```
#clear the envorinment
rm(list=ls())

#set working directory
setwd("E:/data science and machine learning/santander/R file")

#current working directory
getwd()


#loading file
sdf=read.csv("train.csv")
#checking the data type
str(sdf)

#loading some of the libraries
#x=c("ggplot2","Corrgram","DMwR","Caret","randomForest","unbalanced","C50","dummies","e10","MASS","rpart","gbm","ROSE")

#lapply(x,require,character.only=TRUE)

library("VIM")


#finding out number of missing value
missing_val=data.frame(apply(sdf,2,function(x){sum(is.na(x))}))

#giving names in dataframe
missing_val$Columns=row.names(missing_val)
row.names(missing_val)=NULL
names(missing_val)[1]="Missing_Percentage"

#converting to percentage
missing_val$Missing_Percentage=(missing_val$Missing_Percentage/nrow(sdf))*100

#Arranging in descending order
missing_val=missing_val[order(-missing_val$Missing_Percentage),]

#BOXPLOT FOR OUTLIER ANALYSIS

p1=sdf$var_1
p2=sdf$var_2
p3=sdf$var_3
p4=sdf$var_4
p5=sdf$var_5
p6=sdf$var_6
p7=sdf$var_7
p8=sdf$var_8
p9=sdf$var_9
p10=sdf$var_10

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_1","var_2","var_3","var_4","var_5","var_6","var_7","var_8","var_9","var_10"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)

p1=sdf$var_11
p2=sdf$var_12
p3=sdf$var_13
p4=sdf$var_14
p5=sdf$var_15
p6=sdf$var_16
p7=sdf$var_17
p8=sdf$var_18
p9=sdf$var_19
p10=sdf$var_20

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_11","var_12","var_13","var_14","var_15","var_16","var_17","var_18","var_19","var_20"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)

p1=sdf$var_21
p2=sdf$var_22
p3=sdf$var_23
p4=sdf$var_24
```

```
p5=sdf$var_25
p6=sdf$var_26
p7=sdf$var_27
p8=sdf$var_28
p9=sdf$var_29
p10=sdf$var_30

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_21","var_22","var_23","var_24","var_25","var_26","var_27","var_28","var_29","var_30"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)

p1=sdf$var_31
p2=sdf$var_32
p3=sdf$var_33
p4=sdf$var_34
p5=sdf$var_35
p6=sdf$var_36
p7=sdf$var_37
p8=sdf$var_38
p9=sdf$var_39
p10=sdf$var_40

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_31","var_32","var_33","var_34","var_35","var_36","var_37","var_38","var_39","var_40"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_41
p2=sdf$var_42
p3=sdf$var_43
p4=sdf$var_44
p5=sdf$var_45
p6=sdf$var_46
p7=sdf$var_47
p8=sdf$var_48
p9=sdf$var_49
p10=sdf$var_50

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_41","var_42","var_43","var_44","var_45","var_46","var_47","var_48","var_49","var_50"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_51
p2=sdf$var_52
p3=sdf$var_53
p4=sdf$var_54
p5=sdf$var_55
p6=sdf$var_56
p7=sdf$var_57
p8=sdf$var_58
p9=sdf$var_59
p10=sdf$var_60

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_51","var_52","var_53","var_54","var_55","var_56","var_57","var_58","var_59","var_60"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_61
p2=sdf$var_62
p3=sdf$var_63
p4=sdf$var_64
p5=sdf$var_65
p6=sdf$var_66
p7=sdf$var_67
p8=sdf$var_68
```

```
p9=sdf$var_69
p10=sdf$var_70

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_61","var_62","var_63","var_64","var_65","var_66","var_67","var_68","var_69","var_70"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_71
p2=sdf$var_72
p3=sdf$var_73
p4=sdf$var_74
p5=sdf$var_75
p6=sdf$var_76
p7=sdf$var_77
p8=sdf$var_78
p9=sdf$var_79
p10=sdf$var_80

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_71","var_72","var_73","var_74","var_75","var_76","var_77","var_78","var_79","var_80"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_71
