**Assignment No. 18.1**

Required:

1. Use the below given data set

DataSet

2. Perform the below given activities:

a. Create classification model using different decision trees.

b. Verify model goodness of fit.

c. Apply all the model validation techniques.

d. Make conclusions

Since the Data is a Multi Class Categorical Classification model, Numerical based RMSE, MAP, MAPE kind of validation rules could not be performed. Further the ROC Curve, Lift Chart, K-S Statistics, Residual Deviance, Hosmer-Lemeshow test are more applicable for Logistic Regressions and numerical probability prediction modelling. Hence based on what is taught to us, Sensitivity (Recall), Specificity, Positive Predictive Value, Mathews Correlation Coefficient, F1 Score, are calculated as validations for the 9 models which are generated for the given assignment. The following is the results and scores of the respective validations.

The summary of the all the models is presented in the following table

|  |  |  |  |
| --- | --- | --- | --- |
|  | Classification Method | Overall Accuracy | MCC |
| Model 1 | Class | 0.9483 | 0.9274 |
| Model 2 | rpart | 0.9046 | 0.8678 |
| Model 3 | CV, N=10 | 0.7386 | 0.6280 |
| Model 4 | C5.0 | 0.997 | 0.9958 |
| Model 5 | bstTree | 0.5775 | 0.5287 |
| Model 6 | C5.0 Rules | 0.9901 | 0.9860 |
| Model 7 | C5.0 Tree | 0.9857 | 0.9791 |
| Model 8 | cTree | 0.9722 | 0.9611 |
| Model 9 | cTree2 | 0.838 | 0.7717 |

> pred1 <- predict(fit1,bitest[,-52],type="class")

> confusionMatrix(pred1,bitest$classe)

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 334 16 2 0 0

B 14 226 13 0 0

C 0 4 16 0 0

D 0 0 0 58 0

E 0 0 0 3 320

Overall Statistics

Accuracy : 0.9483

95% CI : (0.9328, 0.9612)

No Information Rate : 0.3459

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

Kappa : 0.9273

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.9598 0.9187 0.51613 0.95082 1.0000

Specificity 0.9726 0.9645 0.99590 1.00000 0.9956

Pos Pred Value 0.9489 0.8933 0.80000 1.00000 0.9907

Neg Pred Value 0.9786 0.9734 0.98479 0.99684 1.0000

Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Detection Rate 0.3320 0.2247 0.01590 0.05765 0.3181

Detection Prevalence 0.3499 0.2515 0.01988 0.05765 0.3211

Balanced Accuracy 0.9662 0.9416 0.75601 0.97541 0.9978

> CMpred1 <- confusionMatrix(pred1, bitest$classe)

> CMpred1$byClass[,7]

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

0.9542857 0.9058116 0.6274510 0.9747899 0.9953344

> mcc(preds = pred1, actuals = bitest$classe)

[1] 0.927466

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

> confusionMatrix(pred2,bitest$classe)

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 322 11 0 0 4

B 26 235 31 0 0

C 0 0 0 0 0

D 0 0 0 52 15

E 0 0 0 9 301

Overall Statistics

Accuracy : 0.9046

95% CI : (0.8847, 0.922)

No Information Rate : 0.3459

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

Kappa : 0.8658

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.9253 0.9553 0.00000 0.85246 0.9406

Specificity 0.9772 0.9250 1.00000 0.98413 0.9869

Pos Pred Value 0.9555 0.8048 NaN 0.77612 0.9710

Neg Pred Value 0.9611 0.9846 0.96918 0.99042 0.9727

Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Detection Rate 0.3201 0.2336 0.00000 0.05169 0.2992

Detection Prevalence 0.3350 0.2903 0.00000 0.06660 0.3082

Balanced Accuracy 0.9512 0.9401 0.50000 0.91829 0.9638

> CMpred2 <- confusionMatrix(pred2, bitest$classe)

> CMpred2$byClass[,7]

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

0.9401460 0.8736059 NA 0.8125000 0.9555556

> mcc(preds = pred2, actuals = bitest$classe)

[1] 0.8678342

> train\_control <-trainControl(method="cv", number=10)

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

> confusionMatrix(pred3$classe, pred3$pred3)

