***affairs***

**predict the probability of extra marital affair.**

**Inferences from the Data Set:**

Data Set talks about the extra marital affair is a binary variable with respect to around 9 Independent variables & 601 observations.

**Columns:**

affairs

gender

age

yearsmarried

children

religiousness

education

occupation

rating

**Data Set Size:** 601

Data give is found to be a binary data for which a logistic regression can be performed getting deeper into the data analysis and its behavior

summary(Affairs)

summary(Affairs) talks about the about columns of dataset i.e; range of column(min-max),

Mean, Median, 1st Quartile and 3rd quartile

**affairs:**

Ranges between 0 -12

For this Affairs the mean is 1.456, it is just the average of the Affairs data

The median for the given data is 0, it speaks about the center of data

A comparison between mean and median tell us that data is skewed (median=0<mean-1.456), if data was not skewed, we would have considered mean but hear it is skewed so we take Median to talk about data.

The Data is Right Skewed, Skewness= 2.341136

**age:**

Ranges between 17.50 - 57

For this age the mean is 32.49 , it is just the average of the age data

The median for the given data is 32, it speaks about the center of data

A comparison between mean and median tell us that data is skewed (median=32 < mean-32.49), if data was not skewed, we would have considered mean but hear it is skewed so we take Median to talk about data.

Skewness =0.88

Iike above all columns of Affairs data set below are summary of Affairs dataset.

summary(Affairs)

affairs gender age yearsmarried children

Min. : 0.000 female:315 Min. :17.50 Min. : 0.125 no :171

1st Qu.: 0.000 male :286 1st Qu.:27.00 1st Qu.: 4.000 yes:430

Median : 0.000 Median :32.00 Median : 7.000

Mean : 1.456 Mean :32.49 Mean : 8.178

3rd Qu.: 0.000 3rd Qu.:37.00 3rd Qu.:15.000

Max. :12.000 Max. :57.00 Max. :15.000

religiousness education occupation rating

Min. :1.000 Min. : 9.00 Min. :1.000 Min. :1.000

1st Qu.:2.000 1st Qu.:14.00 1st Qu.:3.000 1st Qu.:3.000

Median :3.000 Median :16.00 Median :5.000 Median :4.000

Mean :3.116 Mean :16.17 Mean :4.195 Mean :3.932

3rd Qu.:4.000 3rd Qu.:18.00 3rd Qu.:6.000 3rd Qu.:5.000

Max. :5.000 Max. :20.00 Max. :7.000 Max. :5.000

transform affairs column into a dichotomous factor called ynaffair with the following code.

Affairs$ynaffair[Affairs$affairs > 0] <- 1

Affairs$ynaffair[Affairs$affairs == 0] <- 0

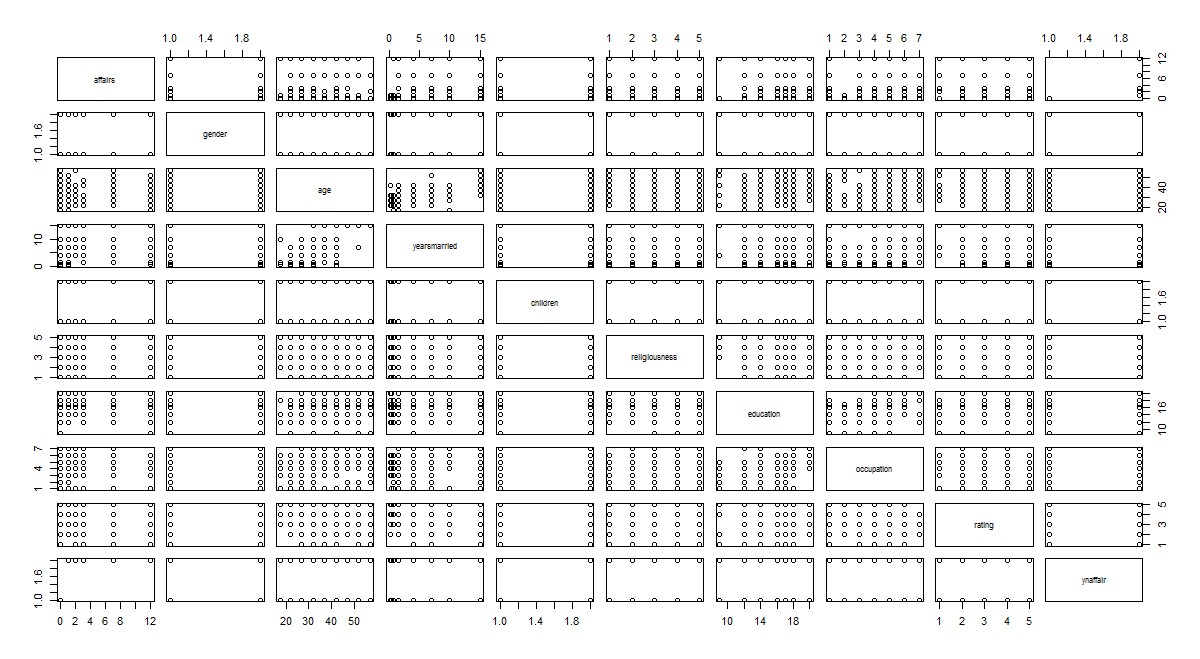
Affairs$ynaffair <- factor(Affairs$ynaffair,

levels=c(0,1),

labels=c("No","Yes"))

because the affairs column is talks about extra marital affair which is binary in nature.