p2=sdf$var_72
p3=sdf$var_73
p4=sdf$var_74
p5=sdf$var_75
p6=sdf$var_76
p7=sdf$var_77
p8=sdf$var_78
p9=sdf$var_79
p10=sdf$var_80

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_71","var_72","var_73","var_74","var_75","var_76","var_77","var_78","var_79","var_80"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_81
p2=sdf$var_82
p3=sdf$var_83
p4=sdf$var_84
p5=sdf$var_85
p6=sdf$var_86
p7=sdf$var_87
p8=sdf$var_88
p9=sdf$var_89
p10=sdf$var_90

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_81","var_82","var_83","var_84","var_85","var_86","var_87","var_88","var_89","var_90"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_91
p2=sdf$var_92
p3=sdf$var_93
p4=sdf$var_94
p5=sdf$var_95
p6=sdf$var_96
p7=sdf$var_97
p8=sdf$var_98
p9=sdf$var_99
p10=sdf$var_100

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
```

```
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_91","var_92","var_93","var_94","var_95","var_96","var_97","var_98","var_99","var_100"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_101
p2=sdf$var_102
p3=sdf$var_103
p4=sdf$var_104
p5=sdf$var_105
p6=sdf$var_106
p7=sdf$var_107
p8=sdf$var_108
p9=sdf$var_109
p10=sdf$var_110

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_101","var_102","var_103","var_104","var_105","var_106","var_107","var_108","var_109","var_110"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_111
p2=sdf$var_112
p3=sdf$var_113
p4=sdf$var_114
p5=sdf$var_115
p6=sdf$var_116
p7=sdf$var_117
p8=sdf$var_118
p9=sdf$var_119
p10=sdf$var_120

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_111","var_112","var_113","var_114","var_115","var_116","var_117","var_118","var_119","var_120"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_121
p2=sdf$var_122
p3=sdf$var_123
p4=sdf$var_124
p5=sdf$var_125
p6=sdf$var_126
p7=sdf$var_127
p8=sdf$var_128
p9=sdf$var_129
p10=sdf$var_130

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_121","var_122","var_123","var_124","var_125","var_126","var_127","var_128","var_129","var_130"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_131
p2=sdf$var_132
p3=sdf$var_133
p4=sdf$var_134
p5=sdf$var_135
p6=sdf$var_136
p7=sdf$var_137
p8=sdf$var_138
p9=sdf$var_139
p10=sdf$var_140

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_131","var_132","var_133","var_134","var_135","var_136","var_137","var_138","var_139","var_140"),
        las=2,
        col="Orange",
        border="blue",
```

```
        notch=TRUE
)
p1=sdf$var_141
p2=sdf$var_142
p3=sdf$var_143
p4=sdf$var_144
p5=sdf$var_145
p6=sdf$var_146
p7=sdf$var_147
p8=sdf$var_148
p9=sdf$var_149
p10=sdf$var_150

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_141","var_142","var_143","var_144","var_145","var_146","var_147","var_148","var_149","var_150"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_151
p2=sdf$var_152
p3=sdf$var_153
p4=sdf$var_154
p5=sdf$var_155
p6=sdf$var_156
p7=sdf$var_157
p8=sdf$var_158
p9=sdf$var_159
p10=sdf$var_160

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_151","var_152","var_153","var_154","var_155","var_156","var_157","var_158","var_159","var_160"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_161
p2=sdf$var_162
p3=sdf$var_163
p4=sdf$var_164
p5=sdf$var_165
p6=sdf$var_166
p7=sdf$var_167
p8=sdf$var_168
p9=sdf$var_169
p10=sdf$var_170

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_161","var_162","var_163","var_164","var_165","var_166","var_167","var_168","var_169","var_170"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_171
