**# Data Transformation**

library(caTools)

library(tidyverse)

# Raw data

rawdata <- read.csv("UCI\_Credit\_Card.csv")

# Removing ID

newdata <- rawdata[,-1]

# Removing Education level 5 and 6

newdata <- newdata %>% filter(EDUCATION <= 4)

# Removing Marriage level 0

newdata <- newdata %>% filter(MARRIAGE != 0)

# Binning PAY\_0-PAY\_6 into 3 categories:

## 0: Payed Duly 1: Payed within 2 months 2: Payed after 2 months

for (i in seq(6,11)) {

newdata[,i][newdata[,i] <= 0] <- 0

newdata[,i][newdata[,i] ==2] <- 1

newdata[,i][newdata[,i] >=3] <- 2

}

# Changing -ve bill\_amount to 0

for (i in seq(12,17)) {

newdata[,i][newdata[,i] < 0] <- 0

}

logdata <- newdata

# Log of LIMIT\_BAL, BILL\_AMT1-6, PAY\_AMT1-6

logdata$LIMIT\_BAL <- log(newdata$LIMIT\_BAL)

for (i in seq(12,23)) {

logdata[,i] <- log(newdata[,i]+1)

}

#Splitting Data

set.seed(123)

fractionTraining <- 2/3

fractionValidation <- 6/30

fractionTest <- 4/30

sampleSizeTraining <- floor(fractionTraining \* nrow(newdata))

sampleSizeValidation <- floor(fractionValidation \* nrow(newdata))

sampleSizeTest <- floor(fractionTest \* nrow(newdata))

indicesTraining <- sort(sample(seq\_len(nrow(newdata)), size=sampleSizeTraining))

indicesNotTraining <- setdiff(seq\_len(nrow(newdata)), indicesTraining)

indicesValidation <- sort(sample(indicesNotTraining, size=sampleSizeValidation))

indicesTest <- setdiff(indicesNotTraining, indicesValidation)

# Without log data

trainingdata <- newdata[indicesTraining, ]

validationdata <- newdata[indicesValidation, ]

testdata <- newdata[indicesTest, ]

# Log data

logtrainingdata <- logdata[indicesTraining, ]

logvalidationdata <- logdata[indicesValidation, ]

logtestdata <- logdata[indicesTest, ]

# Exporting to csv

write.csv(trainingdata, "TrainingSet.csv")

write.csv(validationdata, "ValidationSet.csv")

write.csv(testdata, "TestSet.csv")

write.csv(logtrainingdata, "LogTrainingSet.csv")

write.csv(logvalidationdata, "LogValidationSet.csv")

write.csv(logtestdata, "LogTestSet.csv")

**# Logit Regression**

library(caret)

library(lattice)

library(ggplot2)

library(rpart)

library(rpart.plot)

library(rattle)

library(tree)

library(gmodels)

library(RColorBrewer)

#loding the data and viewing.

a\_train<-read.csv("C:\\Users\\ashwi\\Desktop\\Project\\LogTrainingSet.csv")

a\_valid<-read.csv("C:\\Users\\ashwi\\Desktop\\Project\\LogValidationSet.csv")

#change the numerical to factor

a\_train$SEX<-as.factor(a\_train$SEX)

a\_train$EDUCATION<-as.factor(a\_train$EDUCATION)

a\_train$MARRIAGE<-as.factor(a\_train$MARRIAGE)

a\_valid$SEX<-as.factor(a\_valid$SEX)

a\_valid$EDUCATION<-as.factor(a\_valid$EDUCATION)

a\_valid$MARRIAGE<-as.factor(a\_valid$MARRIAGE)

View(a\_valid)

logit.class<-glm(default.payment.next.month ~LIMIT\_BAL+AGE+PAY\_0+PAY\_2+PAY\_3+

PAY\_4+PAY\_5+PAY\_6+PAY\_AMT1+PAY\_AMT2+PAY\_AMT3+PAY\_AMT4+PAY\_AMT5+

PAY\_AMT6+SEX+EDUCATION+

MARRIAGE,

family=binomial(link="logit"),data=a\_train)

summary(logit.class)

par(mfrow=c(2,2))

plot(logit.class)

logit<-predict.glm(logit.class,a\_valid)

odds<-exp(logit)

head(odds)

probability<-predict.glm(logit.class,a\_valid,type="response")

head(probability)

coefficients(logit.class)[1]

p <-ifelse(logit.class$fitted.values >= 0.5,0,1)

table(p, a\_train$default.payment.next.month)

p <-ifelse(logit.class$fitted.values >= 0.9,0,1)

table(p, a\_train$default.payment.next.month)

p <-ifelse(logit.class$fitted.values >= 0.1,0,1)

p <-ifelse(logit.class$fitted.values >= 0,0,1)

p <-ifelse(logit.class$fitted.values >= 0.03,0,1)

