> ##Reading the dataset

> german\_credit<-read.csv("german.csv", header=TRUE)

> library(rattle)

> names(german\_credit)<-normVarNames(names(german\_credit))

>

> ##Structure of dataframe

> str(german\_credit)

'data.frame': 1000 obs. of 21 variables:

$ status\_of\_existing\_checking\_account : Factor w/ 4 levels "A11","A12","A13",..: 1 2 4 1 1 4 4 2 4 2 ...

$ duration\_in\_month : int 6 48 12 42 24 36 24 36 12 30 ...

$ credit\_history : Factor w/ 5 levels "A30","A31","A32",..: 5 3 5 3 4 3 3 3 3 5 ...

$ purpose : Factor w/ 10 levels "A40","A41","A410",..: 5 5 8 4 1 8 4 2 5 1 ...

$ credit\_amount : int 1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...

$ savings\_account\_bonds : Factor w/ 5 levels "A61","A62","A63",..: 5 1 1 1 1 5 3 1 4 1 ...

$ present\_employment\_since : Factor w/ 5 levels "A71","A72","A73",..: 5 3 4 4 3 3 5 3 4 1 ...

$ installment\_rate\_in\_percentage\_of\_disposable\_income : int 4 2 2 2 3 2 3 2 2 4 ...

$ personal\_status\_and\_sex : Factor w/ 4 levels "A91","A92","A93",..: 3 2 3 3 3 3 3 3 1 4 ...

$ other\_debtors\_guarantors : Factor w/ 3 levels "A101","A102",..: 1 1 1 3 1 1 1 1 1 1 ...

$ present\_residence\_since : int 4 2 3 4 4 4 4 2 4 2 ...

$ property : Factor w/ 4 levels "A121","A122",..: 1 1 1 2 4 4 2 3 1 3 ...

$ age\_in\_years : int 67 22 49 45 53 35 53 35 61 28 ...

$ other\_installment\_plans : Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3 ...

$ housing : Factor w/ 3 levels "A151","A152",..: 2 2 2 3 3 3 2 1 2 2 ...

$ number\_of\_existing\_credits\_at\_this\_bank : int 2 1 1 1 2 1 1 1 1 2 ...

$ job\_status : Factor w/ 4 levels "A171","A172",..: 3 3 2 3 3 2 3 4 2 4 ...

$ number\_of\_people\_being\_liable\_to\_provide\_maintenance\_for: int 1 1 2 2 2 2 1 1 1 1 ...

$ telephone : Factor w/ 2 levels "A191","A192": 2 1 1 1 1 2 1 2 1 1 ...

$ foreign\_worker : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...

$ default\_status : int 0 1 0 0 1 0 0 0 0 1 ...

>

> ##Data exploration

> summary(german\_credit)

status\_of\_existing\_checking\_account duration\_in\_month credit\_history purpose credit\_amount

A11:274 Min. : 4.0 A30: 40 A43 :280 Min. : 250

A12:269 1st Qu.:12.0 A31: 49 A40 :234 1st Qu.: 1366

A13: 63 Median :18.0 A32:530 A42 :181 Median : 2320

A14:394 Mean :20.9 A33: 88 A41 :103 Mean : 3271

3rd Qu.:24.0 A34:293 A49 : 97 3rd Qu.: 3972

Max. :72.0 A46 : 50 Max. :18424

(Other): 55

savings\_account\_bonds present\_employment\_since installment\_rate\_in\_percentage\_of\_disposable\_income

A61:603 A71: 62 Min. :1.000

A62:103 A72:172 1st Qu.:2.000

A63: 63 A73:339 Median :3.000

A64: 48 A74:174 Mean :2.973

A65:183 A75:253 3rd Qu.:4.000

Max. :4.000

personal\_status\_and\_sex other\_debtors\_guarantors present\_residence\_since property

