**Attirbutes that have been removed**

verification\_status\_joint,

acc\_open\_past\_24mths,

avg\_cur\_bal,

bc\_open\_to\_buy,

bc\_util,

chargeoff\_within\_12\_mths,

dti\_joint,

acc\_now\_delinq,

delinq\_amnt,

mo\_sin\_old\_il\_acct,

mo\_sin\_old\_rev\_tl\_op,

mo\_sin\_rcnt\_rev\_tl\_op,

mo\_sin\_rcnt\_tl,

mort\_acc,

mths\_since\_recent\_bc,

mths\_since\_recent\_bc\_dlq,

mths\_since\_recent\_inq,

mths\_since\_recent\_revol\_delinq,

emp\_title,

num\_accts\_ever\_120\_pd,

num\_actv\_bc\_tl,

num\_actv\_rev\_tl,

num\_bc\_sats,

num\_bc\_tl,

num\_il\_tl,

num\_op\_rev\_tl,

num\_rev\_accts,

num\_rev\_tl\_bal\_gt\_0,

num\_sats,

num\_tl\_120dpd\_2m,

num\_tl\_30dpd,

num\_tl\_90g\_dpd\_24m,

num\_tl\_op\_past\_12m,

pct\_tl\_nvr\_dlq,

total\_bal\_ex\_mort,

total\_bc\_limit,

total\_il\_high\_credit\_limit.

**Code used**

LoanDataSet <- read.csv("E:/study/Data Mining/project/LoanDataSet.csv")

View(LoanDataSet)

LoanDataSet <- subset(LoanDataSet, , -c(verification\_status\_joint, acc\_open\_past\_24mths, avg\_cur\_bal, bc\_open\_to\_buy, bc\_util, chargeoff\_within\_12\_mths))

LoanDataSet <- subset(LoanDataSet, , -c(dti\_joint, acc\_now\_delinq, delinq\_amnt, mo\_sin\_old\_il\_acct, mo\_sin\_old\_rev\_tl\_op, mo\_sin\_rcnt\_rev\_tl\_op))

LoanDataSet <- subset(LoanDataSet, , -c(mo\_sin\_rcnt\_tl, mort\_acc, mths\_since\_recent\_bc, mths\_since\_recent\_bc\_dlq, mths\_since\_recent\_inq, mths\_since\_recent\_revol\_delinq))

LoanDataSet <- subset(LoanDataSet, , -c(emp\_title, num\_accts\_ever\_120\_pd, num\_actv\_bc\_tl, num\_actv\_rev\_tl, num\_bc\_sats, num\_bc\_tl, num\_il\_tl, num\_op\_rev\_tl, num\_rev\_accts, num\_rev\_tl\_bal\_gt\_0, num\_sats, num\_tl\_120dpd\_2m, num\_tl\_30dpd, num\_tl\_90g\_dpd\_24m, num\_tl\_op\_past\_12m, pct\_tl\_nvr\_dlq, total\_bal\_ex\_mort, total\_bc\_limit, total\_il\_high\_credit\_limit))

LoanDataSet <- subset(LoanDataSet, , -c(emp\_length, url, desc, title))

library(sqldf)

sqldf("select count(\*) from LoanDataSet")

PPlanLoanDataSet <- sqldf("select \* from LoanDataSet where pymnt\_plan='y'")

View(PPlanLoanDataSet)

VerifiedLDS <- sqldf("select \* from LoanDataSet where verification\_status != 'Not Verified'")

View(VerifiedLDS)

VerifiedLDS <-

VerifiedLDS <- sqldf("select \* from VerifiedLDS where member\_id > 0")

VerifiedLDS <- subset(VerifiedLDS, , -c(pymnt\_plan, verification\_status))

sqldf("select count(\*) from VerifiedLDS as v1, verifiedLDS as v2 where v1.id = v2.id and v1.member\_id!=v2.member\_id")

sqldf("select \* from VerifiedLDS as v1, verifiedLDS as v2 where v1.id = v2.id and v1.member\_id!=v2.member\_id")

sqldf("select count(\*) from VerifiedLDS as v1, verifiedLDS as v2 where v1.id != v2.id and v1.member\_id==v2.member\_id")

sqldf("select count(\*) from VerifiedLDS group by id having count(\*)>1")

sqldf("select count(\*) from VerifiedLDS group by member\_id having count(\*)>1")

savehistory("F:/notes/Data\_Mining\_Tools\_and\_Techniques/Project/Final dataset/Cleaning.Rhistory")

