**CHI SQUARE TEST**

library(ade4)

library(dplyr)

library(data.table)

df <- read.table("bank\_full\_cleaned.csv", sep=",", stringsAsFactors = FALSE,header = TRUE)

head(df)

chisq.test(df$job, df$y, correct=FALSE)

# Pearson's Chi-squared test

#

# data: df$job and df$y

# X-squared = 836.11, df = 11, p-value < 2.2e-16

chisq.test(df$marital, df$y, correct=FALSE)

# Pearson's Chi-squared test

#

# data: df$marital and df$y

# X-squared = 196.5, df = 2, p-value < 2.2e-16

chisq.test(df$education, df$y, correct=FALSE)

# Pearson's Chi-squared test

#

# data: df$education and df$y

# X-squared = 238.92, df = 3, p-value < 2.2e-16

chisq.test(df$default, df$y, correct=FALSE)

# Pearson's Chi-squared test

#

# data: df$default and df$y

# X-squared = 22.724, df = 1, p-value = 1.871e-06

chisq.test(df$housing, df$y, correct=FALSE)

# Pearson's Chi-squared test

#

# data: df$housing and df$y

# X-squared = 875.69, df = 1, p-value < 2.2e-16

chisq.test(df$contact, df$y, correct=FALSE)

# Pearson's Chi-squared test

#

# data: df$contact and df$y

# X-squared = 1035.7, df = 2, p-value < 2.2e-16

chisq.test(df$loan, df$y, correct=FALSE)

# Pearson's Chi-squared test

#

# data: df$loan and df$y

# X-squared = 210.19, df = 1, p-value < 2.2e-16

**MODEL 1**

1. Outlier treatment on duration done
2. Outlier treatment on balance done
3. Biased sampling not done

**#code for overall ---- iteration 1**

**install.packages('caTools')**

**library(caTools)**

**library(ade4)**

**library(dplyr)**

**library(data.table)**

**#dependent variable if the change was positive two days in a row!**

**###########################**

**# CUSTOMIZE DATA FILE: read.table reads a file in table format and creates a data frame from it**

**df <- read.table("df\_final\_ish.csv", sep=",", stringsAsFactors = FALSE,header = TRUE)**

**head(df)**

**df['zero\_previous'] <- ifelse(df$previous == 0,1,0)**

**head(df)**

**df['pdays'] <- ifelse(df$pdays == -1,1,0)**

**head(df)**

**df\_new <- subset(df,select = -c(day,month,pdays,previous,job\_unknown,education\_unknown,contact\_unknown,poutcome\_unknown,marital\_single,X,zero\_previous) )**

