I have a dataset containing family information of married couples, which have around 10 variables & 600+ observations.

Independent variables are ~ gender, age, years married, children, religion etc.

I have one response variable which is number of extra marital affairs.

Now, I want to know what all factor influences the chances of extra marital affair.

Since extra marital affair is a binary variable (either a person will have or not),

so, we can fit logistic regression model here to predict the probability of extra marital affair.

BUSSINESS PROBLEM:

To predict the probability of extra marital affair.

DATA CLEANSING AND MANIPULATION PROCESS:

As the given data consists of the index column remove the column.

Convert the data columns which is not in the form of binary format because extra marital affair is a binary variable.

MODEL CREATION:

Create a model using all the variables.

model <- glm(naffairs~.,data = affairs,family = 'binomial')

SUMMARY OF MODEL:

# Null deviance: 675.38 on 600 degrees of freedom

#Residual deviance: 602.21 on 586 degrees of freedom

#AIC: 632.21

The model consists of variables with high p values than the required significant values p values.

So, remove the variables with the high p values and build the model with the other variables.

Difference between the null and residual deviance is more it says that the model created is good.

Model 2:

model2 <- glm(naffairs~kids+vryunhap+unhap+avgmarr+hapavg+antirel+notrel+slghtrel+yrsmarr1+yrsmarr2+yrsmarr3+yrsmarr4,data=affairs,family = 'binomial')

summary:

#Null deviance: 675.38 on 600 degrees of freedom

#Residual deviance: 602.53 on 588 degrees of freedom

#AIC: 628.53

Compared to the previous model there is a slight change in the AIC value, change in the AIC values greater than two tells us that the model built is good.

Still there are variables with the highest p values, so, try removing the variables with the high p values and build the better model.

Model 3:

model3 <- glm(naffairs~vryunhap+unhap+avgmarr+hapavg+antirel+notrel+slghtrel+yrsmarr1+yrsmarr2+yrsmarr3+yrsmarr4,data=affairs,family = 'binomial')

summary:

# Null deviance: 675.38 on 600 degrees of freedom

#Residual deviance: 602.82 on 589 degrees of freedom

#AIC: 626.82

Slight change in the AIC values .

MODEL 4:

model4 <- glm(naffairs~vryunhap+unhap+avgmarr+hapavg+antirel+notrel+slghtrel+yrsmarr1+yrsmarr2+yrsmarr3,data=affairs,family = 'binomial')

summary:

# Null deviance: 675.38 on 600 degrees of freedom

#Residual deviance: 603.33 on 590 degrees of freedom

#AIC: 625.33

MODEL 5:

model5 <- glm(naffairs~vryunhap+unhap+avgmarr+hapavg+antirel+slghtrel+yrsmarr1+yrsmarr2,data=affairs,family = 'binomial')

summary:

#Null deviance: 675.38 on 600 degrees of freedom

#Residual deviance: 607.49 on 592 degrees of freedom

#AIC: 625.49

CONFUSION MATRIX:

Confusion Matrix and Statistics

Reference

Prediction 0 1

# 0 446 129

# 1 5 21

Accuracy: 0.777

95% CI: (0.7416, 0.8097)

No Information Rate: 0.7504

P-Value [Acc > NIR]: 0.07071

Kappa: 0.178

Mcnemar's Test P-Value: < 2e-16

Sensitivity: 0.14000

Specificity: 0.98891

Pos Pred Value: 0.80769

Neg Pred Value: 0.77565

Prevalence : 0.24958

Detection Rate : 0.03494

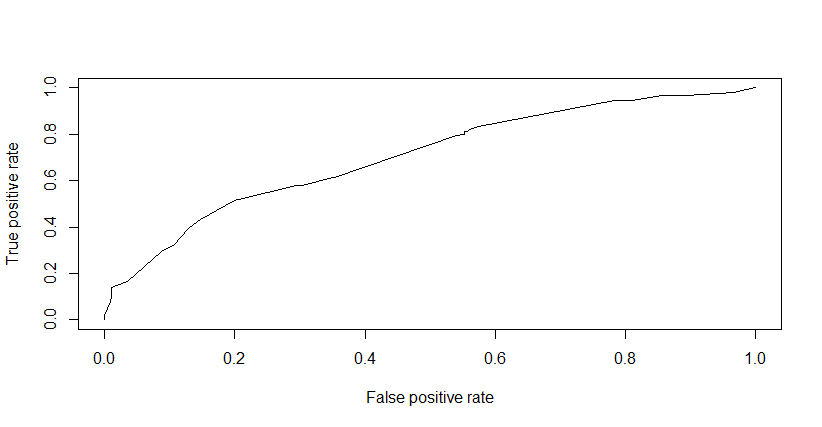
Detection Prevalence : 0.04326

Balanced Accuracy : 0.56446

'Positive' Class : 1

ROC CURVE:

**Cut-off Value using ROC curve**

****

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

Getting cut off or threshold value along with true positive and false positive rates in a data frame.

According to the roc curve we can say that the model is good based on the area under the curve.