Seoul Air 2017 DM with Standardization

Minjung Kang

Wed Dec 12 23:23:25 2018

rm(list=ls())

> A <- read\_csv("KMJ037\_(SeoulAir)c.csv")  
> A$Y <-as.factor(A$Y)  
> A$Time <-as.factor(A$Time)  
> A$Date <-as.factor(A$Date)  
> A$District <-as.factor(A$District)  
> A$Y <- factor(A$Y, levels=c("AQG", "IT3", "IT2", "IT1", "Off.T"))  
>   
> A$PM2.5 <- NULL  
> A$Date <- NULL  
>   
> set.seed(123)

Preprocessing

> prePro <- preProcess(A[,3:7], method=c("center","scale"))  
> X <- predict(prePro, A[,3:7])  
>   
> myA <- data.frame(X, Y = A$Y)

> trData <- splitDataFUN(myA)[[1]]  
> teData <- splitDataFUN(myA)[[2]]  
>   
> IDvar <- setdiff(names(trData), c("Y"))  
> varEq <- paste0(IDvar, collapse = "+")  
> myModel <- as.formula(paste("Y", varEq, sep = " ~ "))

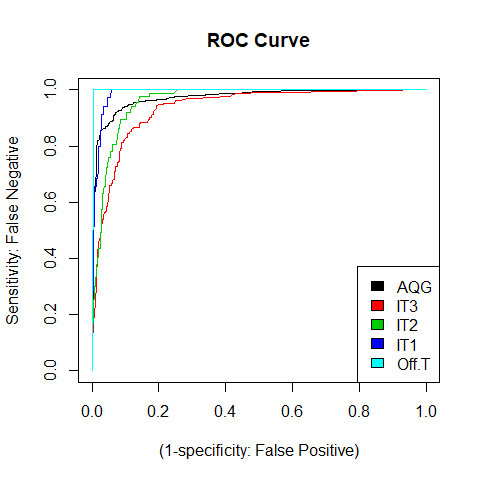
Random Forest

> model.rf <- caret::train(myModel, method = "rf", data = trData)  
> model.rf

Random Forest   
  
2870 samples  
 5 predictor  
 5 classes: 'AQG', 'IT3', 'IT2', 'IT1', 'Off.T'   
  
No pre-processing  
Resampling: Bootstrapped (25 reps)   
Summary of sample sizes: 2870, 2870, 2870, 2870, 2870, 2870, ...   
Resampling results across tuning parameters:  
  
 mtry Accuracy Kappa   
 2 0.8365887 0.6454015  
 3 0.8332502 0.6392319  
 5 0.8290961 0.6319285  
  
Accuracy was used to select the optimal model using the largest value.  
The final value used for the model was mtry = 2.

> test.rf <- test\_rocFUN(model.rf, teData)

Confusion Matrix and Statistics  
  
 Reference  
Prediction AQG IT3 IT2 IT1 Off.T  
 AQG 825 46 6 0 0  
 IT3 41 168 31 7 0  
 IT2 2 16 39 3 0  
 IT1 0 2 11 24 3  
 Off.T 0 0 0 1 0  
  
Overall Statistics  
   
 Accuracy : 0.862   
 95% CI : (0.8415, 0.8809)  
 No Information Rate : 0.7086   
 P-Value [Acc > NIR] : < 2.2e-16   
   
 Kappa : 0.6935   
 Mcnemar's Test P-Value : NA   
  
Statistics by Class:  
  
 Class: AQG Class: IT3 Class: IT2 Class: IT1  
Sensitivity 0.9505 0.7241 0.44828 0.68571  
Specificity 0.8543 0.9204 0.98155 0.98655  
Pos Pred Value 0.9407 0.6802 0.65000 0.60000  
Neg Pred Value 0.8764 0.9346 0.95880 0.99072  
Prevalence 0.7086 0.1894 0.07102 0.02857  
Detection Rate 0.6735 0.1371 0.03184 0.01959  
Detection Prevalence 0.7159 0.2016 0.04898 0.03265  
Balanced Accuracy 0.9024 0.8223 0.71491 0.83613  
 Class: Off.T  
Sensitivity 0.0000000  
Specificity 0.9991817  
Pos Pred Value 0.0000000  
Neg Pred Value 0.9975490  
Prevalence 0.0024490  
Detection Rate 0.0000000  
Detection Prevalence 0.0008163  
Balanced Accuracy 0.4995908