A91: 50 A101:907 Min. :1.000 A121:282

A92:310 A102: 41 1st Qu.:2.000 A122:232

A93:548 A103: 52 Median :3.000 A123:332

A94: 92 Mean :2.845 A124:154

3rd Qu.:4.000

Max. :4.000

age\_in\_years other\_installment\_plans housing number\_of\_existing\_credits\_at\_this\_bank

Min. :19.00 A141:139 A151:179 Min. :1.000

1st Qu.:27.00 A142: 47 A152:713 1st Qu.:1.000

Median :33.00 A143:814 A153:108 Median :1.000

Mean :35.55 Mean :1.407

3rd Qu.:42.00 3rd Qu.:2.000

Max. :75.00 Max. :4.000

job\_status number\_of\_people\_being\_liable\_to\_provide\_maintenance\_for telephone foreign\_worker

A171: 22 Min. :1.000 A191:596 A201:963

A172:200 1st Qu.:1.000 A192:404 A202: 37

A173:630 Median :1.000

A174:148 Mean :1.155

3rd Qu.:1.000

Max. :2.000

default\_status

Min. :0.0

1st Qu.:0.0

Median :0.0

Mean :0.3

3rd Qu.:1.0

Max. :1.0

>

> credit\_integer<-subset(german\_credit,select = c( default\_status,

+ duration\_in\_month,

+ credit\_amount,

+ installment\_rate\_in\_percentage\_of\_disposable\_income,

+ present\_residence\_since,

+ age\_in\_years,

+ number\_of\_existing\_credits\_at\_this\_bank,

+ number\_of\_people\_being\_liable\_to\_provide\_maintenance\_for

+ ))

>

>

> names<-names(credit\_integer)

> names<-as.list(names)

> ##Descriptive statistics

> descriptive <- function(x,y)

+ {

+ hist(x, col="red", main="Histogram ")

+ boxplot(x, col="yellow", main="Boxplot")

+ qqnorm(x, col="green", main= "QQ plot")

+ plot(density(x), col="blue", main="Density plot")

+ title(y, outer=TRUE)

+ }

>

> par(mfrow=c(2,2), oma=c(0,0,1,0))

> mapply(descriptive,credit\_integer,names)

