SUMMER PROJECT

VINEET

15/06/2021

**ROLL :MQMS2014**

**MODELLING USING PERSONAL LOAN DATA #importing the dataset**

**PROBLEM STATEMENT:**

**WE SHALL FIRST TAKE A LOOK AT PERSONAL LOAN DATASET.**

**HERE WE NEED TO PREDICT THAT FOR A GIVEN INDIVIDUAL WHETHER HE WILL RECEIVE THE LOAN OR NOT DEPENDING ON THE GIVEN VARIABLES.**

library(readxl)  
mydata <- read\_excel("C:/Users/Bineet/Downloads/Personal Loan Data 1set.xlsx")

# APPLYING NAIVE BAYES

library(e1071)  
set.seed(1)  
sampleid=sample(2,999,replace = TRUE,prob=c(0.8,0.2))  
training <- mydata[sampleid==1,]  
test <- mydata[sampleid==2,]

#SEPAQRATE X AND Y

x=training[,-8]  
y=factor(training$`Personal Loan`)

#BUILDING THE MODEL

mymodel <- naiveBayes(x,y)  
  
pred <- predict(mymodel,x)

#comparing Actual vs Fitted values

mytable <- table(y,pred)  
mytable

## pred  
## y 0 1  
## 0 660 60  
## 1 29 52

round(prop.table(mytable)\*100,2)

## pred  
## y 0 1  
## 0 82.40 7.49  
## 1 3.62 6.49

#FOR TRAINING DATA #ACCURACY=88.89% #MISCLASSIFICATION=11.11%

#SEEING THE ACCURACY OF MODEL ON TEST DATA

testx <- test[,-8]  
testy <- factor(test$`Personal Loan`)  
  
predtest <- predict(mymodel,testx)  
  
mytesttable <- table(testy,predtest)  
mytesttable

## predtest  
## testy 0 1  
## 0 165 18  
## 1 6 9

round(prop.table(mytesttable)\*100,2)

## predtest  
## testy 0 1  
## 0 83.33 9.09  
## 1 3.03 4.55

#FOR test DATA #ACCURACY=87.88% #MISCLASSIFICATION=12.12%

**LOGISTIC REGRESSION MODEL Developing the model**

attach(mydata)

## The following objects are masked from training:  
##   
## Age, CCAvg, CD Account, CreditCard, Education, Experience, Family,  
## Income, Mortgage, Online, Personal Loan, Securities Account

result<- factor(mydata$`Personal Loan`)  
mymodel <- glm(`Personal Loan`~.,data=mydata,family = binomial(logit))  
anova(mymodel,test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Personal Loan  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 998 632.21   
## Age 1 3.085 997 629.12 0.0790333 .   
## Experience 1 2.606 996 626.52 0.1064815   
## Income 1 245.398 995 381.12 < 2.2e-16 \*\*\*  
## Family 1 44.069 994 337.05 3.170e-11 \*\*\*  
## CCAvg 1 0.527 993 336.52 0.4676633   
## Education 1 64.931 992 271.59 7.758e-16 \*\*\*  
## Mortgage 1 0.003 991 271.59 0.9578955   
## `Securities Account` 1 0.055 990 271.53 0.8141224   
## `CD Account` 1 15.858 989 255.67 6.827e-05 \*\*\*  
## Online 1 1.100 988 254.57 0.2942182   
## CreditCard 1 14.232 987 240.34 0.0001616 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Dropping the variables age,experience,ccavg,mortgage,securities account,online as they are insignificant in our model for prediction

mymodel <- glm(`Personal Loan`~ Income+Family+Education+`CD Account`+CreditCard,family = binomial(logit)  
)  
nullmodel <- glm(`Personal Loan`~1,family = binomial(logit))  
anova(nullmodel,mymodel,test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: `Personal Loan` ~ 1  
## Model 2: `Personal Loan` ~ Income + Family + Education + `CD Account` +   
## CreditCard  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 998 632.21   
## 2 993 258.50 5 373.71 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p\_value of our MOdel is less than 0.05..Hence it is SIGNIFICANT

pred <- predict(mymodel,type="response")  
predclass <- ifelse(pred>0.5,"1","0")  
myresult <- cbind(mydata,pred,predclass)  
mytable <- table(result,predclass)  
mytable