[[1]]  
[1] 0.9720727  
  
[[1]]  
[1] 0.9361609  
  
[[1]]  
[1] 0.9620074  
  
[[1]]  
[1] 0.9883794  
  
[[1]]  
[1] 0.9976814

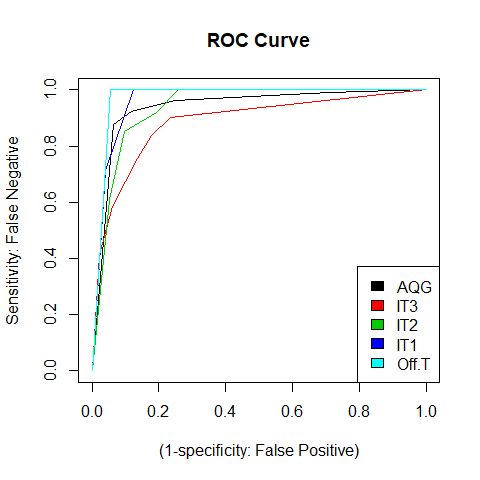
R Partition

> model.rpart <- caret::train(myModel, method = "rpart", data = trData)  
> model.rpart

CART   
  
2870 samples  
 5 predictor  
 5 classes: 'AQG', 'IT3', 'IT2', 'IT1', 'Off.T'   
  
No pre-processing  
Resampling: Bootstrapped (25 reps)   
Summary of sample sizes: 2870, 2870, 2870, 2870, 2870, 2870, ...   
Resampling results across tuning parameters:  
  
 cp Accuracy Kappa   
 0.02731591 0.8159878 0.6063638  
 0.03859857 0.8051270 0.5812293  
 0.27197150 0.7721273 0.4364157  
  
Accuracy was used to select the optimal model using the largest value.  
The final value used for the model was cp = 0.02731591.

> test.rpart <- test\_rocFUN(model.rpart, teData)

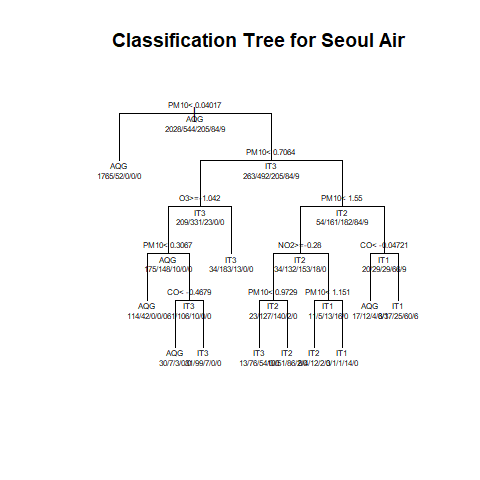
Confusion Matrix and Statistics  
  
 Reference  
Prediction AQG IT3 IT2 IT1 Off.T  
 AQG 802 42 0 0 0  
 IT3 49 136 13 0 0  
 IT2 9 39 53 10 0  
 IT1 8 15 21 25 3  
 Off.T 0 0 0 0 0  
  
Overall Statistics  
   
 Accuracy : 0.8294   
 95% CI : (0.8071, 0.8501)  
 No Information Rate : 0.7086   
 P-Value [Acc > NIR] : < 2.2e-16   
   
 Kappa : 0.6394   
 Mcnemar's Test P-Value : NA   
  
Statistics by Class:  
  
 Class: AQG Class: IT3 Class: IT2 Class: IT1  
Sensitivity 0.9240 0.5862 0.60920 0.71429  
Specificity 0.8824 0.9376 0.94903 0.96050  
Pos Pred Value 0.9502 0.6869 0.47748 0.34722  
Neg Pred Value 0.8268 0.9065 0.96948 0.99133  
Prevalence 0.7086 0.1894 0.07102 0.02857  
Detection Rate 0.6547 0.1110 0.04327 0.02041  
Detection Prevalence 0.6890 0.1616 0.09061 0.05878  
Balanced Accuracy 0.9032 0.7619 0.77911 0.83739  
 Class: Off.T  
Sensitivity 0.000000  
Specificity 1.000000  
Pos Pred Value NaN  
Neg Pred Value 0.997551  
Prevalence 0.002449  
Detection Rate 0.000000  
Detection Prevalence 0.000000  
Balanced Accuracy 0.500000



