**Business Problem:**

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 influence 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.

**Data:**

Data in the form of factors (Yes or No | 1 or 0).

**Pre-processing Data:**

All the data in the form of factors but extra marital affairs has in the form of numbers. So convert them into factors. There was no outlier and NA in the data. Delete the unused feature like ‘X’ from the processing.

**Model Building:**

The dataset contains the 18 features. Our goal is find out whether the person is having Extra Marital Affair or not. End result is discrete and Input data has only two outputs.So I’m proceeding with Binomial Regression. Initially I’, trying with all features (except vryhap ,vryrel yrsmarr6 – due to degree of freedom) and building the model.

model1 <- glm(naffairs~kids+vryunhap+unhap+avgmarr+hapavg+antirel+notrel+slghtrel+smerel+

yrsmarr1+yrsmarr2+yrsmarr3+yrsmarr4+yrsmarr5,data=affairs1,family = "binomial")

**Summary of the model1:**

Null deviance: 675.38 on 600 degrees of freedom

Residual deviance: 602.21 on 586 degrees of freedom

AIC: 632.21

There are few features have probability more than accepted value (0.05). I’m checking the correlation with the affairs.

**Checking the correlation:**

Find the correlation between affairs and kids, affairs and notrel.

model2 <- glm(naffairs~vryunhap+unhap+avgmarr+hapavg+antirel+slghtrel+smerel+

yrsmarr1+yrsmarr2+yrsmarr3+yrsmarr4+yrsmarr5,data=affairs1,family = "binomial")

**Summary of the model2:**

Null deviance: 675.38 on 600 degrees of freedom

Residual deviance: 603.82 on 588 degrees of freedom

AIC: 629.82

Null deviance is increase when we add the more and more features and it increases more and more. In the above case it is same.

Residual deviance is increase when the features have correlation with output variable then only it increases. In the above cases it increases slightly not much and slightly reduces the AIC.

Find the correlation between affairs and years3,4 and 5.

model3 <- glm(naffairs~vryunhap+unhap+avgmarr+hapavg+antirel+slghtrel+smerel+

yrsmarr1+yrsmarr2,data=affairs1,family = "binomial")

**Summary of the model3:**

Null deviance: 675.38 on 600 degrees of freedom

Residual deviance: 604.95 on 591 degrees of freedom

AIC: 624.95

Null deviance is increase when we add the more and more features and it increases more and more. In the above case it is same.

Residual deviance is increase when the features have correlation with output variable then only it increases. In the above cases it increases slightly not much and slightly reduces the AIC

Hence model3 is final model and we go for further calculation.

**Reference**

**Prediction**  **0 1**

**0** 429 118

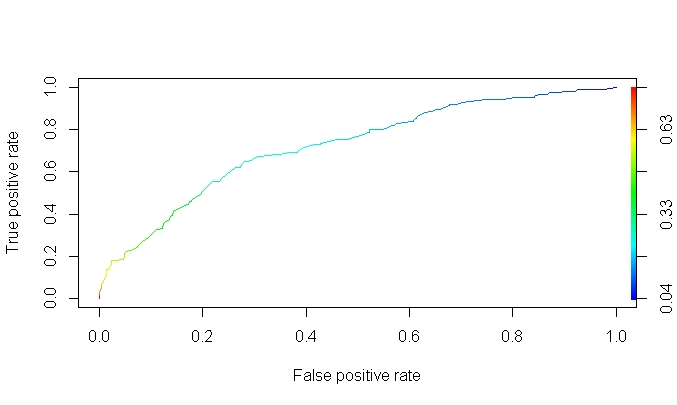
**1**  22 32

|  |  |
| --- | --- |
| True Positive | 32 |
| True Negative | 431 |
| False Positive | 20 |
| False negative | 118 |
| Accuracy | 77.03% |
| Precision | 61.5% |
| Sensitivity | 21.3% |
| Specificity | 95.5% |
| True Positive Rate | 78.6% |
| False Positive Rate | 0.04% |

**Conclusion:**

Initial 2 years and marriage life and happiness of couple are playing major role for attempting extra marital affairs. If they are good and happy in the first 2 years then there will be less chance for extra marital affairs.

**Cut-off Value using ROC curve:**



As per the ROC Curve , more under the true positive rate and less false positive rate is good. So am per my understanding 50% is good.