

Introduction

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, our 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. The goal of this project is to predict the manner in which they did the exercise.

1. **Loading the data**

## [1] 19622 160

## [1] 20 160

We have 19622 values of 160 variables.

1. **Cleaning the data**

Columns with missing value should be removed.

## [1] 19622 60

## [1] 20 60

Data is now reduced to 60 variables.

1. **Preprocessing**
2. Removing the non zero variables

Values near zero do not have so big influence in the predictions

1. Validation set

75% observation training dataset is set to the train model. 70% will be tried later.

## [1] 14718 28

## [1] 4904 28

1. **Train Model**

Random forest with a cross validation of 5 folds is used to avoid overfitting.

## Random Forest

##

## 14718 samples

## 27 predictor

## 5 classes: 'A', 'B', 'C', 'D', 'E'

##

## No pre-processing

## Resampling: Cross-Validated (10 fold)

## Summary of sample sizes: 13246, 13245, 13248, 13245, 13247, 13246, ...

## Resampling results across tuning parameters:

##

## mtry Accuracy Kappa Accuracy SD Kappa SD

## 2 0.9927294 0.9908028 0.001700103 0.002150719

## 14 0.9927966 0.9908879 0.002836707 0.003588893

## 27 0.9889914 0.9860747 0.004497867 0.005690737

##

## Accuracy was used to select the optimal model using the largest value.

## The final value used for the model was mtry = 14.

1. **Cross Validation Testing and Out-of-Sample Error Estimate**

To check its accuracy the training model is applied on the testing model.

Accuracy and Estimated out of sample error

## Reference

## Prediction A B C D E

## A 1392 2 1 0 0

## B 4 944 1 0 0

## C 0 2 848 5 0

## D 0 0 1 803 0

## E 0 0 1 3 897

As it is shown above only few variables out of this model.

## [1] 0.004078303

The estimated accuracy of the model is 99.7% and the estimated out-of-sample error based on our fitted model applied to the cross validation dataset is 0.3%.

## Prediction with Decision Trees

## Confusion Matrix and Statistics

##

## Reference

## Prediction A B C D E

## A 2150 60 7 1 0

## B 61 1260 69 64 0

## C 21 188 1269 143 4

## D 0 10 14 857 78

## E 0 0 9 221 1360

##

## Overall Statistics

##

## Accuracy : 0.8789

## 95% CI : (0.8715, 0.8861)

## No Information Rate : 0.2845

## P-Value [Acc > NIR] : < 2.2e-16

##

## Kappa : 0.8468

## Mcnemar's Test P-Value : NA

##

## Statistics by Class:

##

## Class: A Class: B Class: C Class: D Class: E

## Sensitivity 0.9633 0.8300 0.9276 0.6664 0.9431

## Specificity 0.9879 0.9693 0.9450 0.9845 0.9641

## Pos Pred Value 0.9693 0.8666 0.7809 0.8936 0.8553

## Neg Pred Value 0.9854 0.9596 0.9841 0.9377 0.9869

## Prevalence 0.2845 0.1935 0.1744 0.1639 0.1838

## Detection Rate 0.2740 0.1606 0.1617 0.1092 0.1733

## Detection Prevalence 0.2827 0.1853 0.2071 0.1222 0.2027

## Balanced Accuracy 0.9756 0.8997 0.9363 0.8254 0.9536

## Prediction with Random Forests

## Confusion Matrix and Statistics

##

## Reference

## Prediction A B C D E

## A 2231 2 0 0 0

## B 1 1516 0 0 0

## C 0 0 1367 3 0

## D 0 0 1 1282 1

## E 0 0 0 1 1441

##

## Overall Statistics

##

## Accuracy : 0.9989

## 95% CI : (0.9978, 0.9995)

## No Information Rate : 0.2845

## P-Value [Acc > NIR] : < 2.2e-16

##

## Kappa : 0.9985

## Mcnemar's Test P-Value : NA

##

## Statistics by Class:

##

## Class: A Class: B Class: C Class: D Class: E

## Sensitivity 0.9996 0.9987 0.9993 0.9969 0.9993

## Specificity 0.9996 0.9998 0.9995 0.9997 0.9998

## Pos Pred Value 0.9991 0.9993 0.9978 0.9984 0.9993

## Neg Pred Value 0.9998 0.9997 0.9998 0.9994 0.9998

## Prevalence 0.2845 0.1935 0.1744 0.1639 0.1838

## Detection Rate 0.2843 0.1932 0.1742 0.1634 0.1837

## Detection Prevalence 0.2846 0.1933 0.1746 0.1637 0.1838

## Balanced Accuracy 0.9996 0.9993 0.9994 0.9983 0.9996

## Prediction with Generalized Boosted Regression

## Confusion Matrix and Statistics

##

## Reference

## Prediction A B C D E

## A 2231 4 0 0 0

## B 1 1512 1 0 0

## C 0 2 1358 6 0

## D 0 0 9 1271 0

## E 0 0 0 9 1442

##

## Overall Statistics

##

## Accuracy : 0.9959

## 95% CI : (0.9942, 0.9972)

## No Information Rate : 0.2845

## P-Value [Acc > NIR] : < 2.2e-16

##

## Kappa : 0.9948

## Mcnemar's Test P-Value : NA

##

## Statistics by Class:

##

## Class: A Class: B Class: C Class: D Class: E

## Sensitivity 0.9996 0.9960 0.9927 0.9883 1.0000

## Specificity 0.9993 0.9997 0.9988 0.9986 0.9986

## Pos Pred Value 0.9982 0.9987 0.9941 0.9930 0.9938

## Neg Pred Value 0.9998 0.9991 0.9985 0.9977 1.0000

## Prevalence 0.2845 0.1935 0.1744 0.1639 0.1838

## Detection Rate 0.2843 0.1927 0.1731 0.1620 0.1838

## Detection Prevalence 0.2849 0.1930 0.1741 0.1631 0.1849

## Balanced Accuracy 0.9994 0.9979 0.9957 0.9935 0.9993

1. **Figures**