[[1]]  
[1] 0.938782  
  
[[1]]  
[1] 0.8845301  
  
[[1]]  
[1] 0.9386249  
  
[[1]]  
[1] 0.962485  
  
[[1]]  
[1] 0.9717676

R Partition Plot Tree

> model.rpart.tree <- rpart(myModel, data = trData)  
>   
> plot(model.rpart.tree, uniform=TRUE,   
+ main="Classification Tree for Seoul Air", margin=0.1)  
> text(model.rpart.tree, use.n=TRUE, all=TRUE, cex=.5)



R Part Testing

> Ycol <- which(names(trData) == "Y")  
> predict.Y <- predict(model.rpart, newdata = teData[, - Ycol], type="raw")  
> confusionMatrix(table(predict.Y, teData$Y))

Confusion Matrix and Statistics  
  
   
predict.Y AQG IT3 IT2 IT1 Off.T  
 AQG 802 42 0 0 0  
 IT3 49 136 13 0 0  
 IT2 9 39 53 10 0  
 IT1 8 15 21 25 3  
 Off.T 0 0 0 0 0  
  
Overall Statistics  
   
 Accuracy : 0.8294   
 95% CI : (0.8071, 0.8501)  
 No Information Rate : 0.7086   
 P-Value [Acc > NIR] : < 2.2e-16   
   
 Kappa : 0.6394   
 Mcnemar's Test P-Value : NA   
  
Statistics by Class:  
  
 Class: AQG Class: IT3 Class: IT2 Class: IT1  
Sensitivity 0.9240 0.5862 0.60920 0.71429  
Specificity 0.8824 0.9376 0.94903 0.96050  
Pos Pred Value 0.9502 0.6869 0.47748 0.34722  
Neg Pred Value 0.8268 0.9065 0.96948 0.99133  
Prevalence 0.7086 0.1894 0.07102 0.02857  
Detection Rate 0.6547 0.1110 0.04327 0.02041  
Detection Prevalence 0.6890 0.1616 0.09061 0.05878  
Balanced Accuracy 0.9032 0.7619 0.77911 0.83739  
 Class: Off.T  
Sensitivity 0.000000  
Specificity 1.000000  
Pos Pred Value NaN  
Neg Pred Value 0.997551  
Prevalence 0.002449  
Detection Rate 0.000000  
Detection Prevalence 0.000000  
Balanced Accuracy 0.500000

Support Vector Machine with Linear

> model.svm <- caret::train(myModel, method = "svmLinear", data = trData)  
> model.svm

Support Vector Machines with Linear Kernel   
  
2870 samples  
 5 predictor  
 5 classes: 'AQG', 'IT3', 'IT2', 'IT1', 'Off.T'   
  
No pre-processing  
Resampling: Bootstrapped (25 reps)   
Summary of sample sizes: 2870, 2870, 2870, 2870, 2870, 2870, ...   
Resampling results:  
  
 Accuracy Kappa   
 0.8085434 0.5659475  
  
Tuning parameter 'C' was held constant at a value of 1

> Y.colNumber <- which(names(teData) == "Y")  
> predict.Y2 <- predict(model.svm, teData[, -Y.colNumber])   
> print(confusionMatrix(predict.Y2, as.factor(teData$Y)))

Confusion Matrix and Statistics  
  
 Reference  
Prediction AQG IT3 IT2 IT1 Off.T  
 AQG 826 58 10 2 0  
 IT3 37 164 64 19 1  
 IT2 0 0 0 0 0  
 IT1 5 10 13 14 2  
 Off.T 0 0 0 0 0  
  
Overall Statistics  
   
 Accuracy : 0.8196   
 95% CI : (0.7969, 0.8407)  
 No Information Rate : 0.7086   
 P-Value [Acc > NIR] : < 2.2e-16   
   
 Kappa : 0.5868   
 Mcnemar's Test P-Value : NA   
  
Statistics by Class:  
  
 Class: AQG Class: IT3 Class: IT2 Class: IT1  
Sensitivity 0.9516 0.7069 0.00000 0.40000  
Specificity 0.8039 0.8781 1.00000 0.97479  
Pos Pred Value 0.9219 0.5754 NaN 0.31818  
Neg Pred Value 0.8723 0.9277 0.92898 0.98222  
Prevalence 0.7086 0.1894 0.07102 0.02857  
Detection Rate 0.6743 0.1339 0.00000 0.01143  
Detection Prevalence 0.7314 0.2327 0.00000 0.03592  
Balanced Accuracy 0.8778 0.7925 0.50000 0.68739  
 Class: Off.T  
Sensitivity 0.000000  
Specificity 1.000000  
Pos Pred Value NaN  
Neg Pred Value 0.997551  
Prevalence 0.002449  
Detection Rate 0.000000  
Detection Prevalence 0.000000  
Balanced Accuracy 0.500000