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 314 34 0 0 0

B 73 173 0 0 0

C 2 29 0 0 0

D 10 9 0 0 42

E 19 45 0 0 256

Overall Statistics

Accuracy : 0.7386

95% CI : (0.7102, 0.7655)

No Information Rate : 0.4155

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

Kappa : 0.622

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.7512 0.5966 NA NA 0.8591

Specificity 0.9422 0.8980 0.96918 0.93936 0.9096

Pos Pred Value 0.9023 0.7033 NA NA 0.8000

Neg Pred Value 0.8419 0.8461 NA NA 0.9388

Prevalence 0.4155 0.2883 0.00000 0.00000 0.2962

Detection Rate 0.3121 0.1720 0.00000 0.00000 0.2545

Detection Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Balanced Accuracy 0.8467 0.7473 NA NA 0.8843

> CMpred3 <- confusionMatrix(pred3$classe, pred3$pred3)

> CMpred3$byClass[,7]

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

0.8198433 0.6455224 NA NA 0.8284790

> mcc(preds = pred3$pred3, actuals = pred3$classe)

[1] 0.6279077

> model1 <- train(classe~.,data=bitrain[,-52],trControl = train\_control, method = "C5.0")

> confusionMatrix(pred4$classe, pred4$pred4)

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

> CMpred4 <- confusionMatrix(pred4$classe, pred4$pred4)

> CMpred4$byClass[,7]

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

0.9971347 0.9959184 1.0000000 0.9917355 0.9984399

> mcc(preds = pred4$pred4, actuals = pred4$classe)

[1] 0.9958321

> model2 <- train(classe~.,data=bitrain[,-52],trControl = train\_control, method = "bstTree")

> confusionMatrix(pred5$classe, pred5$pred5)

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 337 11 0 0 0

B 2 244 0 0 0

C 0 31 0 0 0

D 0 61 0 0 0

E 0 320 0 0 0

Overall Statistics

Accuracy : 0.5775

95% CI : (0.5463, 0.6083)

No Information Rate : 0.663

P-Value [Acc > NIR] : 1

Kappa : 0.4143

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.9941 0.3658 NA NA NA

Specificity 0.9835 0.9941 0.96918 0.93936 0.6819

Pos Pred Value 0.9684 0.9919 NA NA NA

Neg Pred Value 0.9970 0.4434 NA NA NA

Prevalence 0.3370 0.6630 0.00000 0.00000 0.0000

Detection Rate 0.3350 0.2425 0.00000 0.00000 0.0000

Detection Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Balanced Accuracy 0.9888 0.6800 NA NA NA

> CMpred5 <- confusionMatrix(pred5$classe, pred5$pred5)

> CMpred5$byClass[,7]

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

0.9810771 0.5345016 NA NA NA

> mcc(preds = pred5$pred5, actuals = pred5$classe)

[1] 0.5287867

> model3 <- train(classe~.,data=bitrain[,-52],trControl = train\_control, method = "C5.0Rules")

> confusionMatrix(pred6$classe, pred6$pred6)

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 347 1 0 0 0

B 3 242 1 0 0

C 3 1 27 0 0

D 0 0 0 60 1

E 0 0 0 0 320

Overall Statistics

Accuracy : 0.9901

95% CI : (0.9818, 0.9952)

No Information Rate : 0.3509

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

Kappa : 0.9861

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.9830 0.9918 0.96429 1.00000 0.9969

Specificity 0.9985 0.9948 0.99591 0.99894 1.0000

Pos Pred Value 0.9971 0.9837 0.87097 0.98361 1.0000

Neg Pred Value 0.9909 0.9974 0.99897 1.00000 0.9985

Prevalence 0.3509 0.2425 0.02783 0.05964 0.3191

Detection Rate 0.3449 0.2406 0.02684 0.05964 0.3181

Detection Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Balanced Accuracy 0.9907 0.9933 0.98010 0.99947 0.9984

> CMpred6 <- confusionMatrix(pred6$classe, pred6$pred6)

> CMpred6$byClass[,7]

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

0.9900143 0.9877551 0.9152542 0.9917355 0.9984399

> mcc(preds = pred6$pred6, actuals = pred6$classe)

