**Business Problem:**

Whether the client has subscribed a term deposit or not

**Data:**

Data is in numerical format, it is combination of continuous and discrete data. Most of features are in the form of factor.

**Pre-processing Data:**

All the data in the form factors and numeric. There was no outlier and NA in the data.

**Model Building:**

The dataset contains the 31 features. Our goal is find out whether the client has subscribed term deposit 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 and building the model.

Model1 <- glm(y~.,data=claimants,family = "binomial")

**Summary of the model1:**

Null deviance: 32631 on 45210 degrees of freedom

Residual deviance: 22640 on 45183 degrees of freedom

AIC: 22696

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

**Checking the correlation:**

Find the correlation between pdays , joblue.collar , joentrepreneu , defaultjohousemaid , joself.employed , joservices with y.

Model2 <- glm( y~ balance + housing + loan + duration + campaign + previous + poutfailure + poutother + poutsuccess + con\_cellular + con\_telephone + divorced + married + joadmin. + jomanagement + joretired + jostudent + jotechnician+ jounemployed,family = "binomial")

**Summary of the model2:**

Null deviance: 32631 on 45210 degrees of freedom

Residual deviance: 22648 on 45191 degrees of freedom

AIC: 22688

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 when we delete the features which don’t have correlation.

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 model2 is final model and we go for further calculation.

**Reference**

**Prediction**  **0 1**

**0** 39485 4339

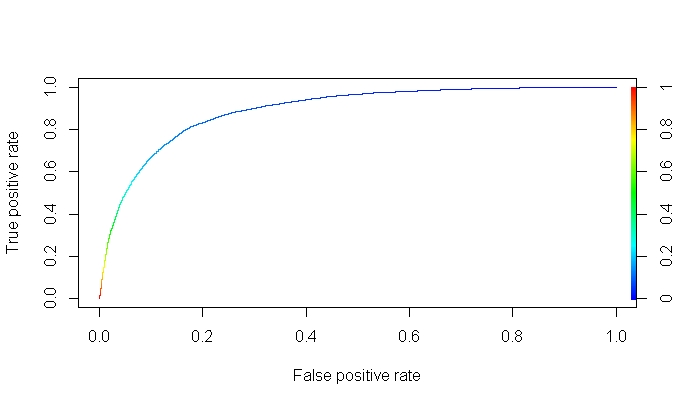
**1**  437 950

|  |  |
| --- | --- |
| True Positive | 950 |
| True Negative | 39845 |
| False Positive | 437 |
| False negative | 4339 |
| Accuracy | 89.4% |
| Precision | 68.4% |
| Sensitivity | 17.9% |
| Specificity | 98.9% |
| True Positive Rate | 82.% |
| False Positive Rate | 0.01% |

**Conclusion:**

I’m not able conclude anything with this data, because I have got false negative cases (4339) more than true positive (950) is one thing. In my binomial regression almost i got accepted probability values for which i have used in the regression. So with this data I am not able to conclude properly. It gives the good accuracy and good True positive rate. With this data we can estimate who are not subscribed the term deposit but can’t predict the client who subscribe term deposit.

**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 70% is good.