**LDSPP to LDSPP2**

LDSPostProcessing2 <- sqldf("select \* from LDSPostProcessing where delinq\_2yrs > -1")

LDSPostProcessing2 <- sqldf("select \* from LDSPostProcessing2 where dti > 0")

LDSPostProcessing2$loan\_status\_scale[LDSPostProcessing2\_intRate\_12$interest\_rate<12] <- 0

LDSPostProcessing2$loan\_status\_scale[LDSPostProcessing2\_intRate\_12$interest\_rate>=12] <- 1

round(LDSPostProcessing2$LTI, digitis = 4)

**Correlation Analysis:**

loan\_amnt <- LDSPostProcessing2 $loan\_amnt

funded\_amnt <- LDSPostProcessing2 $funded\_amnt

cor(loan\_amnt, funded\_amnt)

cor.test(loan\_amnt, funded\_amnt)

funded\_amnt\_inv <- LDSPostProcessing2$funded\_amnt\_inv

cor.test(loan\_amnt, funded\_amnt\_inv)

**Visualization Techniques:**

1) gradeA <- loan[loan[, "grade"] == "A",]

gradeB <- loan[loan[, "grade"] == "B",]

gradeC <- loan[loan[, "grade"] == "C",]

gradeD <- loan[loan[, "grade"] == "D",]

gradeE <- loan[loan[, "grade"] == "E",]

gradeF <- loan[loan[, "grade"] == "F",]

gradeG <- loan[loan[, "grade"] == "G",]

barplot(matrix(c(mean(gradeA[["int\_rate"]]),mean(gradeB[["int\_rate"]]),mean(gradeC[["int\_rate"]]),mean(gradeD[["int\_rate"]]),mean(gradeE[["int\_rate"]]),mean(gradeF[["int\_rate"]]),mean(gradeG[["int\_rate"]])),nr=1),

col=c("aquamarine3"),

names.arg=LETTERS[1:7],main = "HISTORICAL INTEREST RATE BY GRADE",ylab="interest rate avg",xlab="Grade")

2) \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

int\_rate\_2007 <- loan[loan[, "issue\_d"] == "2007",]

int\_rate\_2008 <- loan[loan[, "issue\_d"] == "2008",]

int\_rate\_2009 <- loan[loan[, "issue\_d"] == "2009",]

int\_rate\_2010 <- loan[loan[, "issue\_d"] == "2010",]

int\_rate\_2011 <- loan[loan[, "issue\_d"] == "2011",]

int\_rate\_2012 <- loan[loan[, "issue\_d"] == "2012",]

int\_rate\_2013 <- loan[loan[, "issue\_d"] == "2013",]

int\_rate\_2014 <- loan[loan[, "issue\_d"] == "2014",]

int\_rate\_2015 <- loan[loan[, "issue\_d"] == "2015",]

avg\_int\_rate <- matrix(c(mean(int\_rate\_2007[["int\_rate"]]),mean(int\_rate\_2008[["int\_rate"]]),mean(int\_rate\_2009[["int\_rate"]]),mean(int\_rate\_2010[["int\_rate"]]),mean(int\_rate\_2011[["int\_rate"]]),mean(int\_rate\_2012[["int\_rate"]]),mean(int\_rate\_2013[["int\_rate"]]),mean(int\_rate\_2014[["int\_rate"]]),mean(int\_rate\_2015[["int\_rate"]])),ncol=9,byrow=TRUE)

colnames(avg\_int\_rate) <- c("2007","2008","2009","2010","2011","2012","2013","2014","2015")

rownames(avg\_int\_rate) <- c("int\_rate")

avg\_int\_rate <- as.table(avg\_int\_rate)

barplot(avg\_int\_rate, main = "AVERAGE INTEREST RATE",xlab="Year",ylab="avg of interest rate")

