Gradient Boosting

Gus Vu

2023-05-23

# Gradient Boosting

A GBM model is comprises of several underlying ensemble models like weak decision trees. These decision trees combine together to form a strong model of gradient boosting.

library(gbm, quietly=TRUE)

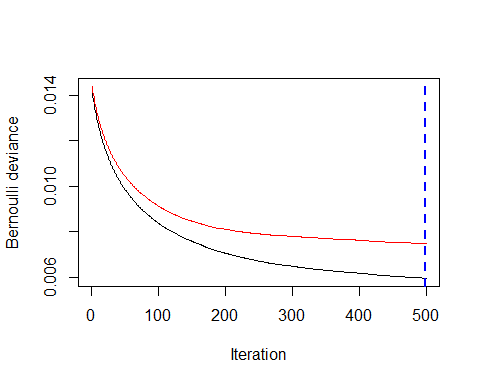
## Warning: package 'gbm' was built under R version 4.2.3

## Loaded gbm 2.1.8.1

# Get the time to train the GBM model  
system.time(  
 model\_gbm <- gbm(Class ~ .  
 , distribution = "bernoulli"  
 , data = rbind(train\_data, test\_data)  
 , n.trees = 500  
 , interaction.depth = 3  
 , n.minobsinnode = 100  
 , shrinkage = 0.01  
 , bag.fraction = 0.5  
 , train.fraction = nrow(train\_data) / (nrow(train\_data) + nrow(test\_data))  
)  
)

## user system elapsed   
## 237.97 0.25 238.33

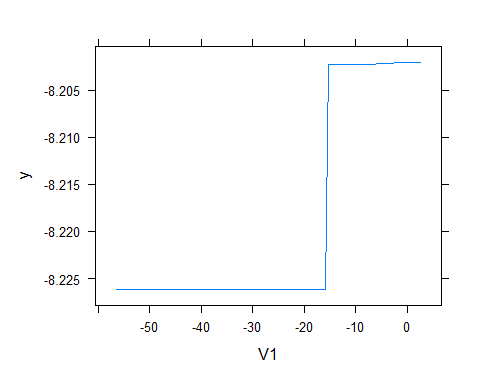
# Determine best iteration based on test data  
gbm.iter = gbm.perf(model\_gbm, method = "test")



The iteration vs Bernuli deviance shows us the performance of the model vs the number of trees. If we had an infinite amount of trees the model would perform well but would over fit. The blue line shows us the optimal amount of trees.

Plot the gbm model

model.influence = relative.influence(model\_gbm, n.trees = gbm.iter, sort. = TRUE)  
  
plot(model\_gbm)

 Partial dependence plots (PDP) show the dependence between the target response and a set of input features of interest, marginalizing over the values of all other input features (the ‘complement’ features). Intuitively, we can interpret the partial dependence as the expected target response as a function of the input features of interest.

Plot and calculate AUC on test data

library(pROC)

## Warning: package 'pROC' was built under R version 4.2.2

## Type 'citation("pROC")' for a citation.

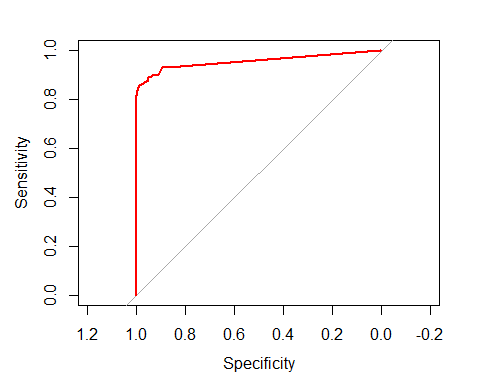
##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

gbm\_test = predict(model\_gbm, newdata = test\_data, n.trees = gbm.iter)  
gbm\_auc = roc(test\_data$Class, gbm\_test, plot = TRUE, col = "red")

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases



gbm\_auc$auc

## Area under the curve: 0.9549

We see that the area under the curve is much more acurate than a random choice of 0.5.