$default\_status

NULL

$duration\_in\_month

NULL

$credit\_amount

NULL

$installment\_rate\_in\_percentage\_of\_disposable\_income

NULL

$present\_residence\_since

NULL

$age\_in\_years

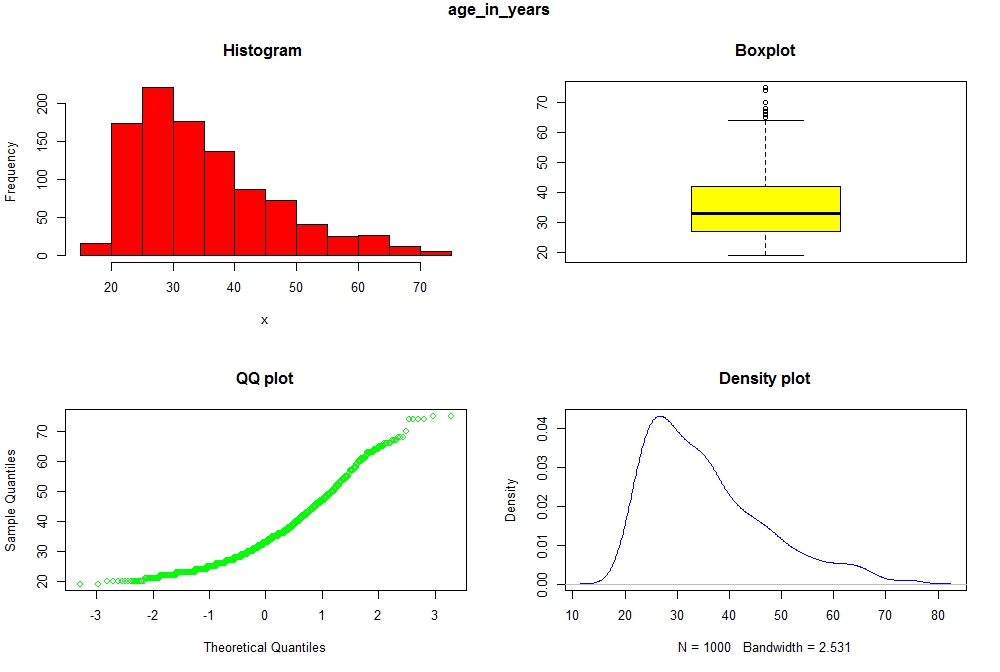
NULL

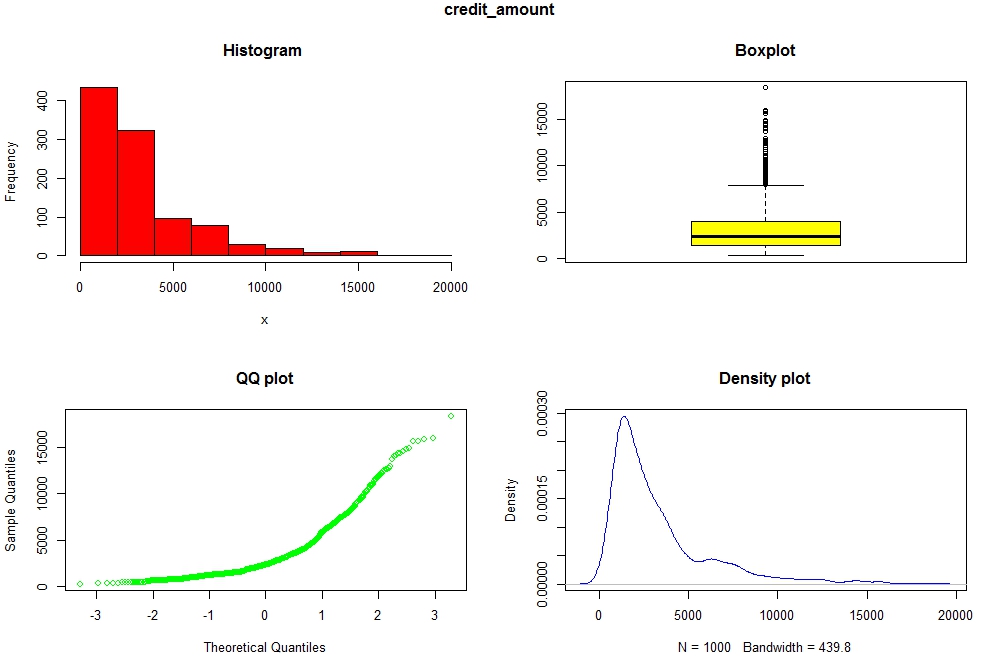
$number\_of\_existing\_credits\_at\_this\_bank