**head(df\_new)**

**normalize <- function(x){**

**return ((x - min(x))/(max(x) - min(x)))**

**}**

**a = df\_new$balance**

**quantile(a, c(.99)) #13164.9**

**b = df\_new$duration**

**quantile(b, c(.95)) #751**

**df\_new['balance\_new']<- ifelse(df\_new$balance > 13164.9,13164.9,df\_new$balance)**

**df\_new['duration\_new']<- ifelse(df\_new$duration > 751,751,df\_new$duration)**

**df\_new1 <- subset(df\_new,select = -c(balance,duration) )**

**data <- sapply(df\_new1,normalize)**

**head(data)**

**#df['dep'] <- ifelse(df$y == "yes",1,0)**

**set.seed(33333) # Set Seed so that same sample can be reproduced in future also**

**# Now Selecting 75% of data as sample from total 'n' rows of the data**

**head(data)**

**sample <- sample.int(n = nrow(data), size = floor(.80\*nrow(data)), replace = F)**

**train <- data[sample, ]**

**test <- data[-sample, ]**

**head(train)**

**head(test)**

**train1 <- as.data.frame(train)**

**class(train1)**

**test1 <- as.data.frame(test)**

**class(test1)**

**#logistic regression model**

**model <- glm (y\_yes ~ ., data = train1, family = binomial)**

**summary(model)**

**predict <- predict(model, type = 'response')**

**#confusion matrix**

**table(train1$y\_yes, predict > 0.5)**

**pred <- predict(model, newdata = test1, type = "response")**

**hist(pred)**

**y\_pred\_num <- ifelse(pred > 0.5, 1, 0)**

**y\_pred <- factor(y\_pred\_num, levels=c(0, 1))**

**y\_act <- test1$y\_yes**

**y\_act**

**table(y\_pred\_num,y\_act)**

**mean(y\_pred == y\_act)**

**#ROCR Curve**

**library(ROCR)**

**ROCRpred <- prediction(predict, train1$y\_yes)**

**ROCRperf <- performance(ROCRpred, 'tpr','fpr')**

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

**OUTPUT**

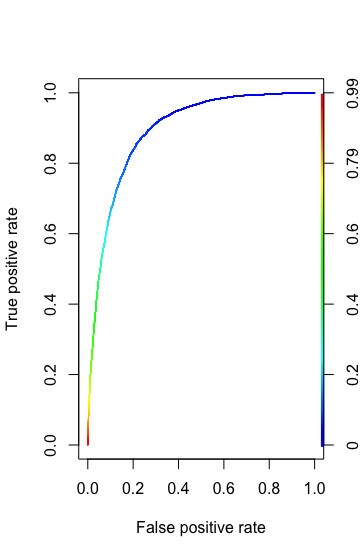
**Accuracy : 0.8999226**

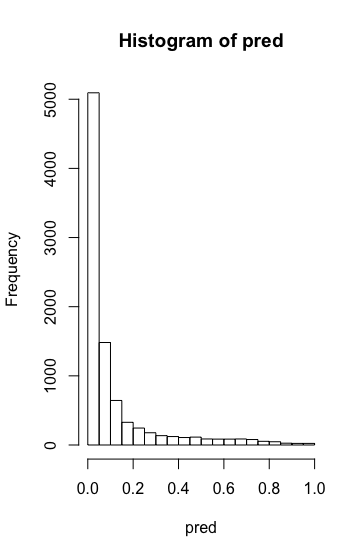
**y\_act**

**y\_pred\_num 0 1**

**0 7764 684**

**1 221 374**

****

****

**MODEL:**

**Call:**

**glm(formula = y\_yes ~ ., family = binomial, data = train1)**

**Deviance Residuals:**

**Min 1Q Median 3Q Max**

**-2.9550 -0.3783 -0.2381 -0.1405 3.2828**

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

**(Intercept) -4.87434 0.28334 -17.203 < 2e-16 \*\*\***

**age 0.03141 0.18674 0.168 0.866411**

**job\_admin. 0.41435 0.26287 1.576 0.114966**

**job\_blue.collar 0.02448 0.26186 0.093 0.925522**

**job\_entrepreneur -0.12632 0.28351 -0.446 0.655905**

**job\_housemaid -0.18450 0.29083 -0.634 0.525839**

**job\_management 0.15367 0.26101 0.589 0.556031**

**job\_retired 0.76743 0.26759 2.868 0.004131 \*\***

**job\_self.employed 0.06001 0.27690 0.217 0.828411**

**job\_services 0.14733 0.26671 0.552 0.580670**

**job\_student 0.96612 0.27593 3.501 0.000463 \*\*\***

**job\_technician 0.11096 0.26093 0.425 0.670653**

**job\_unemployed 0.19832 0.27748 0.715 0.474795**

**marital\_divorced -0.15228 0.07362 -2.068 0.038594 \***

**marital\_married -0.30410 0.05024 -6.053 1.42e-09 \*\*\***

**education\_primary -0.20933 0.11554 -1.812 0.070015 .**

**education\_secondary -0.06073 0.10213 -0.595 0.552049**

**education\_tertiary 0.20657 0.10753 1.921 0.054741 .**

**default\_yes -0.37216 0.19393 -1.919 0.054984 .**

**housing\_yes -0.79386 0.04415 -17.983 < 2e-16 \*\*\***

**loan\_yes -0.53284 0.06422 -8.297 < 2e-16 \*\*\***

**contact\_cellular 1.05042 0.06193 16.961 < 2e-16 \*\*\***

**contact\_telephone 1.02341 0.09817 10.425 < 2e-16 \*\*\***

**campaign -6.56327 0.70119 -9.360 < 2e-16 \*\*\***

**poutcome\_failure 0.42502 0.06136 6.926 4.32e-12 \*\*\***

**poutcome\_other 0.64120 0.08587 7.467 8.18e-14 \*\*\***

**poutcome\_success 2.74966 0.07414 37.088 < 2e-16 \*\*\***

**balance\_new 0.76037 0.17324 4.389 1.14e-05 \*\*\***

**duration\_new 4.48560 0.06975 64.312 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for binomial family taken to be 1)**

**Null deviance: 26104 on 36167 degrees of freedom**

**Residual deviance: 17487 on 36139 degrees of freedom**

**AIC: 17545**

**Number of Fisher Scoring iterations: 6**

**MODEL 2**

1. Outlier treatment on duration done
2. Outlier treatment on balance done
3. Biased sampling done using SMOTE and Caret

library(caret)

set.seed(1234)

library(caTools)

library(ade4)

library(dplyr)

library(data.table)

df <- read.table("df\_final\_ish.csv", sep=",", stringsAsFactors = FALSE,header = TRUE)

head(df)

df['zero\_previous'] <- ifelse(df$previous == 0,1,0)

head(df)

#-1 is new customer

df['pdays'] <- ifelse(df$pdays == -1,0,1)

head(df)

df\_new1 <- subset(df,select = -c(day,month,pdays,job\_unknown,education\_unknown,contact\_unknown,poutcome\_unknown,marital\_single,X,zero\_previous) )

summary(df\_new1$duration)

duration = df\_new1$balance

quantile(duration, c(.99)) #13164.9

duration = df\_new1$balance

quantile(duration, c(.01))

#outler treatment for balance

df\_new1['balance\_new']<- ifelse(df\_new1$balance > 13164.9,13164.9,df\_new1$balance)

df\_new1['duration\_new']<- ifelse(df\_new1$duration > 751,751,df\_new1$duration)

df\_new2 <- subset(df\_new1,select = -c(balance,duration) )

normalize <- function(x){

return ((x - min(x))/(max(x) - min(x)))

}

data <- sapply(df\_new2,normalize)

head(data)

class(data)

library(caret)

set.seed(1234)

sample <- sample.int(n = nrow(data), size = floor(.80\*nrow(data)), replace = F)

train <- data[sample, ]

test <- data[-sample, ]

train1 <- as.data.frame(train)

class(train1)

test1 <- as.data.frame(test)

class(test1)

# write.csv(train1, "train1\_smoted.csv")

# write.csv(test1, "test1\_smoted.csv")

prop.table(table(train1$y\_yes))

## 0 1

#0.8813147 0.1186853

prop.table(table(test1$y\_yes))

## 0 1

##0.8847158 0.1152842

dim(train1)

#22606 30

dim(train1)

library(DMwR)

train1$y\_yes <- as.factor(train1$y\_yes)

train1 <- SMOTE(y\_yes ~ ., train1, perc.over = 100, perc.under=200)

model <- glm (y\_yes ~ ., data = train1, family = binomial)

summary(model)

predict <- predict(model, type = 'response')

#confusion matrix

table(train1$y\_yes, predict > 0.5)

pred <- predict(model, newdata = train1, type = "response")

hist(pred)

y\_pred\_num <- ifelse(pred > 0.5, 1, 0)

y\_pred <- factor(y\_pred\_num, levels=c(0, 1))

y\_act <- train1$y\_yes

y\_act

table(y\_pred\_num,y\_act)

mean(y\_pred == y\_act)

y\_pred

MODEL

**Call:**

**glm(formula = y\_yes ~ ., family = binomial, data = train1)**

**Deviance Residuals:**

**Min 1Q Median 3Q Max**

**-3.5717 -0.5865 0.0105 0.5976 2.8031**

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

**(Intercept) -4.04855 0.31405 -12.891 < 2e-16 \*\*\***

**age -0.13713 0.20698 -0.663 0.507631**

**job\_admin. 0.96743 0.29119 3.322 0.000893 \*\*\***

**job\_blue.collar 0.21179 0.28980 0.731 0.464879**

**job\_entrepreneur 0.05102 0.31248 0.163 0.870314**

**job\_housemaid 0.30744 0.31387 0.980 0.327327**

**job\_management 0.58378 0.28952 2.016 0.043763 \***

**job\_retired 1.31117 0.29691 4.416 1.01e-05 \*\*\***

**job\_self.employed 0.21787 0.30778 0.708 0.479017**

**job\_services 0.38281 0.29535 1.296 0.194934**

**job\_student 1.42967 0.30890 4.628 3.69e-06 \*\*\***

**job\_technician 0.56616 0.28953 1.955 0.050529 .**

**job\_unemployed 0.45735 0.30977 1.476 0.139832**

**marital\_divorced -0.13921 0.08116 -1.715 0.086290 .**

**marital\_married -0.21850 0.05525 -3.955 7.66e-05 \*\*\***

**education\_primary -0.19375 0.12530 -1.546 0.122034**

**education\_secondary 0.05301 0.11149 0.475 0.634457**

**education\_tertiary 0.32745 0.11744 2.788 0.005300 \*\***

**default\_yes -0.44760 0.19726 -2.269 0.023267 \***

**housing\_yes -0.95485 0.04846 -19.703 < 2e-16 \*\*\***

**loan\_yes -0.83070 0.06973 -11.913 < 2e-16 \*\*\***

**contact\_cellular 1.37892 0.06565 21.003 < 2e-16 \*\*\***

**contact\_telephone 1.57217 0.10536 14.922 < 2e-16 \*\*\***

**campaign -10.16670 0.75390 -13.486 < 2e-16 \*\*\***

**previous 5.80311 3.47204 1.671 0.094646 .**

**poutcome\_failure 0.27649 0.07711 3.586 0.000336 \*\*\***

**poutcome\_other 0.57060 0.11387 5.011 5.42e-07 \*\*\***

**poutcome\_success 2.73290 0.11140 24.532 < 2e-16 \*\*\***

**balance\_new 0.87556 0.20765 4.216 2.48e-05 \*\*\***

**duration\_new 5.72622 0.09362 61.166 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for binomial family taken to be 1)**

**Null deviance: 23567 on 16999 degrees of freedom**

**Residual deviance: 13608 on 16970 degrees of freedom**

**AIC: 13668**

**Number of Fisher Scoring iterations: 5**

**ACCURACY:**

**0.8227059**

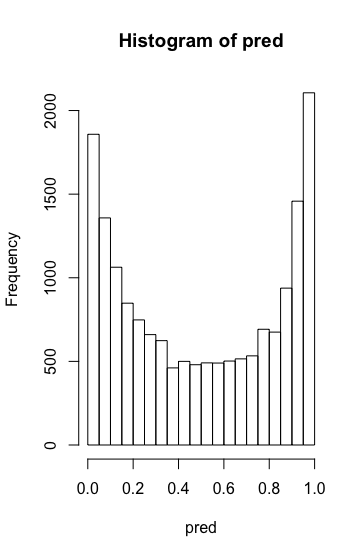
**CONFUSION MATRIX**

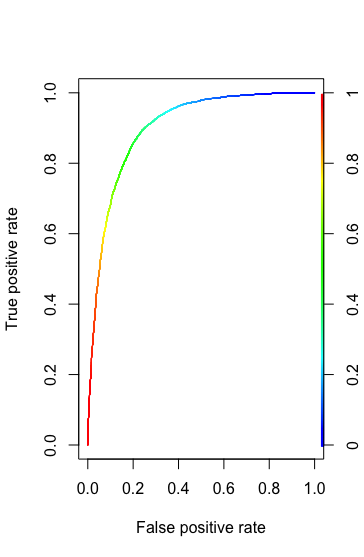
**y\_act**

**y\_pred\_num 0 1**

**0 7043 1557**

**1 1457 6943**

****

****