[1] 0.9860942

> model4 <- train(classe~.,data=bitrain[,-52],trControl = train\_control, method = "C5.0Tree"); model4

> confusionMatrix(pred7$classe, pred7$pred7)

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 347 1 0 0 0

B 3 238 5 0 0

C 3 1 27 0 0

D 0 0 0 60 1

E 0 0 0 1 319

Overall Statistics

Accuracy : 0.9851

95% CI : (0.9755, 0.9916)

No Information Rate : 0.3509

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

Kappa : 0.9791

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.9830 0.9917 0.84375 0.98361 0.9969

Specificity 0.9985 0.9896 0.99589 0.99894 0.9985

Pos Pred Value 0.9971 0.9675 0.87097 0.98361 0.9969

Neg Pred Value 0.9909 0.9974 0.99487 0.99894 0.9985

Prevalence 0.3509 0.2386 0.03181 0.06064 0.3181

Detection Rate 0.3449 0.2366 0.02684 0.05964 0.3171

Detection Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Balanced Accuracy 0.9907 0.9906 0.91982 0.99127 0.9977

> CMpred7 <- confusionMatrix(pred7$classe, pred7$pred7)

> CMpred7$byClass[,7]

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

0.9900143 0.9794239 0.8571429 0.9836066 0.9968750

> mcc(preds = pred7$pred7, actuals = pred7$classe)

[1] 0.9791718

> model5 <- train(classe~.,data=bitrain[,-52],trControl = train\_control, method = "ctree")

> confusionMatrix(pred8$classe, pred8$pred8)

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 340 8 0 0 0

B 8 231 7 0 0

C 0 3 28 0 0

D 0 0 0 60 1

E 0 0 0 1 319

Overall Statistics

Accuracy : 0.9722

95% CI : (0.96, 0.9814)

No Information Rate : 0.3459

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

Kappa : 0.9611

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.9770 0.9545 0.80000 0.98361 0.9969

Specificity 0.9878 0.9804 0.99691 0.99894 0.9985

Pos Pred Value 0.9770 0.9390 0.90323 0.98361 0.9969

Neg Pred Value 0.9878 0.9855 0.99282 0.99894 0.9985

Prevalence 0.3459 0.2406 0.03479 0.06064 0.3181

Detection Rate 0.3380 0.2296 0.02783 0.05964 0.3171

Detection Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Balanced Accuracy 0.9824 0.9675 0.89846 0.99127 0.9977

> CMpred8 <- confusionMatrix(pred8$classe, pred8$pred8)

> CMpred8$byClass[,7]

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

0.9770115 0.9467213 0.8484848 0.9836066 0.9968750

> mcc(preds = pred8$pred8, actuals = pred8$classe)

[1] 0.9611262

>model6 <- train(classe~.,data=bitrain[,-52],trControl = train\_control, method = "ctree2"); >model6

> confusionMatrix(pred9$classe, pred9$pred9)

Confusion Matrix and Statistics

Reference

Prediction A B C D E

A 338 10 0 0 0

B 61 185 0 0 0

C 3 28 0 0 0

D 0 0 0 0 61

E 0 0 0 0 320

Overall Statistics

Accuracy : 0.838

95% CI : (0.8137, 0.8602)

No Information Rate : 0.3996

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

Kappa : 0.7642

Mcnemar's Test P-Value : NA

Statistics by Class:

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

Sensitivity 0.8408 0.8296 NA NA 0.8399

Specificity 0.9834 0.9221 0.96918 0.93936 1.0000

Pos Pred Value 0.9713 0.7520 NA NA 1.0000

Neg Pred Value 0.9027 0.9500 NA NA 0.9111

Prevalence 0.3996 0.2217 0.00000 0.00000 0.3787

Detection Rate 0.3360 0.1839 0.00000 0.00000 0.3181

Detection Prevalence 0.3459 0.2445 0.03082 0.06064 0.3181

Balanced Accuracy 0.9121 0.8758 NA NA 0.9199

> CMpred9 <- confusionMatrix(pred9$classe, pred9$pred9)

> CMpred9$byClass[,7]

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

0.9013333 0.7889126 NA NA 0.9129815

> mcc(preds = pred9$pred9, actuals = pred9$classe)

[1] 0.7716827