3) a2007 <- nrow(int\_rate\_2007[int\_rate\_2007[, "grade"] == "A",]) \* 100/ nrow(int\_rate\_2007)

b2007 <- nrow(int\_rate\_2007[int\_rate\_2007[, "grade"] == "B",]) \* 100/ nrow(int\_rate\_2007)

c2007 <- nrow(int\_rate\_2007[int\_rate\_2007[, "grade"] == "C",]) \* 100/ nrow(int\_rate\_2007)

d2007 <- nrow(int\_rate\_2007[int\_rate\_2007[, "grade"] == "D",]) \* 100/ nrow(int\_rate\_2007)

e2007 <- nrow(int\_rate\_2007[int\_rate\_2007[, "grade"] == "E",]) \* 100/ nrow(int\_rate\_2007)

f2007 <- nrow(int\_rate\_2007[int\_rate\_2007[, "grade"] == "F",]) \* 100/ nrow(int\_rate\_2007)

g2007 <- nrow(int\_rate\_2007[int\_rate\_2007[, "grade"] == "G",]) \* 100/ nrow(int\_rate\_2007)

a2008 <- nrow(int\_rate\_2008[int\_rate\_2008[, "grade"] == "A",]) \* 100/ nrow(int\_rate\_2008)

b2008 <- nrow(int\_rate\_2008[int\_rate\_2008[, "grade"] == "B",]) \* 100/ nrow(int\_rate\_2008)

c2008 <- nrow(int\_rate\_2008[int\_rate\_2008[, "grade"] == "C",]) \* 100/ nrow(int\_rate\_2008)

d2008 <- nrow(int\_rate\_2008[int\_rate\_2008[, "grade"] == "D",]) \* 100/ nrow(int\_rate\_2008)

e2008 <- nrow(int\_rate\_2008[int\_rate\_2008[, "grade"] == "E",]) \* 100/ nrow(int\_rate\_2008)

f2008 <- nrow(int\_rate\_2008[int\_rate\_2008[, "grade"] == "F",]) \* 100/ nrow(int\_rate\_2008)

g2008 <- nrow(int\_rate\_2008[int\_rate\_2008[, "grade"] == "G",]) \* 100/ nrow(int\_rate\_2008)

a2009 <- nrow(int\_rate\_2009[int\_rate\_2009[, "grade"] == "A",]) \* 100/ nrow(int\_rate\_2009)

b2009 <- nrow(int\_rate\_2009[int\_rate\_2009[, "grade"] == "B",]) \* 100/ nrow(int\_rate\_2009)

c2009 <- nrow(int\_rate\_2009[int\_rate\_2009[, "grade"] == "C",]) \* 100/ nrow(int\_rate\_2009)

d2009 <- nrow(int\_rate\_2009[int\_rate\_2009[, "grade"] == "D",]) \* 100/ nrow(int\_rate\_2009)

e2009 <- nrow(int\_rate\_2009[int\_rate\_2009[, "grade"] == "E",]) \* 100/ nrow(int\_rate\_2009)

f2009 <- nrow(int\_rate\_2009[int\_rate\_2009[, "grade"] == "F",]) \* 100/ nrow(int\_rate\_2009)

g2009 <- nrow(int\_rate\_2009[int\_rate\_2009[, "grade"] == "G",]) \* 100/ nrow(int\_rate\_2009)

a2010 <- nrow(int\_rate\_2010[int\_rate\_2010[, "grade"] == "A",]) \* 100/ nrow(int\_rate\_2010)

b2010 <- nrow(int\_rate\_2010[int\_rate\_2010[, "grade"] == "B",]) \* 100/ nrow(int\_rate\_2010)

c2010 <- nrow(int\_rate\_2010[int\_rate\_2010[, "grade"] == "C",]) \* 100/ nrow(int\_rate\_2010)

d2010 <- nrow(int\_rate\_2010[int\_rate\_2010[, "grade"] == "D",]) \* 100/ nrow(int\_rate\_2010)

e2010 <- nrow(int\_rate\_2010[int\_rate\_2010[, "grade"] == "E",]) \* 100/ nrow(int\_rate\_2010)

f2010 <- nrow(int\_rate\_2010[int\_rate\_2010[, "grade"] == "F",]) \* 100/ nrow(int\_rate\_2010)

g2010 <- nrow(int\_rate\_2010[int\_rate\_2010[, "grade"] == "G",]) \* 100/ nrow(int\_rate\_2010)

a2011 <- nrow(int\_rate\_2011[int\_rate\_2011[, "grade"] == "A",]) \* 100/ nrow(int\_rate\_2011)

b2011 <- nrow(int\_rate\_2011[int\_rate\_2011[, "grade"] == "B",]) \* 100/ nrow(int\_rate\_2011)

