setwd("D:\\DS 1\\logistic regression")

data=read.csv("titanic.csv")

str(data)

summary(data)

attach(data)

hist(age)

data$age[is.na(data$age)]=29

summary(data)

sum(is.na(data))

View(data)

data$male=ifelse(data$sex=="male",1,0)

table(embarked)

data$embarked\_c=ifelse(data$embarked=="C",1,0)

data$embarked\_s=ifelse(data$embarked=="S",1,0)

newdata=data[-c(3,4,9)]

View(newdata)

names(newdata)

attach(newdata)

boxplot((pclass),(survived),(age),(sibsp),(parch),(fare),(male),(embarked\_c),(embarked\_s))

boxplot(pclass)

bx=boxplot(age)

quantile(age,seq(0,1,0.02))

bx$stats

newdata$age=ifelse(newdata$age<4,7,newdata$age)

newdata$age=ifelse(newdata$age>52,52,newdata$age)

boxplot(newdata$age)

bx=boxplot(newdata$sibsp)

quantile(newdata$sibsp,seq(0,1,0.02))

bx$stats

newdata$sibsp=ifelse(newdata$sibsp>2,2,newdata$sibsp)

boxplot(newdata$sibsp)

bx=boxplot(newdata$parch)

quantile(newdata$parch,seq(0,1,0.02))

bx$stats

newdata$parch=ifelse(newdata$parch>0,0,newdata$parch)

boxplot(newdata$parch)

bx=boxplot(newdata$fare)

quantile(newdata$fare,seq(0,1,0.02))

bx$stats

newdata$fare=ifelse(newdata$fare>65,59,newdata$fare)

boxplot(newdata$fare)

#lets divide the data into test and train

set.seed(222)

t=sample(1:nrow(newdata),0.7\*nrow(newdata))

t\_train=newdata[t,]

t\_test=newdata[-t,]

#checking the multi-collinearity

library(car)

mod<- lm(survived ~ ., data=t\_train)

mod1 <- glm(as.factor(survived) ~ ., family="binomial", data=t\_train)

summary(mod1)

stpmod = step(mod1, direction = "both")

formula(stpmod)

summary(stpmod)

#checking the probability for each observation by creating a variable names score

mod2 <- glm(as.factor(survived) ~ pclass + age + sibsp + male + embarked\_s, family="binomial", data=t\_train)

summary(mod2)

t\_train$score=predict(mod2,newdata=t\_train,type = "response")

head(t\_train$score)

tail(t\_train$score)

#Lets try to analyse the confusion matrix and model accuracy

library(lattice)

library(ggplot2)

library(caret)

library(e1071)

prediction<-ifelse(t\_train$score>=0.6,1,0)

confusionMatrix(as.factor(prediction),as.factor(t\_train$survived),positive = "1")

str(prediction)

str(t\_train$survived)

##Mcfadden test

library(pscl)

pR2(mod2)

# Concordance Test #

library(InformationValue)

library(caret)

concor <- Concordance(t\_train$survived,t\_train$score)

concor

#lets check the AUC and ROC

##AUC

library(InformationValue)

plotROC(actuals = t\_train$survived,predictedScores = as.numeric(fitted(mod2)))

ks\_plot(actuals = t\_train$survived,predictedScores = as.numeric(fitted(mod2)))

ks\_stat(actuals = t\_train$survived,predictedScores = as.numeric(fitted(mod2)))

t\_test$score2= predict(mod2, t\_test, type="response")

View(t\_test)