p2=sdf$var_172
p3=sdf$var_173
p4=sdf$var_174
p5=sdf$var_175
p6=sdf$var_176
p7=sdf$var_177
p8=sdf$var_178
p9=sdf$var_179
p10=sdf$var_180

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_171","var_172","var_173","var_174","var_175","var_176","var_177","var_178","var_179","var_180"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
p1=sdf$var_181
p2=sdf$var_182
p3=sdf$var_183
```

```
p4=sdf$var_184
p5=sdf$var_185
p6=sdf$var_186
p7=sdf$var_187
p8=sdf$var_188
p9=sdf$var_189
p10=sdf$var_190

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_181","var_182","var_183","var_184","var_185","var_186","var_187","var_188","var_189","var_190"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)

p1=sdf$var_191
p2=sdf$var_192
p3=sdf$var_193
p4=sdf$var_194
p5=sdf$var_195
p6=sdf$var_196
p7=sdf$var_197
p8=sdf$var_198
p9=sdf$var_199
p10=sdf$var_0

boxplot(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,
        main="box plot for Santander",
        at=c(1,2,3,4,5,6,7,8,9,10),
        names=c("var_191","var_192","var_193","var_194","var_195","var_196","var_197","var_198","var_199","var_0"),
        las=2,
        col="Orange",
        border="blue",
        notch=TRUE
)
sdf1=sdf
numeric_index=sapply(sdf1,is.numeric)
numeric_data=sdf1[,numeric_index]
cnames=colnames(numeric_data)

#to check for outliers & remove it
for(i in cnames)
{
val=sdf1[,i][sdf1[,i] %in% boxplot.stats(sdf1[,i])$out]
sdf1=sdf1[which(!sdf1[,i] %in% val),]
}


#FEATURE SELECTION
#install.packages("corrgram",dependencies = TRUE)
#library(knitr)
#library(fpca)
#library(corrgram)

#numeric_index=sapply(sdf,is.numeric)
#numeric_data=sdf[,numeric_index]

#corrgram(sdf[,numeric_index],order=TRUE,upper.panel=panel.pie,text.panel=panel.txt,main="Correlation Plot")

n=density(sdf$var_0)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_1)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_2)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_3)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_4)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_5)
plot(n)
polygon(n,col='skyblue',border="dark blue")
```

```
n=density(sdf$var_6)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_7)
plot(n)
polygon(n,col='skyblue',border="dark blue")


n=density(sdf$var_8)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_9)
plot(n)
polygon(n,col='skyblue',border="dark blue")



n=density(sdf$var_34)
plot(n)
polygon(n,col='skyblue',border="dark blue")


n=density(sdf$var_47)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_68)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_77)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_109)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_116)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_145)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_161)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_187)
plot(n)
polygon(n,col='skyblue',border="dark blue")

n=density(sdf$var_199)
plot(n)
polygon(n,col='skyblue',border="dark blue")



#SAMPLING & REMOVING ID_CODE

numeric_index=sapply(sdf1,is.numeric)
sdf1=sdf1[,numeric_index]

numeric_index=sapply(sdf,is.numeric)
sdf=sdf[,numeric_index]

install.packages("caret",repos="https://cran.r-project.org/")
install.packages("caTools",repos="https://cran.r-project.org/")
library(caTools)

set.seed(1234)
#train.index=createDataPartition(sdf2$target,p=0.70,list=FALSE)