c2011 <- nrow(int\_rate\_2011[int\_rate\_2011[, "grade"] == "C",]) \* 100/ nrow(int\_rate\_2011)

d2011 <- nrow(int\_rate\_2011[int\_rate\_2011[, "grade"] == "D",]) \* 100/ nrow(int\_rate\_2011)

e2011 <- nrow(int\_rate\_2011[int\_rate\_2011[, "grade"] == "E",]) \* 100/ nrow(int\_rate\_2011)

f2011 <- nrow(int\_rate\_2011[int\_rate\_2011[, "grade"] == "F",]) \* 100/ nrow(int\_rate\_2011)

g2011 <- nrow(int\_rate\_2011[int\_rate\_2011[, "grade"] == "G",]) \* 100/ nrow(int\_rate\_2011)

a2012 <- nrow(int\_rate\_2012[int\_rate\_2012[, "grade"] == "A",]) \* 100/ nrow(int\_rate\_2012)

b2012 <- nrow(int\_rate\_2012[int\_rate\_2012[, "grade"] == "B",]) \* 100/ nrow(int\_rate\_2012)

c2012 <- nrow(int\_rate\_2012[int\_rate\_2012[, "grade"] == "C",]) \* 100/ nrow(int\_rate\_2012)

d2012 <- nrow(int\_rate\_2012[int\_rate\_2012[, "grade"] == "D",]) \* 100/ nrow(int\_rate\_2012)

e2012 <- nrow(int\_rate\_2012[int\_rate\_2012[, "grade"] == "E",]) \* 100/ nrow(int\_rate\_2012)

f2012 <- nrow(int\_rate\_2012[int\_rate\_2012[, "grade"] == "F",]) \* 100/ nrow(int\_rate\_2012)

g2012 <- nrow(int\_rate\_2012[int\_rate\_2012[, "grade"] == "G",]) \* 100/ nrow(int\_rate\_2012)

a2013 <- nrow(int\_rate\_2013[int\_rate\_2013[, "grade"] == "A",]) \* 100/ nrow(int\_rate\_2013)

b2013 <- nrow(int\_rate\_2013[int\_rate\_2013[, "grade"] == "B",]) \* 100/ nrow(int\_rate\_2013)

c2013 <- nrow(int\_rate\_2013[int\_rate\_2013[, "grade"] == "C",]) \* 100/ nrow(int\_rate\_2013)

d2013 <- nrow(int\_rate\_2013[int\_rate\_2013[, "grade"] == "D",]) \* 100/ nrow(int\_rate\_2013)

e2013 <- nrow(int\_rate\_2013[int\_rate\_2013[, "grade"] == "E",]) \* 100/ nrow(int\_rate\_2013)

f2013 <- nrow(int\_rate\_2013[int\_rate\_2013[, "grade"] == "F",]) \* 100/ nrow(int\_rate\_2013)

g2013 <- nrow(int\_rate\_2013[int\_rate\_2013[, "grade"] == "G",]) \* 100/ nrow(int\_rate\_2013)

a2014 <- nrow(int\_rate\_2014[int\_rate\_2014[, "grade"] == "A",]) \* 100/ nrow(int\_rate\_2014)

b2014 <- nrow(int\_rate\_2014[int\_rate\_2014[, "grade"] == "B",]) \* 100/ nrow(int\_rate\_2014)

c2014 <- nrow(int\_rate\_2014[int\_rate\_2014[, "grade"] == "C",]) \* 100/ nrow(int\_rate\_2014)

d2014 <- nrow(int\_rate\_2014[int\_rate\_2014[, "grade"] == "D",]) \* 100/ nrow(int\_rate\_2014)

e2014 <- nrow(int\_rate\_2014[int\_rate\_2014[, "grade"] == "E",]) \* 100/ nrow(int\_rate\_2014)

f2014 <- nrow(int\_rate\_2014[int\_rate\_2014[, "grade"] == "F",]) \* 100/ nrow(int\_rate\_2014)

g2014 <- nrow(int\_rate\_2014[int\_rate\_2014[, "grade"] == "G",]) \* 100/ nrow(int\_rate\_2014)

a2015 <- nrow(int\_rate\_2015[int\_rate\_2015[, "grade"] == "A",]) \* 100/ nrow(int\_rate\_2015)