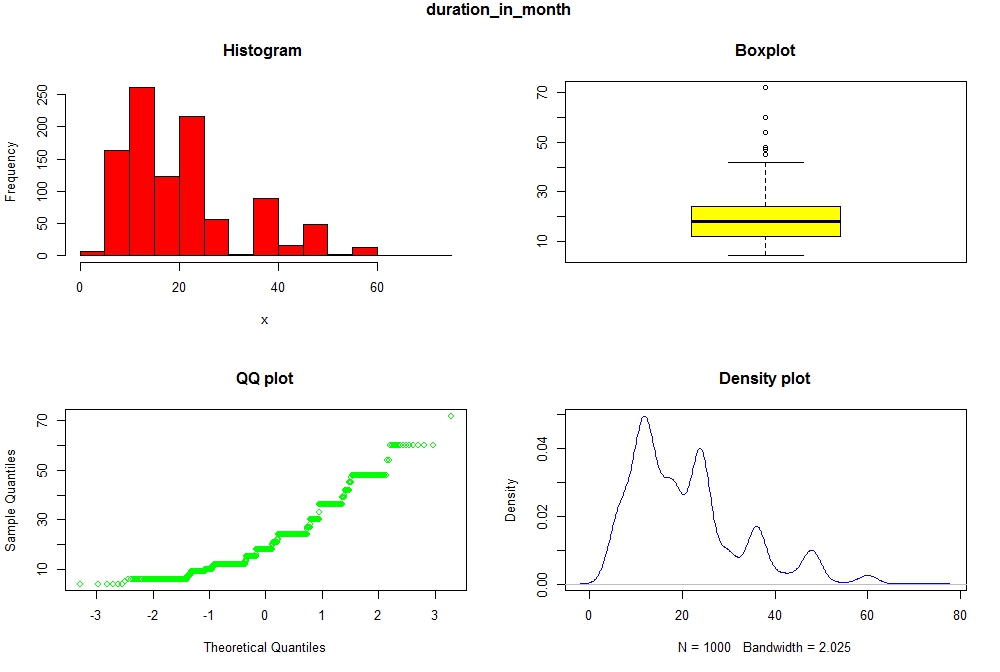
NULL

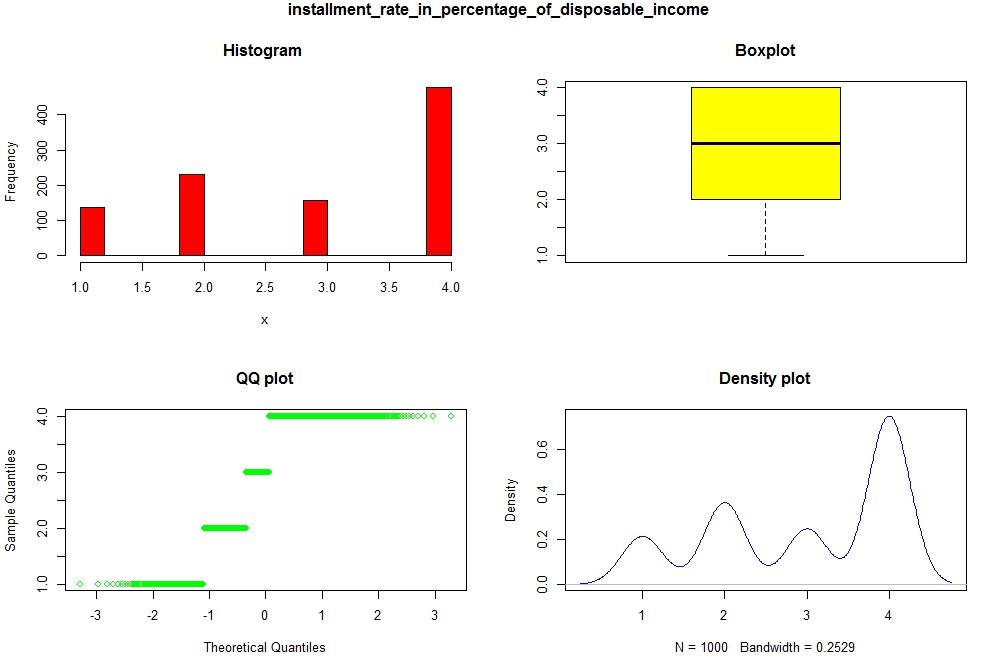
$number\_of\_people\_being\_liable\_to\_provide\_maintenance\_for

