**LOGISTIC REGRESSION**

#LOGISTIC REGRESSION

getwd()

setwd("C:\\Users\\10ani\\Desktop\\DS\_classes")

df= read.csv("HR\_Data.csv")

str(df)

df1=df[,-c(21,22,23,24,25,26,27,28,29,30,31,32,33,34,35)]

Level= as.factor(df$JobLevel)

df$JobSatisfaction= as.factor(df$JobSatisfaction)

df$PerformanceRating = as.factor(df$PerformanceRating)

df$Attrition\_txt= NULL

library(plotly)

plot\_ly(data=df,x=~df$Age,type ='histogram')

plot\_ly(data=df,x=~df$MonthlyIncome,type ='histogram')

mean\_sal = aggregate(df$MonthlyIncome, list(JobLevel = df$JobLevel), mean)

df\_mean = merge(df, mean\_sal)

names(df\_mean)[20] = c("mean\_sal")

df\_mean$sal\_diff = df\_mean$MonthlyIncome-df\_mean$mean\_sal

df\_mean$comp\_ratio = df\_mean$MonthlyIncome / df\_mean$mean\_sal

help("table")

cont = table(df\_mean$PerformanceRating, df\_mean$Attrition)

cont

write.csv(df\_mean, "hello.csv")

str(df\_mean)

corr = cor(df\_mean[ ,-c(4,5,6,7,8,12)])

write.csv(corr, "hr\_correlation.csv")

df\_mean$YearsAtCompany = NULL

set.seed(345)

ind = sort(sample(nrow(df\_mean), nrow(df\_mean)\*0.8))

train = df\_mean[ind, ]

test = df\_mean[-ind, ]

glm.fit = glm(Attrition~., train, family = binomial)

pred = predict(glm.fit, test,type="response")

glm.new = step(glm.fit)

summary(glm.fit)

summary(glm.new)

conf\_mat = table(test$Attrition, pred>0.5)

accuracy = sum(diag(conf\_mat))/sum(conf\_mat)

accuracy

i=0

for(i in seq(from=0.1, to=.7, by=0.05))

{

conf\_mat = table(test$Attrition, pred>i)

accuracy = sum(diag(conf\_mat))/sum(conf\_mat)

print(c(i,accuracy))

}

library("pscl")

pR2(glm.new)

# try to improve the McFadden value to 70

library(lmtest)

lrtest(glm.fit,glm.new)

# has less confidence.

library(pROC)

trainpred = glm.fit$fitted.values

trainROC = roc(response = train$Attrition, predictor = trainpred, plot = TRUE, auc = TRUE)

trainROC$auc

library(caret)

varImp(glm.fit)

table1 = table(df$JobSatisfaction, df$Attrition)