#####Installing libraries

install.packages(c("dplyr","car","hflights","lubridate","tidyr","xlsx","stringr",

"esquisse","vcd","ggplot2","nortest","sas7bdat","psych",

"stringi","tree","cvTools","randomForest","knitr","xtable",

"gbm","forecast","caret","ranger","data.table"))

g=c("dplyr","car","hflights","lubridate","tidyr","xlsx","stringr",

"esquisse","vcd","ggplot2","nortest","sas7bdat","psych",

"stringi","tree","cvTools","randomForest","knitr","xtable",

"gbm","forecast","caret","ranger","data.table")

lapply(g, library, character.only = TRUE)

#####

setwd("C:\\Users\\ADMIN\\Desktop\\R\_project1")

bank\_train=read.csv("C:\\Users\\ADMIN\\Desktop\\R\_project1\\bank-full\_train.csv",stringsAsFactors = F)

bank\_test=read.csv("C:\\Users\\ADMIN\\Desktop\\R\_project1\\bank-full\_test.csv",stringsAsFactors = F)

mean(bank\_train$age)

var(bank\_train$balance)

setdiff(colnames(bank\_train),colnames(bank\_test))

#y

bank\_test$y=NA

bank\_train$data="train"

bank\_test$data="test"

bank=rbind(bank\_train,bank\_test)

glimpse(bank)

bank=bank %>%

select(-ID)

table(bank$age)

quantile(bank$age)

bank$age=ifelse(bank$age <= 33, 1,

ifelse(bank$age <= 39, 2,

ifelse(bank$age <= 48, 3, 4)))

CreateDummies=function(data,var,freq\_cutoff=0){

t=table(data[,var])

t=t[t>freq\_cutoff]

t=sort(t)

categories=names(t)[-1]

for( cat in categories){

name=paste(var,cat,sep="\_")

name=gsub(" ","",name)

name=gsub("-","\_",name)

name=gsub("\\?","Q",name)

name=gsub("<","LT\_",name)

name=gsub("\\+","",name)

data[,name]=as.numeric(data[,var]==cat)

}

data[,var]=NULL

return(data)

}

bank=CreateDummies(bank,"job")

bank=CreateDummies(bank,"marital")

bank=CreateDummies(bank,"education")

bank=CreateDummies(bank,"contact")

bank=CreateDummies(bank,"month")

bank=CreateDummies(bank,"poutcome")

table(bank$y)

bank$y=(bank$y=="yes") + 0

table(bank$default)

bank$default=(bank$default=="yes") +0

table(bank$loan)

bank$loan=(bank$loan=="yes") +0

table(bank$housing)

bank$housing=(bank$housing=="yes") +0

glimpse(bank)

z=sapply(bank,function(x) is.character(x))

z=z[z==T]

z

glimpse(bank)

#####CHECKING FOR NA VALUES.

lapply(bank,function(x) sum(is.na(x)))

#####FILTERING DATA

bank\_train=bank %>%

filter(data=='train') %>%

select(-data)

bank\_test=bank%>%

filter(data=='test') %>%

select(-data,-y)

#####

set.seed(2211)

s=sample(1:nrow(bank\_train),0.8\*nrow(bank\_train))

bank\_train1=bank\_train[s,]

bank\_train2=bank\_train[-s,]

fit=lm(y~.,data=bank\_train1)

vif(fit)

sort(vif(fit),decreasing = T)

fit=lm(y~.-month\_may-job\_blue\_collar,data=bank\_train1)

vif(fit)

sort(vif(fit),decreasing = T)

#####

log\_fit=glm(y~.-month\_may-job\_blue\_collar,data=bank\_train1,family="binomial")

log\_fit=step(log\_fit)

formula(log\_fit)

log\_fit=glm(y ~ balance + housing + loan + day + duration + campaign + job\_student +

job\_housemaid + job\_retired + job\_admin. + job\_technician +

job\_management + marital\_married + education\_primary +

education\_tertiary + contact\_unknown + month\_mar + month\_sep +

month\_oct + month\_jan + month\_feb + month\_apr + month\_nov +

month\_jun + month\_aug + month\_jul + poutcome\_other + poutcome\_failure +

poutcome\_unknown,data=bank\_train1,family="binomial")

summary(log\_fit)

#####

library(pROC)

val.score=predict(log\_fit,newdata = bank\_train2,type='response')

auc(roc(bank\_train2$y,val.score))

#Area under the curve: 0.9027

#####

log\_fit\_final=glm(y ~ balance + housing + loan + day + duration + campaign + job\_student +

job\_housemaid + job\_retired + job\_admin. + job\_technician +

job\_management + marital\_married + education\_primary +

education\_tertiary + contact\_unknown + month\_mar + month\_sep +

month\_oct + month\_jan + month\_feb + month\_apr + month\_nov +

month\_jun + month\_aug + month\_jul + poutcome\_other + poutcome\_failure +

poutcome\_unknown,data=bank\_train,family = "binomial")

formula(log\_fit\_final)

summary(log\_fit\_final)

variable=predict(log\_fit\_final,bank\_train,type="response")

test.prob.score= as.numeric(predict(log\_fit\_final,newdata = bank\_test,type='response')>0.3)

test.prob.score=ifelse(test.prob.score==1,'Yes','No')

write.table(test.prob.score,"Ganesh\_patil\_P5\_part2.csv",row.names = F,col.names="y")

#####

library(ROCR)

RP=prediction(variable,bank\_train$y)

RPE=performance(RP,"tpr","fpr")

plot(RPE,colorize=T,print.cutoffs.at=seq(0.1,by=0.1))

# 0.3 THRESHOLD OBTAINED FROM GRAPH

#CONFUSION MATRIX

table(ActualValue=bank\_train2$y,PredictedValue=val.score>0.3)

ks=(5286/(5286+299))-(355/(355+390))

round(ks,2)

#0.47

library(ggplot2)

bank\_train$score=predict(log\_fit\_final,bank\_train,type="response")

ggplot(bank\_train,aes(x=score,y=y,color=factor(y)))+geom\_point()+geom\_jitter()

k=read.csv("Ganesh\_Patil\_P5\_part2.csv")

table(k$y)