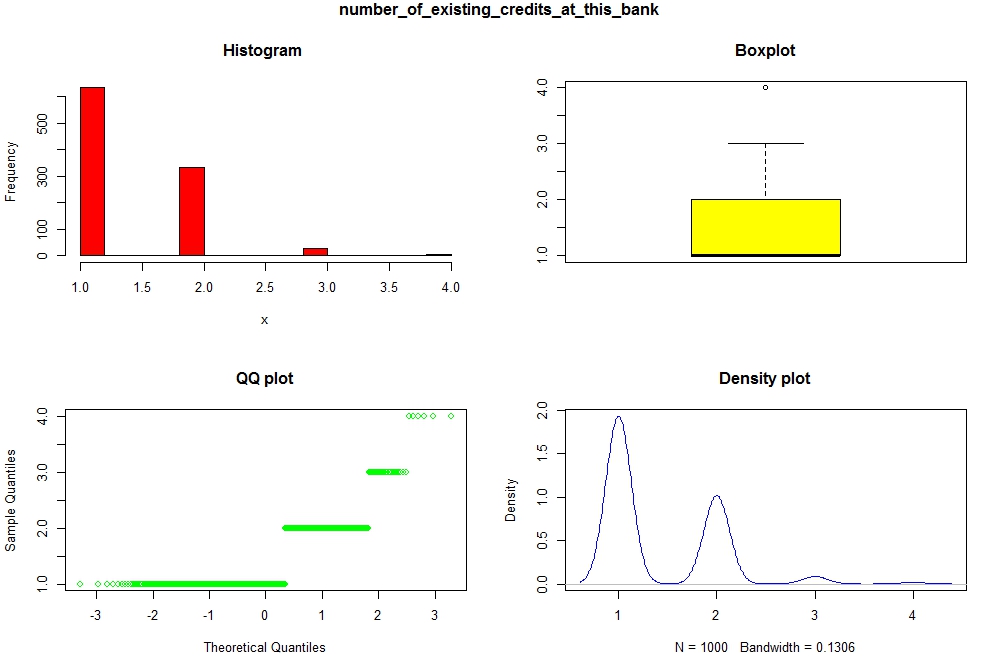
NULL

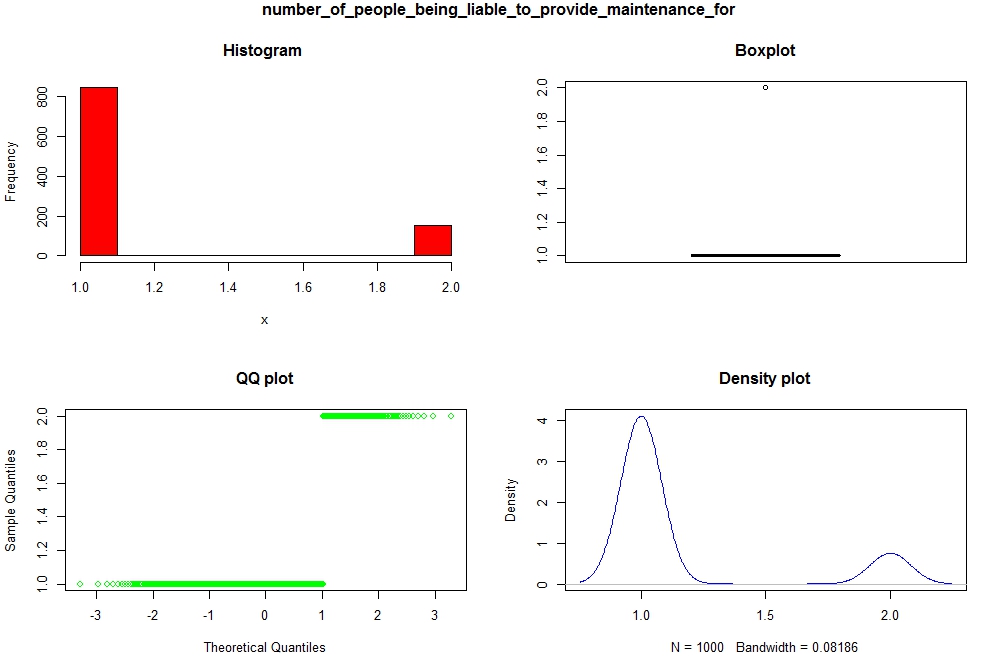


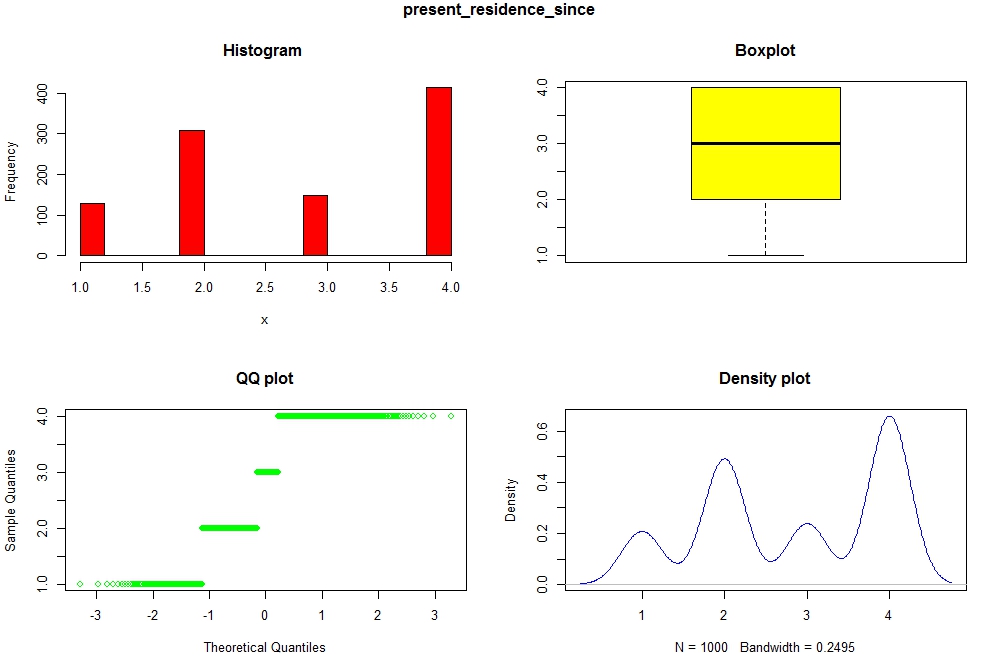












> ##Dummy variable

> dummy\_eca<-model.matrix(~status\_of\_existing\_checking\_account, german\_credit)[,-1]

> dummy\_ch<-model.matrix(~credit\_history, german\_credit)[,-1]

> dummy\_p<-model.matrix(~purpose, german\_credit)[,-1]

> dummy\_sab<-model.matrix(~savings\_account\_bonds, german\_credit)[,-1]

> dummy\_pms<-model.matrix(~present\_employment\_since, german\_credit)[,-1]

> dummy\_psas<-model.matrix(~personal\_status\_and\_sex, german\_credit)[,-1]

> dummy\_odg<-model.matrix(~other\_debtors\_guarantors, german\_credit)[,-1]

> dummy\_ppt<-model.matrix(~property, german\_credit)[,-1]

> dummy\_oip<-model.matrix(~other\_installment\_plans, german\_credit)[,-1]

> dummy\_h<-model.matrix(~housing, german\_credit)[,-1]

> dummy\_js<-model.matrix(~job\_status, german\_credit)[,-1]

> dummy\_fw<-model.matrix(~foreign\_worker, german\_credit)[,-1]

> dummy\_npbl<-model.matrix(~number\_of\_people\_being\_liable\_to\_provide\_maintenance\_for, german\_credit)[,-1]

>

> data<-cbind(credit\_integer,dummy\_eca,dummy\_ch,

+ dummy\_p, dummy\_sab,dummy\_pms, dummy\_psas,dummy\_odg, dummy\_ppt,

+ dummy\_oip, dummy\_h, dummy\_js,dummy\_fw, dummy\_npbl )

> ##Data split

> library(caret)

> set.seed(100)

>

> train<-createDataPartition(y=data$default\_status, p = .70,list = FALSE)

> training<-data[train,]

> testing<-data[-train,]

> model<-glm(default\_status ~ ., data=training ,family=binomial(link='logit'))

> summary(model)

Call:

glm(formula = default\_status ~ ., family = binomial(link = "logit"),

data = training)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5236 -0.7253 -0.3737 0.7383 2.6140

Coefficients: (1 not defined because of singularities)

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

(Intercept) -3.058e-01 1.330e+00 -0.230 0.81808

duration\_in\_month 2.924e-02 1.101e-02 2.655 0.00793 \*\*

credit\_amount 1.363e-04 5.187e-05 2.627 0.00861 \*\*

installment\_rate\_in\_percentage\_of\_disposable\_income 3.190e-01 1.048e-01 3.043 0.00234 \*\*

present\_residence\_since -3.098e-02 1.034e-01 -0.299 0.76456

age\_in\_years -8.570e-03 1.118e-02 -0.766 0.44349

number\_of\_existing\_credits\_at\_this\_bank 1.018e-01 2.294e-01 0.444 0.65721

number\_of\_people\_being\_liable\_to\_provide\_maintenance\_for 3.670e-01 2.943e-01 1.247 0.21237

status\_of\_existing\_checking\_accountA12 -5.145e-01 2.635e-01 -1.953 0.05088 .

status\_of\_existing\_checking\_accountA13 -1.004e+00 4.300e-01 -2.335 0.01954 \*

status\_of\_existing\_checking\_accountA14 -1.742e+00 2.741e-01 -6.355 2.08e-10 \*\*\*

credit\_historyA31 3.858e-01 7.084e-01 0.545 0.58604

credit\_historyA32 -1.788e-02 5.762e-01 -0.031 0.97524

credit\_historyA33 -3.134e-01 6.297e-01 -0.498 0.61870

credit\_historyA34 -1.132e+00 5.901e-01 -1.919 0.05503 .

purposeA41 -9.287e-01 4.482e-01 -2.072 0.03829 \*

purposeA410 -1.280e+00 8.670e-01 -1.476 0.13985

purposeA42 -5.128e-01 3.021e-01 -1.697 0.08964 .

purposeA43 -7.589e-01 2.982e-01 -2.545 0.01093 \*

purposeA44 -5.155e-01 7.979e-01 -0.646 0.51823

purposeA45 1.366e-01 6.885e-01 0.198 0.84271

purposeA46 6.471e-01 4.668e-01 1.386 0.16568

purposeA48 -7.921e-01 1.234e+00 -0.642 0.52096

purposeA49 -3.104e-01 4.017e-01 -0.773 0.43966

savings\_account\_bondsA62 -4.730e-01 3.440e-01 -1.375 0.16911

savings\_account\_bondsA63 -3.668e-01 4.707e-01 -0.779 0.43583

savings\_account\_bondsA64 -1.038e+00 5.742e-01 -1.807 0.07073 .

savings\_account\_bondsA65 -7.608e-01 3.099e-01 -2.455 0.01408 \*

present\_employment\_sinceA72 -7.125e-01 5.219e-01 -1.365 0.17226

present\_employment\_sinceA73 -5.367e-01 5.029e-01 -1.067 0.28590

present\_employment\_sinceA74 -1.320e+00 5.397e-01 -2.446 0.01443 \*

present\_employment\_sinceA75 -7.163e-01 5.039e-01 -1.422 0.15517

personal\_status\_and\_sexA92 -8.834e-03 4.784e-01 -0.018 0.98527

personal\_status\_and\_sexA93 -7.635e-01 4.737e-01 -1.612 0.10702

personal\_status\_and\_sexA94 -4.579e-01 5.574e-01 -0.821 0.41140

other\_debtors\_guarantorsA102 9.098e-01 4.776e-01 1.905 0.05678 .

other\_debtors\_guarantorsA103 -1.286e+00 5.499e-01 -2.339 0.01932 \*

propertyA122 3.423e-01 3.052e-01 1.121 0.26211

propertyA123 4.380e-01 2.817e-01 1.555 0.11994

propertyA124 9.518e-01 5.000e-01 1.904 0.05694 .

other\_installment\_plansA142 -3.154e-01 4.804e-01 -0.657 0.51147

other\_installment\_plansA143 -7.070e-01 2.786e-01 -2.537 0.01117 \*

housingA152 -3.154e-01 2.768e-01 -1.139 0.25452

housingA153 -7.393e-01 5.597e-01 -1.321 0.18657

job\_statusA172 1.034e+00 8.172e-01 1.265 0.20582

job\_statusA173 6.924e-01 7.872e-01 0.880 0.37910

job\_statusA174 3.115e-01 7.660e-01 0.407 0.68426

dummy\_fw -7.170e-01 8.162e-01 -0.878 0.37972

dummy\_npbl NA NA NA NA

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 883.54 on 699 degrees of freedom