**plot(Affairs)**



The above diagram infer that the Affairs of dataset.

We use **glm() function from Base Package in R-Studio** to estimate the affairs using the other 9 independent variables gender, age, yearsmarried, children, religiousness, education, occupation, rating whereas in **python sm.logit() is used from the statsmodels.formula.api package**

summary(model)

Call:

glm(formula = ynaffair ~ gender + age + yearsmarried + children +

religiousness + education + occupation + rating, family = "binomial",

data = Affairs)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5713 -0.7499 -0.5690 -0.2539 2.5191

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.37726 0.88776 1.551 0.120807

gendermale 0.28029 0.23909 1.172 0.241083

age -0.04426 0.01825 -2.425 0.015301 \*

yearsmarried 0.09477 0.03221 2.942 0.003262 \*\*

childrenyes 0.39767 0.29151 1.364 0.172508

religiousness -0.32472 0.08975 -3.618 0.000297 \*\*\*

education 0.02105 0.05051 0.417 0.676851

occupation 0.03092 0.07178 0.431 0.666630

rating -0.46845 0.09091 -5.153 2.56e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 675.38 on 600 degrees of freedom

Residual deviance: 609.51 on 592 degrees of freedom

AIC: 627.51

Number of Fisher Scoring iterations: 4

**P-values:**

coefficient p-values are used to determine which terms to keep in the model

**Confusion matrix table:**

In logistic regression that affairs values we are predicting in the equation.

prob <- predict(model,Affairs,type="response")

Basically the output is 1 r 0 whether having extra marital affair or not. We need to convert the data into binary class.

In order to convert we use cut-off value is 0.5( threshold value)

confusion <- table(prob > 0.5, Affairs$ynaffair)

confusion

No Yes

FALSE 435 125

TRUE 16 25

**Accuracy f mde:**

Accuracy f mde is confirmed frm cnfusin matrix tabe i.e; sum of diagnal of confusion/sum of confusion.

Accuracy <- sum(diag(confusion)/sum(confusion))

**Accuracy = 0.765391**

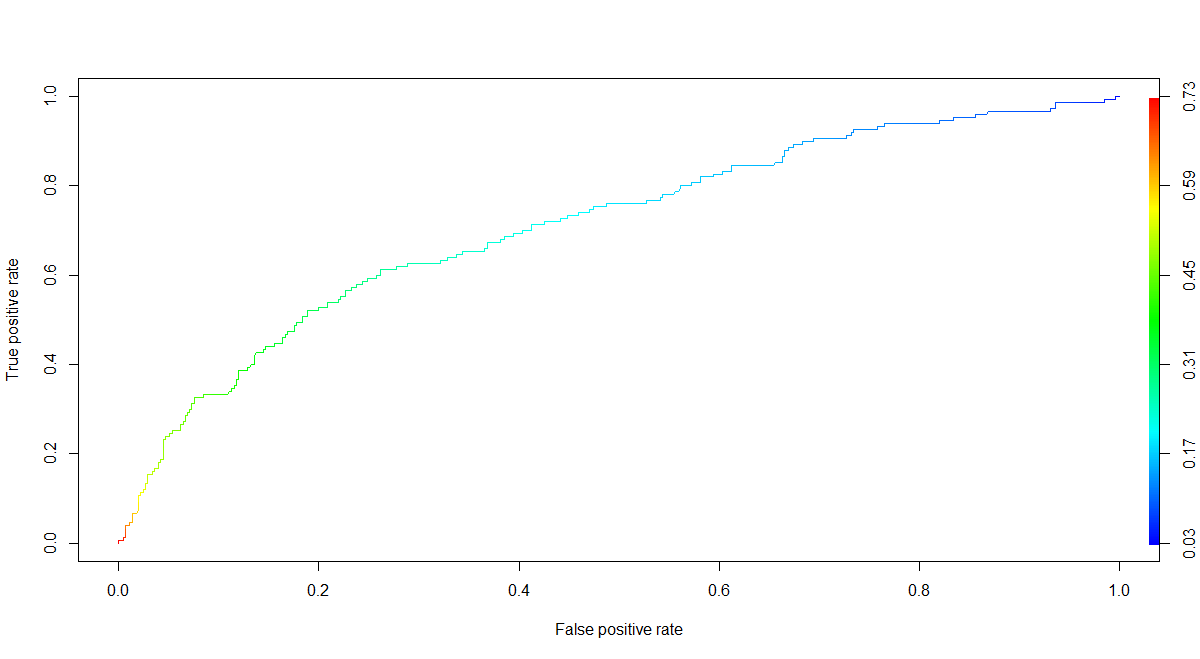
Specificity =435/(435+16)=**0.964** #TN/(TN+FP) OR True\_Negetive\_Rate