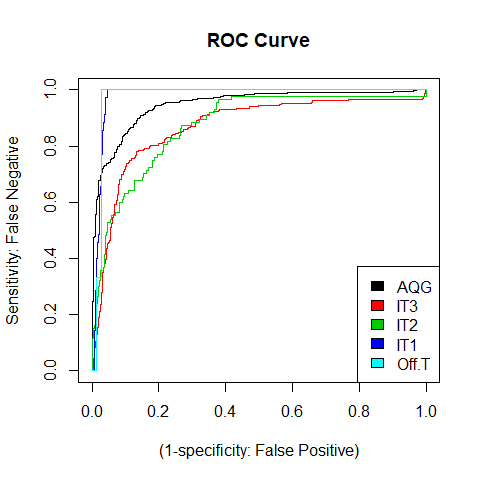
Linear Discriminant Analysis

> model.lda <- caret::train(myModel, method = "lda", data = trData)  
> model.lda

Linear Discriminant Analysis   
  
2870 samples  
 5 predictor  
 5 classes: 'AQG', 'IT3', 'IT2', 'IT1', 'Off.T'   
  
No pre-processing  
Resampling: Bootstrapped (25 reps)   
Summary of sample sizes: 2870, 2870, 2870, 2870, 2870, 2870, ...   
Resampling results:  
  
 Accuracy Kappa   
 0.8036403 0.5530759

> test.lda <- test\_rocFUN(model.lda, teData)

Confusion Matrix and Statistics  
  
 Reference  
Prediction AQG IT3 IT2 IT1 Off.T  
 AQG 830 70 13 10 0  
 IT3 29 145 47 1 0  
 IT2 1 7 13 3 0  
 IT1 4 3 12 17 2  
 Off.T 4 7 2 4 1  
  
Overall Statistics  
   
 Accuracy : 0.8212   
 95% CI : (0.7986, 0.8423)  
 No Information Rate : 0.7086   
 P-Value [Acc > NIR] : < 2.2e-16   
   
 Kappa : 0.5837   
 Mcnemar's Test P-Value : 6.197e-13   
  
Statistics by Class:  
  
 Class: AQG Class: IT3 Class: IT2 Class: IT1  
Sensitivity 0.9562 0.6250 0.14943 0.48571  
Specificity 0.7395 0.9225 0.99033 0.98235  
Pos Pred Value 0.8992 0.6532 0.54167 0.44737  
Neg Pred Value 0.8742 0.9133 0.93838 0.98484  
Prevalence 0.7086 0.1894 0.07102 0.02857  
Detection Rate 0.6776 0.1184 0.01061 0.01388  
Detection Prevalence 0.7535 0.1812 0.01959 0.03102  
Balanced Accuracy 0.8479 0.7737 0.56988 0.73403  
 Class: Off.T  
Sensitivity 0.3333333  
Specificity 0.9860884  
Pos Pred Value 0.0555556  
Neg Pred Value 0.9983430  
Prevalence 0.0024490  
Detection Rate 0.0008163  
Detection Prevalence 0.0146939  
Balanced Accuracy 0.6597109



[[1]]  
[1] 0.9474919  
  
[[1]]  
[1] 0.8688448  
  
[[1]]  
[1] 0.8749167  
  
[[1]]  
[1] 0.9792797  
  
[[1]]  
[1] 0.9765412

Neural Network

> model.nn <- caret::train(myModel, method = "nnet", data = trData, trace = FALSE, maxit = 100)  
> model.nn

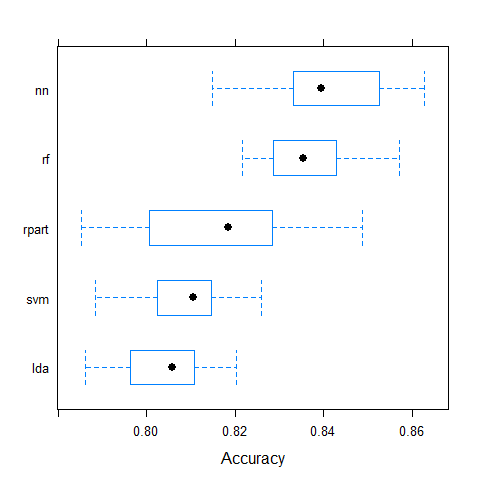
Neural Network   
  
2870 samples  
 5 predictor  
 5 classes: 'AQG', 'IT3', 'IT2', 'IT1', 'Off.T'   
  
