Example

The dataset (training) is a collection of data about some of the passengers (889 to be precise), and the goal of the competition is to predict the survival (either 1 if the passenger survived or 0 if they did not) based on some features such as the class of service, the sex, the age etc.

The training set will be used to fit our model which we will be testing over the testing set.

train <- data[1:800,]

test <- data[801:889,]

model <- glm(Survived ~.,family=binomial(link='logit'),data=train)

summary(model)

Assessing the predictive ability of the model

In the steps above, we briefly evaluated the fitting of the model, now we would like to see how the model is doing when predicting y on a new set of data. By setting the parameter type='response', R will output probabilities in the form of P(y=1|X). Our decision boundary will be 0.5. If P(y=1|X) > 0.5 then y = 1 otherwise y=0. Note that for some applications different thresholds could be a better option.

# If prob > 0.5 then 1, else 0. Threshold can be set for better results

fitted.results <- predict(model,newdata=subset(test,select=c(2,3,4,5,6,7,8)),type='response')

fitted.results <- ifelse(fitted.results > 0.5,1,0)

misClasificError <- mean(fitted.results != test$Survived)

print(paste('Accuracy',1-misClasificError))

"Accuracy 0.842696629213483"

As a last step, we are going to plot the ROC curve and calculate the AUC (area under the curve) which are typical performance measurements for a binary classifier.  
The ROC is a curve generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC is the area under the ROC curve. As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5.

# Confusion matrix

library(caret)

confusionMatrix(data=fitted.results, reference=test$Survived)

library(ROCR)

# ROC and AUC

p <- predict(model, newdata=subset(test,select=c(2,3,4,5,6,7,8)), type="response")

pr <- prediction(p, test$Survived)

# TPR = sensitivity, FPR=specificity

prf <- performance(pr, measure = "tpr", x.measure = "fpr")

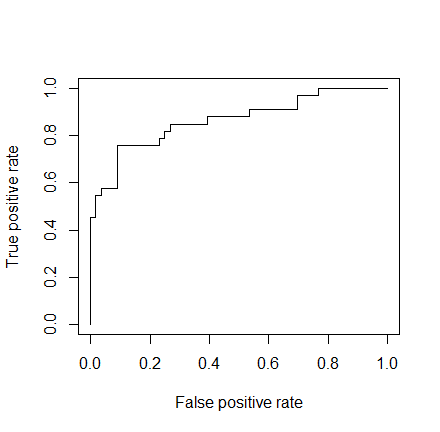
plot(prf)

auc <- performance(pr, measure = "auc")

auc <- [auc@y.values[[1](mailto:auc@y.values[[1)]]

auc

0.8647186

And here is the ROC plot:  
[](http://i2.wp.com/datascienceplus.com/wp-content/uploads/2015/09/Rplot011.png)

About prediction:

problem: in the data set [mtcars](http://www.r-tutor.com/r-introduction/data-frame), estimate the probability of a vehicle being fitted with a manual transmission if it has a 120hp engine and weights 2800 lbs.

solutions:

> am.glm = glm(formula=am ~ hp + wt, data=mtcars, family=binomial)

then wrap the test parameters inside a [data frame](http://www.r-tutor.com/r-introduction/data-frame) newdata

> newdata = data.frame(hp=120, wt=2.8)

> predict(am.glm, newdata, type="response")

0.64181

For an automobile with 120hp engine and 2800 lbs weight, the probability of it being fitted with a manual transmission is about 64%.