Sensitivity <-25/(25+125)=**0.16** #TP/(TP+FN) OR True\_Positive\_Rate

False\_Positive\_Rate <- 1-Specificity=**0.03**

By using accuracy of model which is 76.5 % .we can say that it is the best fit model to find out extra marital affairs.

**ROC (Receiver operating characteristic):**



ROCused to evaluate the betterness of the logistic model more area under ROC curve better is the model. We will use ROC curve for any classification technique not only for logistic.

rocrpred <- prediction(prob,Affairs$ynaffair)

rocrpred

rocrperf <- performance(rocrpred,'tpr','fpr')

rocrperf

plot(rocrperf,colorize=T,text.adj=c(-0.2,1.7))

**R code:**

install.packages("AER")

library(AER)

data("Affairs",package = "AER")

library(moments)

Affairs <- read.csv(file.choose())

View(Affairs)

summary(Affairs)

attach(Affairs)

skewness(Affairs$affairs)

skewness(Affairs$age)

plot(Affairs)

colnames(Affairs)

**#transform affairs column into a dichotomous factor called ynaffair with the following code.**

Affairs$ynaffair[Affairs$affairs > 0] <- 1

Affairs$ynaffair[Affairs$affairs == 0] <- 0

Affairs$ynaffair <- factor(Affairs$ynaffair,

levels=c(0,1),

labels=c("No","Yes"))

table(Affairs$ynaffair)

**# GLM function use sigmoid curve to produce desirable results**

**# The output of sigmoid function lies in between 0-1**

**#(GLM)generalised linear model**

model <- glm(ynaffair ~ gender+age+yearsmarried+children+religiousness+education+occupation+rating, data = Affairs, family = "binomial")

summary(model)

**# To calculate the odds ratio manually we going r going to take exp of coef(model)**

exp(coef(model))

**# Confusion matrix table**

prob <- predict(model,Affairs,type="response")

prob

**# We are going to use NULL and Residual Deviance to compare the between different models**

**# Confusion matrix and considering the threshold value as 0.5**

confusion <- table(prob > 0.5, Affairs$ynaffair)

confusion

**# Model Accuracy**

Accuracy <- sum(diag(confusion)/sum(confusion))

Accuracy

Specificity =435/(435+16)=0.964 #TN/(TN+FP) OR True\_Negetive\_Rate

Sensitivity <-25/(25+125)=0.16 #TP/(TP+FN) OR True\_Positive\_Rate

False\_Positive\_Rate <- 1-Specificity=0.03

**# Creating empty vectors to store predicted classes based on threshold value**

pred\_values <- NULL

yes\_no <- NULL

pred\_values <- ifelse(prob > 0.5, 1, 0)

yes\_no <- ifelse(prob > 0.5,"yes","no")

# Creating new column to store the above values

Affairs[ , "prob"] <- prob

Affairs[ , "pred\_values"] <- pred\_values

Affairs[ , "yes\_no"] <- yes\_no

View(Affairs[ ,c(1,11:12)])

table(Affairs$ynaffair,Affairs$pred\_values)

**# ROC(Receiver perativing characteristic) Curve => used to evaluate the betterness of the logistic model**

**# more area under ROC curve better is the model**

**# We will use ROC curve for any classification technique not only for logistic**

install.packages("ROCR")

library(ROCR)

rocrpred <- prediction(prob,Affairs$ynaffair)

rocrpred

rocrperf <- performance(rocrpred,'tpr','fpr')

rocrperf

plot(rocrperf,colorize=T,text.adj=c(-0.2,1.7))

**# More area under the ROC Curve better is the logistic regression model obtained**

str(rocrperf)

rocr\_cutoff <- data.frame(cut\_off = rocrperf@alpha.values[[1]],fpr=rocrperf@x.values,tpr=rocrperf@y.values)

colnames(rocr\_cutoff) <- c("cut\_off","FPR","TPR")

View(rocr\_cutoff)

library(dplyr)

rocr\_cutoff$cut\_off <- round(rocr\_cutoff$cut\_off,6)

**# Sorting data frame with respect to tpr in decreasing order**

rocr\_cutoff <- arrange(rocr\_cutoff,desc(TPR))

View(rocr\_cutoff)