**Assignment No. 19.2**

1. Use the below given data set

DataSet

2. Perform the below given activities:

a. Create classification model using different classifiers

b. Verify model goodness of fit

c. Apply all the model validation techniques.

I have chosen two approaches for classifying, basing on the previous assignments experience. One is the multinomial logistics regression, and the other C5.0 based decision tree classification, which has created overall model accuracy and the MCC value very high. So in this assignment since the data is the same, the previous algorithm and the output will be used.

> model <- train(classe~.,data=bitrain[,-52],trContol = train\_control,method ="C5.0")

> pred3 <- predict(model,bitest[,-52])

> pred3 <- cbind(bitest,pred3)

> CMpred3 <- confusionMatrix(as.factor(pred3$classe), as.factor(pred3$pred3))

> CMpred3$byClass[,7]

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

0.9971347 0.9959184 1.0000000 0.9917355 0.9984399

> mcc(preds = pred3$pred3, actuals = as.factor(pred3$classe))

[1] 0.9958321

> CMpred3

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 348 0 0 0 0

B 2 244 0 0 0

C 0 0 31 0 0

D 0 0 0 60 1

E 0 0 0 0 320

Overall Statistics

Accuracy : 0.997

95% CI : (0.9913, 0.9994)

No Information Rate : 0.3479

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

Kappa : 0.9958

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.9943 1.0000 1.00000 1.00000 0.9969

Specificity 1.0000 0.9974 1.00000 0.99894 1.0000

Pos Pred Value 1.0000 0.9919 1.00000 0.98361 1.0000

Neg Pred Value 0.9970 1.0000 1.00000 1.00000 0.9985

Prevalence 0.3479 0.2425 0.03082 0.05964 0.3191

Detection Rate 0.3459 0.2425 0.03082 0.05964 0.3181

Detection Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Balanced Accuracy 0.9971 0.9987 1.00000 0.99947 0.9984

The above results and the validation metrics show clearly that the decision tree approach using the C5.0 algorithm is showing a very high overall accuracy. In terms of the Sensitivity, Specificity, Mathews Correlation Coefficient, and the F1 Score.

> fit1 <- multinom(classe~.,data = compbicepsdata[,-52])

# weights: 260 (204 variable)

initial value 6476.378160

iter 10 value 2705.232017

iter 20 value 1505.079428

iter 30 value 1200.663807

iter 40 value 1077.016253

iter 50 value 1041.946052

iter 60 value 1012.673881

iter 70 value 994.620772

iter 80 value 981.990370

iter 90 value 975.324537

iter 100 value 967.878856

final value 967.878856

stopped after 100 iterations

Residual Deviance: 1935.758

AIC: 2343.758

> pred1 <- predict(fit1,compbicepsdata[,-52])

> pred1 <- cbind(compbicepsdata,pred1)

> confusionMatrix(as.factor(pred1$classe),as.factor(pred1$pred1))

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 1296 62 0 2 5

B 17 864 20 0 0

C 10 14 88 0 0

D 0 0 0 275 1

E 24 1 0 29 1316

Overall Statistics

Accuracy : 0.954

95% CI : (0.9471, 0.9603)

No Information Rate : 0.3347

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

Kappa : 0.9358

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.9621 0.9182 0.81481 0.89869 0.9955

Specificity 0.9742 0.9880 0.99387 0.99973 0.9800

Pos Pred Value 0.9495 0.9589 0.78571 0.99638 0.9606

Neg Pred Value 0.9808 0.9753 0.99489 0.99173 0.9977

Prevalence 0.3347 0.2338 0.02684 0.07604 0.3285

Detection Rate 0.3221 0.2147 0.02187 0.06834 0.3270

Detection Prevalence 0.3392 0.2239 0.02783 0.06859 0.3405

Balanced Accuracy 0.9682 0.9531 0.90434 0.94921 0.9877

#F1 Score

> CMpred1$byClass[,7]

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

0.9557522 0.9381107 0.8000000 0.9450172 0.9777117

> mcc(preds = pred1$pred1, actuals = as.factor(pred1$classe))

[1] 0.9360295

> pR2(fit1)["McFadden"]

fitting null model for pseudo-r2

# weights: 10 (4 variable)

initial value 6476.378160

iter 10 value 5440.939667

iter 10 value 5440.939664

final value 5440.939664

converged

McFadden

0.8221118

The multinomial logistic regression using the entire data shows relatively higher level of overall accuracy, F1 Score, and McFadden’s Pseudo Rsquare, compared to the split sample of train and test out of the same dataset. The following are the results of the train and test data sets.

> fit2 <- multinom(classe~.,data = bitrain[,-52])

# weights: 260 (204 variable)

initial value 4857.283620

iter 10 value 2271.655772

iter 20 value 1341.060718

iter 30 value 1156.584919

iter 40 value 1093.482644

iter 50 value 1060.158061

iter 60 value 1039.054166

iter 70 value 1018.987163

iter 80 value 1010.858725

iter 90 value 1001.984176

iter 100 value 995.588594

final value 995.588594

stopped after 100 iterations

Residual Deviance: 1991.177

AIC: 2399.177

> pred2 <- predict(fit2,bitest[,-52])

> pred2 <- cbind(bitest,pred2)

> pred2$classe <- as.factor(pred2$classe)

> pred2$pred2 <- as.factor(pred2$pred2)

> confusionMatrix(pred2$classe,pred2$pred2)

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 346 1 1 0 0

B 31 213 2 0 0

C 5 2 24 0 0

D 2 0 0 58 1

E 5 0 3 14 298

Overall Statistics

Accuracy : 0.9334

95% CI : (0.9162, 0.948)

No Information Rate : 0.3867

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

Kappa : 0.9067

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.8895 0.9861 0.80000 0.80556 0.9967

Specificity 0.9968 0.9582 0.99283 0.99679 0.9689

Pos Pred Value 0.9943 0.8659 0.77419 0.95082 0.9313

Neg Pred Value 0.9347 0.9961 0.99385 0.98519 0.9985

Prevalence 0.3867 0.2147 0.02982 0.07157 0.2972

Detection Rate 0.3439 0.2117 0.02386 0.05765 0.2962

Detection Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Balanced Accuracy 0.9431 0.9722 0.89641 0.90117 0.9828

#F1 Score

> CMpred2$byClass[,7]

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

0.9389417 0.9220779 0.7868852 0.8721805 0.9628433

> mcc(preds = pred2$pred2, actuals = as.factor(pred2$classe))

[1] 0.9086954

> pR2(fit2)["McFadden"]

fitting null model for pseudo-r2

# weights: 10 (4 variable)

initial value 4857.283620

iter 10 value 4076.494148

final value 4076.494064

converged

McFadden

0.7557733