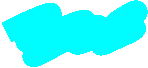
Creating a very basic model in R using the caret package:



1.Creating a training data set , we use 75% of the total values there in the spam dataset to create a training dataset and 25% of the total values to create a testing dataset.

inTrain<-createDataPartition(y=spam$type,p=0.75,list=FALSE)

y=spam$type means that we are creating samples on the basis of spam type .

p= 0.75 means we use 75% of the total values there in the spam dataset to create a training dataset and 25% of the total values to create a testing dataset.

list= false means we do not want the datatype of inTrain to be list.

2. Assigning training and testing values to different variables:

training<-spam[inTrain,]

testing<-spam[-inTrain,]

3.

dim(training)

[1] 3451 58

dim(testing)

[1] 1150 58

3451+1150

[1] 4601

3451/4601

[1] 0.7500543 #75% of the total values there in the spam dataset are used to create a training dataset

4. We train our model to predict type of mail that is spam or non spam on the basis of all the predictors there in the training dataset .The dot in type ~. indicates that we want all our predictors(columns there in the training dataset) to predict type.

The method we use is generalized linear models or glm.

modelFit<-train(type~.,data=training,method="glm")

5. We test our model on the testing dataset .

predictions<-predict(modelFit,newdata=testing)

6. Now check the results of our prediction of the testing data using a confusion matrix. Here we compare actual values of email type there in the testing dataset to the values of email type we have predicted for the testing dataset.

confusionMatrix(predictions,testing$type)

Confusion Matrix and Statistics

Reference

Prediction nonspam spam

nonspam 667 44

spam 30 409

Accuracy : 0.9357

95% CI : (0.9199, 0.9491)

No Information Rate : 0.6061

P-Value [Acc > NIR] : <2e-16

Kappa : 0.8645

Mcnemar's Test P-Value : 0.1307

Sensitivity : 0.9570

Specificity : 0.9029

Pos Pred Value : 0.9381

Neg Pred Value : 0.9317

Prevalence : 0.6061

Detection Rate : 0.5800

Detection Prevalence : 0.6183

Balanced Accuracy : 0.9299

'Positive' Class : nonspam

Preprocessing in R:



Data Slicing:

