**Phase3: Development part1**

**Loading and Processing Dataset RoC Analysis**

# **Processing data in a Dataset**

* Datasets provides many methods to modify a Dataset, be it to reorder, split or shuffle the dataset or to apply data processing functions or evaluation functions to its elements.
* We’ll start by presenting the methods which change the order or number of elements before presenting methods which access and can change the content of the elements themselves.

**Statistics**

AUC, negative group, missing values, positive classification, cutoff value, strength of conviction, two-sided asymptotic confidence interval, distribution, standard error, independent-group design, paired-sample design, nonparametric assumption, bi-negative exponential distribution assumption, midpoint, cut point, PR curve, stepwise interpolation, asymptotic significance (2-tail), Sensitivity and (1-Specicity), Precision and Recall.

**Methods**

The areas under two ROC curves, that are generated from either independent groups or paired subjects, are compared. Comparing two ROC curves can provide more information in the accuracy resulted from two comparative diagnostic approaches.

**How do you do a roc analysis?**

**Obtaining an ROC analysis**

* Click Classification to define the cutoff value, test direction, and standard error of area under the curve.
* Click Statistics to select which statistics to include in the procedure.
* Click Plots to define plotting for the ROC and Precision-Recall curves.

**Is roc good for imbalanced datasets?**

* ROC curves can sometimes be misleading in some very imbalanced applications. A ROC curve can still look pretty good (ie better than random) while misclassifying most or all of the minority class. In contrast, PR curves are specifically tailored for the detection of rare events and are pretty useful in those scenarios.

**How to Create a ROC Curve in Excel (Step-by-Step)**

1. Step 1: Enter the Data. First, let's enter some raw data:
2. Step 2: Calculate the Cumulative Data. ...
3. Step 3: Calculate False Positive Rate & True Positive Rate. ...
4. Step 4: Create the ROC Curve. ...
5. Step 5: Calculate the AUC

Data considerations

* **Data**

PR curves plot Precision versus Recall, and tend to be more informative when the observed data samples are highly skewed. A simple linear interpolation may mistakenly yield an overly-optimistic estimate of a PR curve.

* **Assumptions**

The prediction will be in the correct order when a test variable is observed for one subject that is randomly selected from the case group and the other is randomly selected from the control group. Each defined group will contain at least one valid observation. Only a single grouping variable is used for a single procedure.

**How do you manually plot a ROC curve?**

All we need to do, based on different threshold values, is to compute True Positive Rate (TPR) and False Positive Rate (FPR) values for each of the thresholds and then plot TPR against FPR. When you obtain True Positive Rate and False Positive Rate for each of thresholds, all you need to is plot them! is the formula fo

**What is the formula for calculating ROC?**

In finance, the calculation for ROC can also be computed as a return over time, in that it can takes the current value of a stock or index and divides it by the value from an earlier period. Subtract one and multiply the resulting number by 100 to give it a percentage representation.

**Limitations of ROC curves**

Confidence scores used to build ROC curves may be difficult to assign. False-positive and false-negative diagnoses have different misclassification costs. Excessive ROC curve extrapolation is undesirable. Net benefit methods may provide more meaningful and clinically interpretable results than ROC AUC.

**Two parameters of the ROC curve**

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: **True Positive Rate. False Positive Rate.**

**Algorithms**

1. Step 1 - Load the necessary libraries. ...
2. Step 2 - Read a csv dataset. ...
3. Step 3- Create train and test dataset. ...
4. Step 4 -Create a model for logistics using the training dataset. ...
5. Step 5- Make predictions on the model using the test dataset. ...
6. Step 6 - Model Diagnostics. ...
7. Step 7 - Create AUC and ROC for test data(pROC lib)

### **Key metrics on the ROC curve include:**

1. **True Positive Rate (TPR)**: This is the ratio of correctly predicted positive instances to the total number of actual positives. It corresponds to the y-axis of the ROC curve.
2. **False Positive Rate (FPR)**: This is the ratio of incorrectly predicted positive instances to the total number of actual negatives. It corresponds to the x-axis of the ROC curve.

### **Constructing a ROC Curve**

Let's break down the steps involved in constructing a ROC curve using Python and sci-kit-learn:

### **Step 1: Importing Necessary Packages**

I started by importing the essential libraries, including NumPy, Pandas, CSV, random, and Matplotlib. These libraries facilitate data manipulation, analysis, and visualization.

### **Step 2: Generating Synthetic Data for ROC Curve Analysis**

I then generate a synthetic dataset. The dataset I created had two columns: 'probability' (predicted probabilities) and 'actual\_label' (true labels). This dataset serves as the foundation for constructing the ROC curve.

### **Step 3: Loading the Synthetic Data into a Data Frame**

I proceeded to load the generated data into a Pandas data frame to facilitate data manipulation and analysis.

### **Step 4: Visualizing Data Distribution and Overlapping in Scatter Plot**

Next, I create a scatter plot to visualize the relationship between 'probability' values and 'actual\_label' outcomes. This step illustrates the challenge of finding a single separation line due to data overlap.

### **Step 5: Constructing and Analyzing the ROC Curve**

Below are the steps to create the ROC Curve:

* Extract model outputs and actual labels from the data.
* Compute ROC metrics using the roc\_curve function, including FPRs, TPRs, and thresholds.
* Calculate AUROC (Area Under the ROC Curve) using the AUC function.
* Visualize the ROC curve with TPR vs. FPR. Include a baseline for random guessing.
* Display the calculated AUROC value to summarize the model's overall performance.

## [Method I: Using plot() function](https://www.digitalocean.com/community/tutorials/plot-roc-curve-r-programming#method-i-using-plot-function)

rm(list = ls())

#Setting the working directory

setwd("D:/Edwisor\_Project - Loan\_Defaulter/")

getwd()

#Load the dataset

dta = read.csv("bank-loan.csv",header=TRUE)

### Data SAMPLING ####

library(caret)

set.seed(101)

split = createDataPartition(data$default, p = 0.80, list = FALSE)

train\_data = data[split,]

test\_data = data[-split,]

#error metrics -- Confusion Matrix

err\_metric=function(CM)

{

TN =CM[1,1]

TP =CM[2,2]

FP =CM[1,2]

FN =CM[2,1]

precision =(TP)/(TP+FP)

recall\_score =(FP)/(FP+TN)

f1\_score=2\*((precision\*recall\_score)/(precision+recall\_score))

accuracy\_model =(TP+TN)/(TP+TN+FP+FN)

False\_positive\_rate =(FP)/(FP+TN)

False\_negative\_rate =(FN)/(FN+TP)

print(paste("Precision value of the model: ",round(precision,2)))

print(paste("Accuracy of the model: ",round(accuracy\_model,2)))

print(paste("Recall value of the model: ",round(recall\_score,2)))

print(paste("False Positive rate of the model: ",round(False\_positive\_rate,2)))

print(paste("False Negative rate of the model: ",round(False\_negative\_rate,2)))

print(paste("f1 score of the model: ",round(f1\_score,2)))

}

# 1. Logistic regression

logit\_m =glm(formula = default~. ,data =train\_data ,family='binomial')

summary(logit\_m)

logit\_P = predict(logit\_m , newdata = test\_data[-13] ,type = 'response' )

logit\_P <- ifelse(logit\_P > 0.5,1,0) # Probability check

CM= table(test\_data[,13] , logit\_P)

print(CM)

err\_metric(CM)

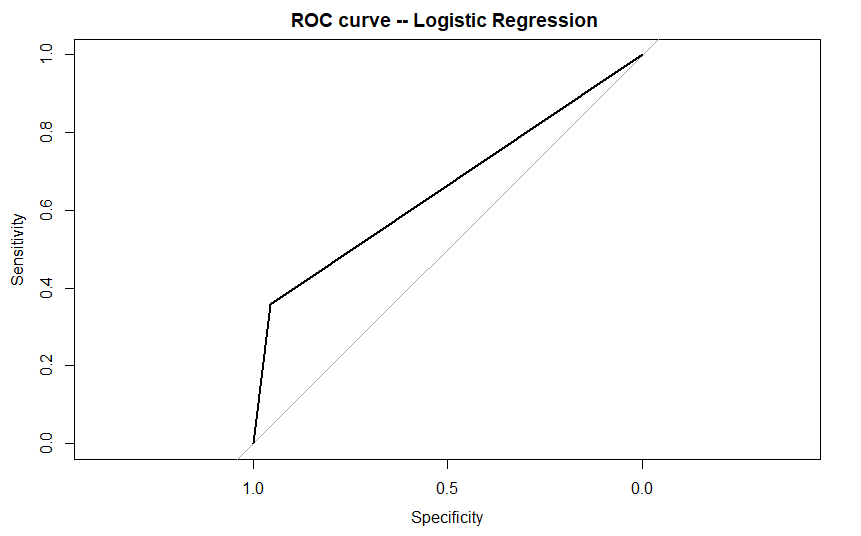
#ROC-curve using pROC library

library(pROC)

roc\_score=roc(test\_data[,13], logit\_P) #AUC score

plot(roc\_score ,main ="ROC curve -- Logistic Regression ")

**Output**



## [Method II: Using roc.plot() function](https://www.digitalocean.com/community/tutorials/plot-roc-curve-r-programming#method-ii-using-roc-plot-function)

**install.packages("verification")**

**library(verification)**

**x<- c(0,0,0,1,1,1)**

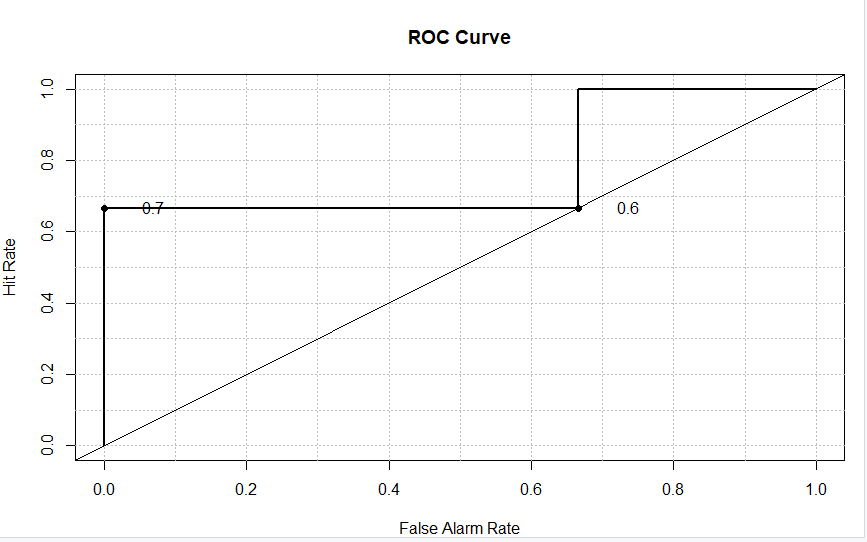
**y<- c(.7, .7, 0, 1,5,.6)**

**data<-data.frame(x,y)**

**names(data)<-c("yes","no")**

**roc.plot(data$yes, data$no)**

**Output**

****

**DATASET EXAMPLE**

