**Name:** Rajib Mandal(SID: TA17002)

**Data Set Information**

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The goal is to build models which predict if the client will subscribe a term deposit(y)

**Data Cleaning and preparation**

The following tasks are performed during this phase

1. Check if there is missing value in any of the given fields in the data set

bankdata<-read.csv("bankdata1.csv",header=TRUE)

naColumns <- function(df)

{

colnames(df)[unlist(lapply(df, function(x) any(is.na(x))))]

}

naColumns(bankdata)

1. Calculate mode in the job field and replace unknown value with mode

job\_mode<-names(table(bankdata$job))[table(bankdata$job)==max(table(bankdata$job))]

bankdata$job[bankdata$job == "unknown"] <- job\_mode

1. Remove month data from the data set
2. Transform all the categorical fields with numerical value 0 &1
3. Create dummy variable for job

bankdata$job\_status<-ifelse(bankdata$job!='unemployed',1,0)

bankdata$job<-NULL

1. Create dummy variable marital

bankdata$sing\_status<-ifelse(bankdata$marital=="single",1,0)

bankdata$mar\_status<-ifelse(bankdata$marital=="married",1,0)

bankdata$marital<-NULL;

1. create dummy variable for education

bankdata$education\_primary<-ifelse(bankdata$education=='primary',1,0)

bankdata$education\_secondary<-ifelse(bankdata$education=='secondary',1,0)

bankdata$education\_tertiary<-ifelse(bankdata$education=='tertiary',1,0)

bankdata$education<-NULL

1. create dummy variables for default

bankdata$default\_yes<-ifelse(bankdata$default=='yes',1,0)

bankdata$default<-NULL

1. create dummy variable for housing

bankdata$housing\_loan<-ifelse(bankdata$housing=='yes',1,0)

bankdata$housing<-NULL

1. create dummy variable for loan

bankdata$loan\_approval<-ifelse(bankdata$loan=='yes',1,0)

bankdata$loan<-NULL

1. create dummy variable for contact

bankdata$contact\_cellular<-ifelse(bankdata$contact=='cellular',1,0)

bankdata$contact\_telephone<-ifelse(bankdata$contact=='telephone',1,0)

bankdata$contact<-NULL

1. create dummy variable for outcome

bankdata$p\_outcome\_fail<-ifelse(bankdata$poutcome=='failure',1,0)

bankdata$p\_outcome\_success<-ifelse(bankdata$poutcome=='success',1,0)

bankdata$p\_outcome\_other<-ifelse(bankdata$poutcome=='other',1,0)

1. create dummy variable for y

bankdata$churn\_y\_deposit<-ifelse(bankdata$y=='yes',1,0)

bankdata$y<-NULL

1. Check for multicollinearity

res <-glm(churn\_y\_deposit~age+balance+day+duration+campaign+pdays+

p\_outcome\_fail+ p\_outcome\_success+p\_outcome\_other+mar\_status+sing\_status+ education\_primary+education\_tertiary+education\_secondary+default\_yes+loan\_approval+housing\_loan+contact\_cellular+contact\_telephone+job\_status

,data=bankdata,family = "binomial")

vif(res)

Since the values all the variables are less than 10, we can conclude there is no relation between explanatory variables.

|  |  |
| --- | --- |
| Variable Name | Values |
| age | 1.519295 |
| balance | 1.034195 |
| day | 1.019289 |
| duration | 1.173969 |
| campaign | 1.043238 |
| pdays | 3.37609 |
| p\_outcome\_fail | 2.800721 |
| p\_outcome\_success | 1.304001 |
| p\_outcome\_other | 1.750624 |
| mar\_status | 2.577838 |
| sing\_status | 3.069774 |
| education\_primary | 3.276973 |
| education\_tertiary | 5.674969 |
| education\_secondary | 6.051041 |
| default\_yes | 1.013234 |
| loan\_approval | 1.034935 |
| housing\_loan | 1.188188 |
| contact\_cellular | 1.629395 |
| contact\_telephone | 1.563408 |
| job\_status | 1.008203 |

**Model Building**

1. Creation of testing and training dataset from the given data set with 70:30 ratio

spl<-sample.split(bankdata,SplitRatio = 0.7)

bankdata\_training<-subset(bankdata,spl==TRUE)

bankdata\_testing<-subset(bankdata,spl==FALSE)