View(a\_valid)

pr<-ifelse(probability <= 0.3,0,1)

install.packages("e1071")

library(e1071)

confusionMatrix(factor(a\_valid[,25]),factor(pr))

library(ROCR)

rocChart(pr,a\_valid[,25])

**# KNN**

library(caret)

library(lattice)

library(ggplot2)

library(rpart)

library(rpart.plot)

library(rattle)

library(tree)

library(gmodels)

library(RColorBrewer)

#loding the data and viewing.

a\_train<-read.csv("C:\\Users\\ashwi\\Desktop\\Project\\TrainingSet.csv")

a\_valid<-read.csv("C:\\Users\\ashwi\\Desktop\\Project\\ValidationSet.csv")

sum(is.na(a\_train))

n\_train = as.data.frame(scale(a\_train[,-25]))

n\_valid = as.data.frame(scale(a\_valid[,-25]))

#change the numerical to factor

n\_train$SEX<-as.factor(n\_train$SEX)

n\_train$EDUCATION<-as.factor(n\_train$EDUCATION)

n\_train$MARRIAGE<-as.factor(n\_train$MARRIAGE)

n\_valid$SEX<-as.factor(n\_valid$SEX)

n\_valid$EDUCATION<-as.factor(n\_valid$EDUCATION)

n\_valid$MARRIAGE<-as.factor(n\_valid$MARRIAGE)

#creating dummy variables for sex,education and marriage

dummy\_cat<-fastDummies::dummy\_cols(n\_train[3:5])

head(dummy\_cat[4:13])

dummy\_cat\_v<-fastDummies::dummy\_cols(n\_valid[3:5])

head(dummy\_cat\_v[4:13])

#adding the dummies to the dataframe and removing orginal

#and creating a new df with dummies(wd)

n\_train\_wd<-cbind(n\_train[-3:-5],dummy\_cat[4:13])

n\_valid\_wd<-cbind(n\_valid[-3:-5],dummy\_cat\_v[4:13])

defualttrain <- a\_train$default.payment.next.month

defualtvalid<- a\_valid$default.payment.next.month

library(kknn)

library(class)

knn.1 = knn(n\_train\_wd,n\_valid\_wd,cl = defualttrain,k= 1,prob = TRUE)

table(knn.1,defualtvalid)

confusionMatrix(factor(knn.1),factor(defualtvalid))

knn.2 = knn(n\_train\_wd,n\_valid\_wd,cl = defualttrain,k= 2,prob = TRUE)

confusionMatrix(factor(knn.2),factor(defualtvalid))

knn.3 = knn(n\_train\_wd,n\_valid\_wd,cl = defualttrain,k= 3,prob = TRUE)

confusionMatrix(factor(knn.3),factor(defualtvalid))

knn.4 = knn(n\_train\_wd,n\_valid\_wd,cl = defualttrain,k= 4,prob = TRUE)

confusionMatrix(factor(knn.4),factor(defualtvalid))

knn.5 = knn(n\_train\_wd,n\_valid\_wd,cl = defualttrain,k= 5,prob = TRUE)

confusionMatrix(factor(knn.5),factor(defualtvalid))

**#Decision tree**

library(caret)

library(lattice)

library(ggplot2)

library(rpart)

library(rpart.plot)

library(rattle)

library(tree)

library(gmodels)

library(RColorBrewer)

#loding the data and viewing.

a\_train<-read.csv("C:\\Users\\ashwi\\Desktop\\Project\\LogTrainingSet.csv")

a\_valid<-read.csv("C:\\Users\\ashwi\\Desktop\\Project\\LogValidationSet.csv")

View(a\_train)

sum(is.na(a\_train))

#change the numerical to factor

a\_train$SEX<-as.factor(a\_train$SEX)

a\_train$EDUCATION<-as.factor(a\_train$EDUCATION)

a\_train$MARRIAGE<-as.factor(a\_train$MARRIAGE)

a\_valid$SEX<-as.factor(a\_valid$SEX)

a\_valid$EDUCATION<-as.factor(a\_valid$EDUCATION)

a\_valid$MARRIAGE<-as.factor(a\_valid$MARRIAGE)

#creating dummy variables for sex,education and marriage

dummy\_cat<-fastDummies::dummy\_cols(a\_train[3:5])

head(dummy\_cat[4:13])

dummy\_cat\_v<-fastDummies::dummy\_cols(a\_valid[3:5])

head(dummy\_cat\_v[4:13])

#adding the dummies to the dataframe and removing orginal

#and creating a new df with dummies(wd)

a\_train\_wd<-cbind(a\_train[-3:-5],dummy\_cat[4:13])

View(a\_train\_wd)

a\_valid\_wd<-cbind(a\_valid[-3:-5],dummy\_cat\_v[4:13])

View(a\_valid\_wd)

library(C50)