## predclass  
## result 0 1  
## 0 890 13  
## 1 36 60

round(prop.table(mytable)\*100,2)

## predclass  
## result 0 1  
## 0 89.09 1.30  
## 1 3.60 6.01

ACCURACY=95.10% MISCLASSIFICATION=4.9% MODEL IS ACCURATE AS ACCURACY >80%

LOOCV

library(boot)  
mymodel <- glm(factor(`Personal Loan`)~ Income+Family+factor(Education)+factor(`CD Account`)+factor(CreditCard),family = binomial(logit)  
)  
mycv<- cv.glm(mydata,mymodel)  
loocv\_misclassification\_error <- mycv$delta[1]  
loocv\_misclassification\_error

## [1] 0.03120035

K-FOLD CV

set.seed(1)  
mymodel <- glm(factor(`Personal Loan`)~ Income+Family+  
factor(Education)+factor(`CD Account`)+factor(CreditCard),family = binomial(logit))  
mycv<- cv.glm(mydata,mymodel,K=10)  
kcv\_misclassification\_error <- mycv$delta[1]  
kcv\_misclassification\_error

## [1] 0.03056426

CLEARLY THE MODEL PERFORMANCE AS IMPROVED AS THE MISCLASSIFICATION FROM MODEL IS 4.9% FROM LOOCV IT IS 0.031% FROM K-FOLD CV IT IS 0.030% MODEL IS GENERALIZABLE

**K NEARESR NEIGHBOUR MODELS**

library(FNN)  
set.seed(2)  
x=training[,-8]  
y=training$`Personal Loan`  
mymodel <- ownn(x,x,y)  
mymodel$k

## [1] 67

pred <- mymodel$knnpred  
mytable <- table(y,pred)  
mytable

## pred  
## y 0  
## 0 720  
## 1 81

round(prop.table(mytable)\*100,2)

## pred  
## y 0  
## 0 89.89  
## 1 10.11

#TEST DATA ANALYSIS  
testx<- test[,-8]  
testy <- test$`Personal Loan`  
mymodel <- ownn(x,testx,y,prob=T)  
predtest <- mymodel$knnpred  
mytesttable <- table(testy,predtest)  
mytesttable

## predtest  
## testy 0  
## 0 183  
## 1 15

round(prop.table(mytesttable)\*100,2)

## predtest  
## testy 0  
## 0 92.42  
## 1 7.58

FOR TRAINING DATA ACCURACY =89.89% HENCE MODEL IS ACCURATE MISCLASSIFIVATION IS 10.11 %  
FOR TEST data ACCURACY=92.42% MISCLASSIFICATION=7.58%..MODEL IS GENERALIZABLE AS THERE IS SLIGHT DIFFERENCE BETWEEN TRAINING AND TEST DATA MISCLASSIFICATION

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | |
|  | | | NAÏVE BAYES | | | KNN | | LOGISTIC REGRESSION | | |
|  | | | TRAINING | | TEST | TRAINING | TEST | ON MODEL | LOOCV | KFOLD CV |
| ACCURACY | | | 88.89 | | 87.88 | 89.89 | 92.42 | 95.1 |  |  |
| MISCLASSIFICATION | | | 11.11 | | 12.12 | 10.11 | 7.58 | 4.9 | 0.031 | 0.03 |

FOR LOGISTIC REGRESSION MODEL TEST DATA ACCURACY SING LOOCV IS 0.03

**CLEARLY ALL THE MODELS ARE ACCURATE AND GENERALIZABLE.**

**COMPARING THE EFFICIENCY OF THE MODEL ON THE BASIS OF TEST DATA we see that LOGISTIC REGRESSION MODEL PROVIDES THE BEST MODEL**