1. Building best fitted model on training data

**Model:1**

Consider all explanatory variables in the model

churn\_logistic<-glm(churn\_y\_deposit~age+balance+day+duration+campaign+pdays+

p\_outcome\_fail+ p\_outcome\_success+p\_outcome\_other+mar\_status+sing\_status+

education\_primary+education\_tertiary+education\_secondary+ default\_yes+loan\_approval+housing\_loan+contact\_cellular+contact\_telephone+job\_status,data = bankdata\_training,family = "binomial")

summary(churn\_logistic)

|  |  |  |
| --- | --- | --- |
| Summary | | Degree of Freedom |
| Null deviance | 10651.6 | 7705 |
| Residual deviance | 6467 | 7685 |
| AIC | 6509 |  |

**Model:2**

Remove day field from the model

churn\_logistic<-glm(churn\_y\_deposit~age+balance+duration+campaign+pdays+

p\_outcome\_fail+ p\_outcome\_success+p\_outcome\_other+mar\_status+sing\_status+

education\_primary+education\_tertiary+education\_secondary+

default\_yes+housing\_loan+contact\_cellular+contact\_telephone+job\_status,data = bankdata\_training,family = "binomial")

summary(churn\_logistic)

|  |  |  |
| --- | --- | --- |
| Summary | | Degree of Freedom |
| Null deviance | 10651.6 | 7705 |
| Residual deviance | 6472.4 | 7686 |
| AIC | 6512.4 |  |

**Model:3**

Remove day field & loan approval from the model

churn\_logistic<-glm(churn\_y\_deposit~age+balance+duration+campaign+pdays+

p\_outcome\_fail+ p\_outcome\_success+p\_outcome\_other+mar\_status+sing\_status+

education\_primary+education\_tertiary+education\_secondary+

default\_yes+housing\_loan+contact\_cellular+contact\_telephone+job\_status,data = bankdata\_training,family = "binomial")

summary(churn\_logistic)

|  |  |  |
| --- | --- | --- |
| Summary | | Degree of Freedom |
| Null deviance | 10651.6 | 7705 |
| Residual deviance | 6525.6 | 7687 |
| AIC | 6563.6 |  |

Since AIC for Model1 is less than Model2 & Model3 , we choose Model1 as a best fitted model

1. Apply Logistic regression on test data

prediction\_data<-predict(churn\_logistic,bankdata\_testing,type="response")

predicted\_outcome<-ifelse(prediction\_data>0.5,1,0)

churn<-data.frame(bankdata\_testing,prediction\_data,predicted\_outcome)

write.csv(churn,"TEST\_Assignnment\_TA17002.csv")

tb<-table(predicted\_outcome,bankdata\_testing$churn\_y\_deposit)

acc<-sum(diag(tb))/sum(tb)

acc

miss\_class<-1-acc

Here we find the accuracy of the logistic model is 80.63%

1. The performance of the model is 88.72%

roc\_pred<-predict(churn\_logistic,bankdata\_testing,type="response")

roc\_obj<-prediction(roc\_pred,bankdata\_testing$churn\_y\_deposit)

roc\_performance<-performance(roc\_obj,"tpr","fpr")

plot(roc\_performance)

abline(a=0,b=1)

accuracy<-performance(roc\_obj,"auc")

accuracy

1. Evaluating TPR, FPR

For the threshold value **0.5** ,we get **TPR= 0.7426 FPR= 0.1505**

Thus, we can conclude 74% customer who will churn properly classified with Churn label, and 15% of the customers who will remain as customers are improperly labelled as ***Churn***