b2015 <- nrow(int\_rate\_2015[int\_rate\_2015[, "grade"] == "B",]) \* 100/ nrow(int\_rate\_2015)

c2015 <- nrow(int\_rate\_2015[int\_rate\_2015[, "grade"] == "C",]) \* 100/ nrow(int\_rate\_2015)

d2015 <- nrow(int\_rate\_2015[int\_rate\_2015[, "grade"] == "D",]) \* 100/ nrow(int\_rate\_2015)

e2015 <- nrow(int\_rate\_2015[int\_rate\_2015[, "grade"] == "E",]) \* 100/ nrow(int\_rate\_2015)

f2015 <- nrow(int\_rate\_2015[int\_rate\_2015[, "grade"] == "F",]) \* 100/ nrow(int\_rate\_2015)

g2015 <- nrow(int\_rate\_2015[int\_rate\_2015[, "grade"] == "G",]) \* 100/ nrow(int\_rate\_2015)

grade\_mix\_year <- matrix(c(a2007,a2008,a2009,a2010,a2011,a2012,a2013,a2014,a2015,

b2007,b2008,b2009,b2010,b2011,b2012,b2013,b2014,b2015,

c2007,c2008,c2009,c2010,c2011,c2012,c2013,c2014,c2015,

d2007,d2008,d2009,d2010,d2011,d2012,d2013,d2014,d2015,

e2007,e2008,e2009,e2010,e2011,e2012,e2013,e2014,e2015,

f2007,f2008,f2009,f2010,f2011,f2012,f2013,f2014,f2015,

g2007,g2008,g2009,g2010,g2011,g2012,g2013,g2014,g2015),ncol=9,byrow=TRUE)

colnames(grade\_mix\_year) <- c("2007","2008","2009","2010","2011","2012","2013","2014",2015)

rownames(grade\_mix\_year) <- c("A","B","C","D","E","F","G")

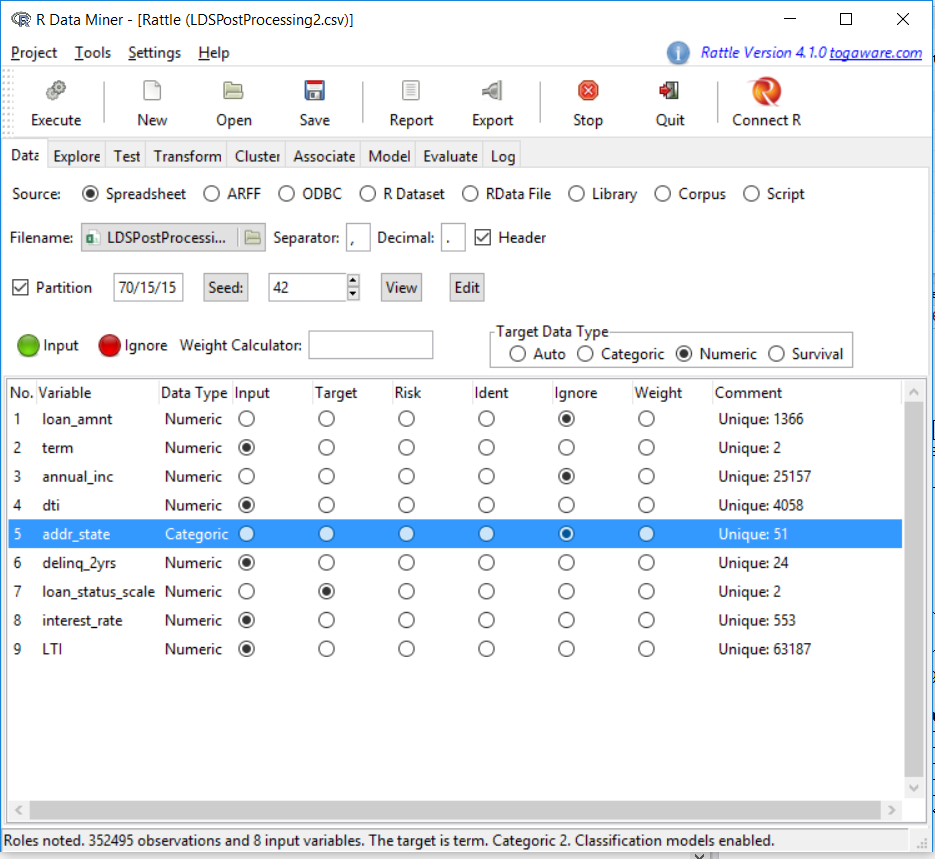
grade\_mix\_year\_table <- as.table(grade\_mix\_year)