train.index=sample.split(sdf1$target,SplitRatio=0.7)
train1=sdf1[train.index,]
test1=sdf1[!train.index,]

train.index=sample.split(sdf$target,SplitRatio=0.7)
train=sdf[train.index,]
test=sdf[!train.index,]
```

```
#LOGISTIC REGRESSION
logit_model=glm(target~.,data=train,family="binomial")

#Summary of the model
summary(logit_model)

#predict using logistic regression
logit_prediction=predict(logit_model,newdata = test,type = "response")

#convert prob
logit_prediction=ifelse(logit_prediction>0.5,1,0)

#confusion metrics
ConfMetric_LR=table(test$target,logit_prediction)

TP=53204
FP=4368
FN=767
TN=1661

#FNR=1.42
FNR=FN*100/(FN+TP)
#FPR=72.44
FPR=FP*100/(FP+TN)

#TPR(Recall)=98.5
TPR=TP*100/(TP+FN)

#TNR=27.5
TNR=TN*100/(TN+FP)

#Accuracy=91.4
accuracy=(TP+TN)/60000

#precision=92.4
Precision=TP*100/(TP+FP)

#ROCR curve
library(ROCR)

ROCRpred=prediction(logit_prediction,test$target)
ROCRperf=performance(ROCRpred,'tpr','fpr')

plot(ROCRperf,colorize=TRUE,text.adj=c(-0.2,1.7))

#AUC
library(mltools)
auc=auc_roc(logit_prediction,test$target)
#auc=0.630



#CONfusion metrics
install.packages(caret)
install.packages("pbkrtest", dependencies = TRUE)

library(ROCR)
library(ggplot2)
library(caret)

#confusion metrics
ConfMetric_RF=table(test$target,RF_Prediction)
confusionMatrix(ConfMetric_RF)


install.packages("caret",dependencies = TRUE)
install.packages("e1071",dependencies = TRUE)




#DECISION TREE
library(C50)
library(caret)
train$target=as.factor(train$target)
C50_model=C5.0(target~.,train,trials=1,rules=TRUE)
summary(C50_model)
C50_Predict=predict(C50_model,test[,-1],type="class")
ConfMatrix_C50=table(test$target,C50_Predict)
confusionMatrix(ConfMatrix_C50)

library(mltools)
auc_DT=auc_roc(C50_Predict,test$target)
```

```
#auc=0.53

TP=53181
TN=472
FN=789
FP=5557

#FNR=1.46
FNR=FN*100/(FN+TP)
#FPR=92.17
FPR=FP*100/(FP+TN)

#TPR(Recall)=98.5
TPR=TP*100/(TP+FN)

#TNR=7.82
TNR=TN*100/(TN+FP)

#Accuracy=89.4
accuracy=(TP+TN)/60000

#precision=90.53
Precision=TP*100/(TP+FP)

#NaiVE BAYES
library(e1071)
library(caret)
NB_model=naiveBayes(train[,2:201],train[,1])
NB_Predictions=predict(NB_model,test[,-1])
confmatrix_NB=table(observed=test[,1],predicted=NB_Predictions)
confusionMatrix(confmatrix_NB)

library(mltools)
auc_nb=auc_roc(NB_Predictions,test$target)
#auc=0.67

TP=53072
TN=2212
FN=899
FP=3817

#FNR=1.66
FNR=FN*100/(FN+TP)
#FPR=63.31
FPR=FP*100/(FP+TN)

#TPR(Recall)=98.33
TPR=TP*100/(TP+FN)

#TNR=36.68
TNR=TN*100/(TN+FP)

#Accuracy=92.14
accuracy=(TP+TN)/60000

#precision=93.29
Precision=TP*100/(TP+FP)



#DEPLOYING FINAL MODEL TO TEST CASE

#loading file
sdf_test=read.csv("test.csv")


library("VIM")


#finding out number of missing value
missing_val=data.frame(apply(sdf_test,2,function(x){sum(is.na(x))}))

#giving names in dataframe
missing_val$Columns=row.names(missing_val)
row.names(missing_val)=NULL
names(missing_val)[1]="Missing_Percentage"

#converting to percentage
missing_val$Missing_Percentage=(missing_val$Missing_Percentage/nrow(sdf_test))*100

#Arranging in descending order
missing_val=missing_val[order(-missing_val$Missing_Percentage),]

numeric_index=sapply(sdf_test,is.numeric)
sdf_test=sdf_test[,numeric_index]
```

```
#Predicting
NB_Predictions=predict(NB_model,sdf_test)



#ADDING THE PREDICT_AMOUNT TO THE DATASET
sdf_final=read.csv("test.csv",header=T)
sdf_final$predicted_target=with(sdf_final,NB_Predictions)
sdf_final


#WRITING THE FINAL DATA INTI HDD
write.csv(sdf_final,"Santander_predict.csv",row.names = T)
```