## Predicting Bank Loan Defaults:Logistic Regression in R

#Read the data from the working directory, create your own working directly to read the dataset.

data1 <- read.csv ("C:/Users/Deep/Desktop/data/

loan\_default.csv",header=TRUE,sep=",")

data2<-data.frame(data1)

#perform exploratory data analysis to know about the data

# display top 6 rows of dataset to see how data look like

head (data2)

# display bottom 6 rows

tail(data2)

# describe the structure of data

str(data2)

#display the column name of the data

names(data2)

# display the datatype

class(data2)

#display the summary or descriptive statistics of the data

summary(data2$Amount)

#Let’s check the missing values present in the data

is.na(data2)

#To find out the correlation between the variables

corr <- cor.test(data2$Default, data2$Term,

method = "pearson" )

corr

#building logistic regression model using glm on full data

fullmodel1 <-glm(Default~.,data = data2,family=binomial

(link=logit ))

summary(fullmodel1)

#removing insignificant variables in order to build final logistic model on full data

fullmodel2<-glm(Default~Checking\_amount+Term+Credit\_score

+Saving\_amount+Age,data = data2,family=binomial(link=logit))

summary(fullmodel2)

#splitting data set into training and validation dataset

in 70:30

train\_obs <- floor (0.7\*nrow (data2))

print(train\_obs)

#Set seed in order to reproduce the sample

set.seed(2)

train\_ind <- sample(seq\_len(nrow(data2)),size=train\_obs)

test <- -train\_ind

#No of observations in train dataset

train\_data<-data2[train\_ind,]

# No of observations in test dataset

test\_data<-data2[-train\_ind,]

testing\_high = data2$Default[test]

#Building logistic regression model using glm on training data

model1<-glm(Default~.,data= train\_data,family=binomial(link=logit))

summary(model1)

#After removing insignificant variable inorder to build final logistic model on training data

model2<-glm(Default~Checking\_amount+Term+Credit\_score

+Emp\_status+Saving\_amount+Age,data= train\_data,family=binomial(link=logit))

summary(model2)

#Check for variance inflation factor, VIF > 5 to 10 -high correlation

#install car package

install.packages("car")

library(car)

vif(model2)

# Predicting the model using test data

Prob <-predict(model2,test\_data,type ="response")

prob1<- data.frame(prob)

# setting the cutoff for probability values

results <- ifelse(prob1 > 0.7,1,0)

#Display the confusion matrix or classification table

table(testing\_high,results)

#Calculating the error rate

misclasificationerror <- mean(results != testing\_high)

misclasificationerror

# Calculating the accuracy rate

accuracyrate <- 1-misclasificationerror

print(accuracyrate)

#MODEL FIT TEST

#Hosmer-Lemeshow Test

#Install MKmisc package

install.packages("MKmisc ")

library(MKmisc)

HLgof.test(fit = fitted(model2), obs = train\_data$Default)

#Install ResourceSelection package

install.packages("ResourceSelection ")

library(ResourceSelection)

hoslem.test(train\_data$Default, fitted(model2), g=10)

#Likelihood Ratio Test

model3<-glm(Default~Checking\_amount+Term+Amount,

data= train\_data,family=binomial(link=logit))

summary(model3)

#Install lmtest package

install.packages("lmtest")

library(lmtest)

lrtest(model1, model3)

# Statistical Tests for Individual Predictors: Wald Test#

#Install survey package

install.packages("survey")

library(survey)

# Wald test for Credit\_Score

regTermTest(model2,"Credit\_score")

# Wald test for Age

regTermTest(model2, "Age")

# conducting Receiver operating characteristic (ROC) test and Area under curve (AUC)

install.packages("ROCR")

library(ROCR)

# Compute AUC for predicting Default with the model

prob <- predict(model2, newdata=test\_data, type="response")

pred <- prediction(prob, test\_data$Default)

pmf <- performance(pred, measure = "tpr", x.measure = "fpr")

plot(pmf,col= "red" )

auc <- performance(pred, measure = "auc")

auc <- auc@y.values[[1]]

auc