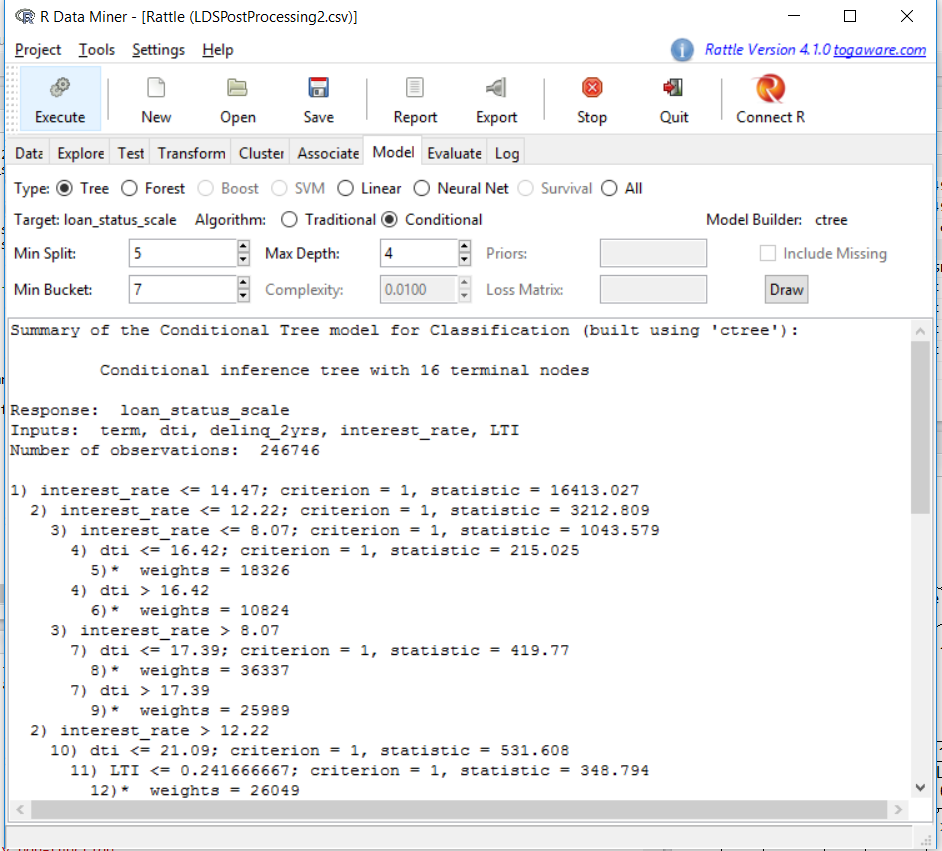
barplot(grade\_mix\_year\_table,main='GRADE MIX OVER TIME',ylab="Percentages for Grades A-G",xlab = "Years")

**Decision Tree:**

library(rattle)

rattle()





**Neural Network:**

library(grid)

library(MASS)

library(neuralnet)

temp1 = sample(352495, 200000)

temp\_data = LDSPostProcessing2[temp1, ]

nn <- neuralnet(formula = loan\_status\_scale ~ interest\_rate + dti + LTI + delinq\_2yrs + term, data = temp\_data, hidden = 3, stepmax = 1e6, linear.output = FALSE, algorithm = "backprop", learningrate = 0.01)

nn$net.result

nn

plot(nn)

nn$result.matrix

**Testing of Neural Network:**

temp = sample(352495, 100000)

test\_data = LDSPostProcessing2[temp, ]

temp\_test <- subset(test\_data, select = c("interest\_rate", "dti", "LTI", "delinq\_2yrs","term"))

nn.results <- compute(nn, temp\_test)

results <- data.frame(actual = test\_data$loan\_status\_scale, prediction = nn59.results$net.result)

results

results$prediction <- round(results$prediction)

results

**Evaluating the Results:**

install.packages("minqa")

install.packages("caret")

library(lattice)

library(ggplot2)

library(minqa)

library(caret)

install.packages("e1071")

library(e1071)

confusionMatrix(test\_data$loan\_status\_scale, results$prediction)