Here is the sample code

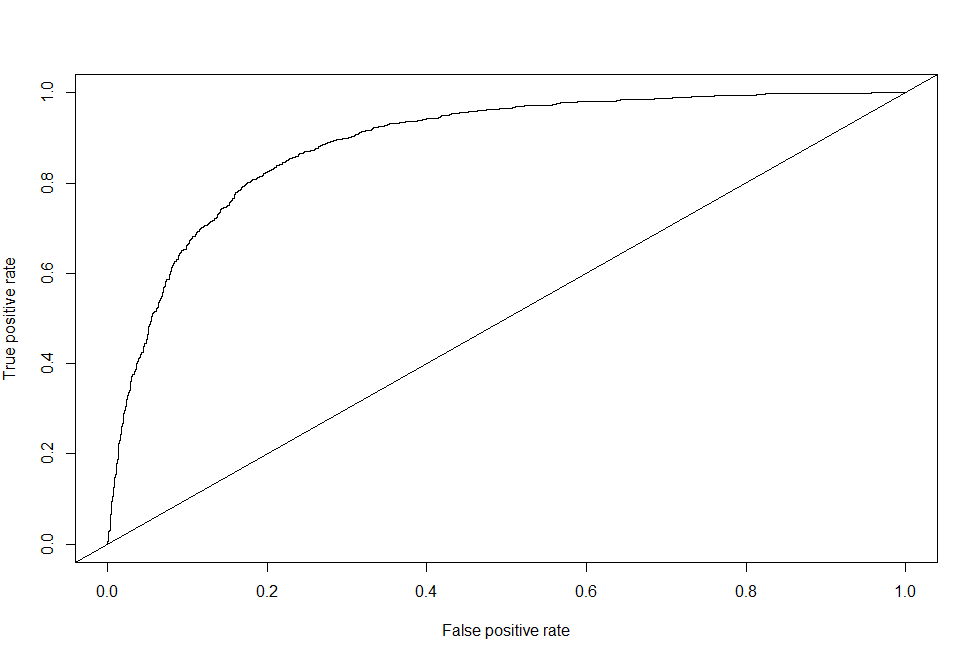
tpr<-round(as.numeric(unlist(roc\_performance@y.values)),4)

fpr<-round(as.numeric(unlist(roc\_performance@x.values)),4)

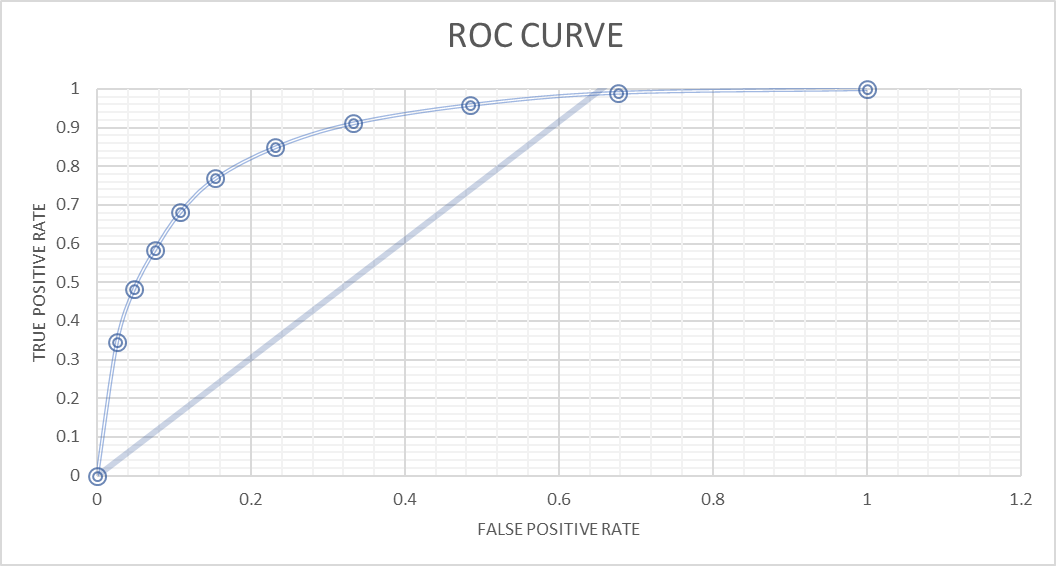
alpha<-round(as.numeric(unlist(roc\_performance@alpha.values)),4)

i<-which(round(alpha,2)==0.5)

paste("Threshold=",(alpha[i]),"TPR=",tpr[i],"FPR=",fpr[i])



ROC Curve generated from Excel(Refer : TEST\_Assignnment\_TA17002.xlsx)



**Appendix**

Refer attached two files for detail analysis

1. Dataminig\_Assignnment\_TA17002.R
2. TEST\_Assignnment\_TA17002.XLSX