No pre-processing  
Resampling: Bootstrapped (25 reps)   
Summary of sample sizes: 2870, 2870, 2870, 2870, 2870, 2870, ...   
Resampling results across tuning parameters:  
  
 size decay Accuracy Kappa   
 1 0e+00 0.8239285 0.6288082  
 1 1e-04 0.8248812 0.6245881  
 1 1e-01 0.8079111 0.5707750  
 3 0e+00 0.8354602 0.6439093  
 3 1e-04 0.8364647 0.6447281  
 3 1e-01 0.8377700 0.6457818  
 5 0e+00 0.8395773 0.6508696  
 5 1e-04 0.8384963 0.6470422  
 5 1e-01 0.8405514 0.6484169  
  
Accuracy was used to select the optimal model using the largest value.  
The final values used for the model were size = 5 and decay = 0.1.

Compare Accuracy

> model.list <- resamples(list(rpart=model.rpart, rf=model.rf, svm = model.svm, lda=model.lda, nn = model.nn))  
> summary(model.list)

Call:  
summary.resamples(object = model.list)  
  
Models: rpart, rf, svm, lda, nn   
Number of resamples: 25   
  
Accuracy   
 Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
rpart 0.7851638 0.8005780 0.8185220 0.8159878 0.8283582 0.8488593 0  
rf 0.8217350 0.8286814 0.8353949 0.8365887 0.8429672 0.8570093 0  
svm 0.7884250 0.8023599 0.8105561 0.8085434 0.8145387 0.8258427 0  
lda 0.7861873 0.7962441 0.8058712 0.8036403 0.8107345 0.8202460 0  
nn 0.8148148 0.8331762 0.8394366 0.8405514 0.8525214 0.8628405 0  
  
Kappa   
 Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
rpart 0.5334822 0.5915511 0.6083653 0.6063638 0.6290484 0.6624129 0  
rf 0.6199019 0.6278482 0.6397750 0.6454015 0.6543859 0.6904171 0  
svm 0.5237096 0.5489129 0.5718364 0.5659475 0.5793284 0.6052963 0  
lda 0.5167399 0.5435694 0.5524764 0.5530759 0.5666331 0.5833202 0  
nn 0.5836316 0.6366977 0.6501617 0.6484169 0.6639214 0.6957958 0

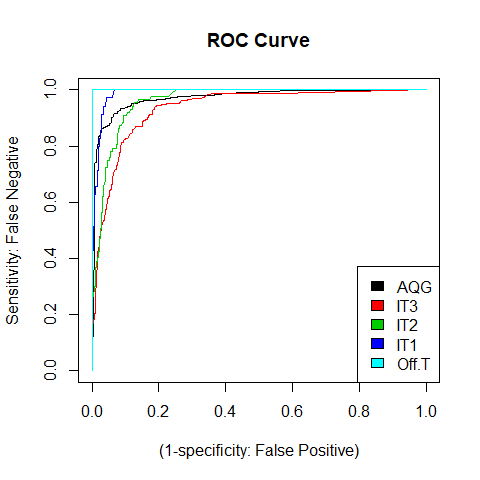
> bwplot(model.list, metric = "Accuracy")



K-fold cross validation

> my\_trControl1 <- trainControl(method = "cv",  
+ number = 5)  
> model.rf1 <- train(Y ~ ., method = "rf", data = trData, trControl = my\_trControl1)  
> test.rf1 <- test\_rocFUN(model.rf1, teData)

Confusion Matrix and Statistics  
  
 Reference  
Prediction AQG IT3 IT2 IT1 Off.T  
 AQG 827 44 5 0 0  
 IT3 39 169 31 8 0  
 IT2 2 17 40 3 0  
 IT1 0 2 11 23 3  
 Off.T 0 0 0 1 0  
  
Overall Statistics  
   
 Accuracy : 0.8645   
 95% CI : (0.844, 0.8832)  
 No Information Rate : 0.7086   
 P-Value [Acc > NIR] : < 2.2e-16   
   