#running a regression tree

set.seed(123)

colnames(a\_train\_wd)

RT1<-rpart(default.payment.next.month ~ LIMIT\_BAL+AGE+PAY\_0+PAY\_2+PAY\_3+

PAY\_4+PAY\_5+PAY\_6+PAY\_AMT1+PAY\_AMT2+PAY\_AMT3+PAY\_AMT4+PAY\_AMT5+

PAY\_AMT6+SEX\_2+SEX\_1+EDUCATION\_2+EDUCATION\_1+EDUCATION\_3+

EDUCATION\_4+EDUCATION\_0+MARRIAGE\_1+MARRIAGE\_2+MARRIAGE\_3,

data = a\_train\_wd,

method="anova",control = rpart.control(maxdepth = 3))

print(RT1)

summary(RT1)

prp(RT1, type = 1, extra = 1, split.font = 1, varlen = -10)

plotcp(RT1)

par(mfrow = c(1,1))

RT1.pred<-predict(RT1,a\_train\_wd,type="vector")

RT1.valid<-predict(RT1,a\_valid\_wd,type="vector")

RMSE(RT1.pred,a\_train\_wd$default.payment.next.month)

RMSE(RT1.valid,a\_valid\_wd$default.payment.next.month)

library(e1071)

View(a\_valid\_wd[22])

pr<-ifelse(RT1.valid >= 0.3,1,0)

confusionMatrix(factor(a\_valid\_wd[,22]),factor(pr))

library(ROCR)

rocChart(pr,a\_valid[,25])

par(mfrow=c(1,2))

boxplot(RT1.pred-a\_train\_wd$default.payment.next.month,main="Boxplot for training data error")

boxplot(RT1.valid-a\_valid\_wd$default.payment.next.month,main="Boxplot for validation data error")

par(mfrow=c(1,1))

RT1.cv<-rpart(default.payment.next.month ~ LIMIT\_BAL+AGE+PAY\_0+PAY\_2+PAY\_3+

PAY\_4+PAY\_5+PAY\_6+PAY\_AMT1+PAY\_AMT2+PAY\_AMT3+PAY\_AMT4+PAY\_AMT5+

PAY\_AMT6+SEX\_2+SEX\_1+EDUCATION\_2+EDUCATION\_1+EDUCATION\_3+

EDUCATION\_4+EDUCATION\_0+MARRIAGE\_1+MARRIAGE\_2+MARRIAGE\_3,

data=a\_train\_wd,method="anova",cp=0.00001,minsplit=2,xval=5)

printcp(RT1.cv)

RT1.pruned<-prune(RT1.cv, cp = RT1.cv$cptable[which.min(RT1.cv$cptable[,"xerror"]),"CP"])

prp(RT1.pruned, type = 1, extra = 1, split.font = 1, varlen = -10)

pruned.valid<-predict(RT1.pruned,a\_valid\_wd,type="vector")

RMSE(pruned.valid,a\_valid\_wd$default.payment.next.month)

length(pruned.valid)

pr<-ifelse(pruned.valid >= 0.3,1,0)

confusionMatrix(factor(a\_valid\_wd[,22]),factor(pr))

library(ROCR)

rocChart(pr,a\_valid[,25])

#Making price as categorical response

TC<-read.csv("C:\\Users\\ashwi\\Desktop\\Project\\TrainingSet.csv")

TC$binneddefault<-as.factor(as.numeric(cut(TC$default.payment.next.month,5)))

View(TC$binneddefault)

TC$SEX<-as.factor(TC$SEX)

TC$EDUCATION<-as.factor(TC$EDUCATION)

TC$MARRIAGE<-as.factor(TC$MARRIAGE)

set.seed(123)

colnames(TC)

CT<-rpart(binneddefault~LIMIT\_BAL+AGE+PAY\_0+PAY\_2+PAY\_3+

PAY\_4+PAY\_5+PAY\_6+PAY\_AMT1+PAY\_AMT2+PAY\_AMT3+PAY\_AMT4+PAY\_AMT5+

PAY\_AMT6+SEX+EDUCATION+

MARRIAGE,data = TC,

method="class",control = rpart.control(maxdepth = 3))

print(CT)

prp(CT, type = 1, extra = 1, split.font = 1, varlen = -10)

plot(CT,margin=0.07)

text(CT,cex=0.8)

plot(RT1,margin=0.07)

text(RT1,cex=0.8)

#Running model on validation data

TC\_valid<-read.csv("C:\\Users\\ashwi\\Desktop\\Project\\ValidationSet.csv")

TC\_valid$SEX<-as.factor(TC\_valid$SEX)

TC\_valid$EDUCATION<-as.factor(TC\_valid$EDUCATION)

TC\_valid$MARRIAGE<-as.factor(TC\_valid$MARRIAGE)

predict(CT,TC\_valid)

predict(RT1,a\_valid\_wd)