 Kappa : 0.6993   
 Mcnemar's Test P-Value : NA   
  
Statistics by Class:  
  
 Class: AQG Class: IT3 Class: IT2 Class: IT1  
Sensitivity 0.9528 0.7284 0.45977 0.65714  
Specificity 0.8627 0.9215 0.98067 0.98655  
Pos Pred Value 0.9441 0.6842 0.64516 0.58974  
Neg Pred Value 0.8825 0.9356 0.95959 0.98988  
Prevalence 0.7086 0.1894 0.07102 0.02857  
Detection Rate 0.6751 0.1380 0.03265 0.01878  
Detection Prevalence 0.7151 0.2016 0.05061 0.03184  
Balanced Accuracy 0.9078 0.8249 0.72022 0.82185  
 Class: Off.T  
Sensitivity 0.0000000  
Specificity 0.9991817  
Pos Pred Value 0.0000000  
Neg Pred Value 0.9975490  
Prevalence 0.0024490  
Detection Rate 0.0000000  
Detection Prevalence 0.0008163  
Balanced Accuracy 0.4995908

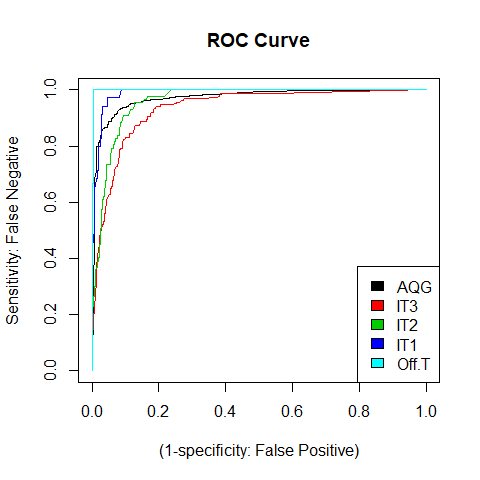


[[1]]  
[1] 0.9713256  
  
[[1]]  
[1] 0.9353535  
  
[[1]]  
[1] 0.9612044  
  
[[1]]  
[1] 0.9885954  
  
[[1]]  
[1] 0.9986361

Cross Validation

> my\_trControl2 <- trainControl(method = "repeatedcv",  
+ number = 5,  
+ repeats = 3)  
> model.rf2 <- train(Y ~ ., method = "rf", data = trData, trControl = my\_trControl2)  
> test.rf2 <- test\_rocFUN(model.rf2, teData)

Confusion Matrix and Statistics  
  
 Reference  
Prediction AQG IT3 IT2 IT1 Off.T  
 AQG 827 46 6 0 0  
 IT3 39 167 32 7 0  
 IT2 2 17 39 4 0  
 IT1 0 2 10 23 3  
 Off.T 0 0 0 1 0  
  
Overall Statistics  
   
 Accuracy : 0.862   
 95% CI : (0.8415, 0.8809)  
 No Information Rate : 0.7086   
 P-Value [Acc > NIR] : < 2.2e-16   
   
 Kappa : 0.6929   
 Mcnemar's Test P-Value : NA   
  
Statistics by Class:  
  
 Class: AQG Class: IT3 Class: IT2 Class: IT1  
Sensitivity 0.9528 0.7198 0.44828 0.65714  
Specificity 0.8543 0.9215 0.97979 0.98739  
Pos Pred Value 0.9408 0.6816 0.62903 0.60526  
Neg Pred Value 0.8815 0.9337 0.95873 0.98989  
Prevalence 0.7086 0.1894 0.07102 0.02857  
Detection Rate 0.6751 0.1363 0.03184 0.01878  
Detection Prevalence 0.7176 0.2000 0.05061 0.03102  
Balanced Accuracy 0.9036 0.8206 0.71403 0.82227  
 Class: Off.T  
Sensitivity 0.0000000  
Specificity 0.9991817  
Pos Pred Value 0.0000000  
Neg Pred Value 0.9975490  
Prevalence 0.0024490  
Detection Rate 0.0000000  
Detection Prevalence 0.0008163  
Balanced Accuracy 0.4995908



[[1]]  
[1] 0.9713385  
  
[[1]]  
[1] 0.9351452  
  
[[1]]  
[1] 0.9610175  
  
[[1]]  
[1] 0.9881032  
  
[[1]]  
[1] 0.9978178