Residual deviance: 643.32 on 652 degrees of freedom

AIC: 739.32

Number of Fisher Scoring iterations: 5

> model\_F<-glm(formula = default\_status ~ duration\_in\_month + credit\_amount +

+ installment\_rate\_in\_percentage\_of\_disposable\_income + status\_of\_existing\_checking\_accountA12 +

+ status\_of\_existing\_checking\_accountA13 + status\_of\_existing\_checking\_accountA14 +

+ credit\_historyA34 + purposeA43 +

+ purposeA46 + savings\_account\_bondsA65 +

+ present\_employment\_sinceA74 + personal\_status\_and\_sexA93 +

+ other\_debtors\_guarantorsA102 +

+ other\_debtors\_guarantorsA103 + other\_installment\_plansA143,

+ family = binomial(link = "logit"), data = training)

>

> summary(model\_F)

Call:

glm(formula = default\_status ~ duration\_in\_month + credit\_amount +

installment\_rate\_in\_percentage\_of\_disposable\_income + status\_of\_existing\_checking\_accountA12 +

status\_of\_existing\_checking\_accountA13 + status\_of\_existing\_checking\_accountA14 +

credit\_historyA34 + purposeA43 + purposeA46 + savings\_account\_bondsA65 +

present\_employment\_sinceA74 + personal\_status\_and\_sexA93 +

other\_debtors\_guarantorsA102 + other\_debtors\_guarantorsA103 +

other\_installment\_plansA143, family = binomial(link = "logit"),

data = training)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.1517 -0.7694 -0.4035 0.8071 2.7863

Coefficients:

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

(Intercept) -2.875e-01 4.290e-01 -0.670 0.50273

duration\_in\_month 3.049e-02 1.009e-02 3.021 0.00252 \*\*

credit\_amount 1.124e-04 4.633e-05 2.425 0.01530 \*

installment\_rate\_in\_percentage\_of\_disposable\_income 2.633e-01 9.737e-02 2.704 0.00686 \*\*

status\_of\_existing\_checking\_accountA12 -6.638e-01 2.363e-01 -2.809 0.00497 \*\*

status\_of\_existing\_checking\_accountA13 -1.160e+00 4.079e-01 -2.843 0.00447 \*\*

status\_of\_existing\_checking\_accountA14 -1.905e+00 2.488e-01 -7.656 1.93e-14 \*\*\*

credit\_historyA34 -1.087e+00 2.388e-01 -4.552 5.33e-06 \*\*\*

purposeA43 -4.900e-01 2.342e-01 -2.092 0.03645 \*

purposeA46 1.076e+00 4.163e-01 2.585 0.00975 \*\*

savings\_account\_bondsA65 -6.662e-01 2.802e-01 -2.377 0.01744 \*

present\_employment\_sinceA74 -6.388e-01 2.766e-01 -2.309 0.02094 \*

personal\_status\_and\_sexA93 -6.162e-01 2.009e-01 -3.068 0.00216 \*\*

other\_debtors\_guarantorsA102 9.089e-01 4.398e-01 2.067 0.03875 \*

other\_debtors\_guarantorsA103 -1.212e+00 5.091e-01 -2.380 0.01729 \*

other\_installment\_plansA143 -6.790e-01 2.299e-01 -2.954 0.00314 \*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 883.54 on 699 degrees of freedom

Residual deviance: 674.25 on 684 degrees of freedom

AIC: 706.25

Number of Fisher Scoring iterations: 5

> vif(model\_F)

duration\_in\_month

1.648814

credit\_amount

1.913026

installment\_rate\_in\_percentage\_of\_disposable\_income

1.234368

status\_of\_existing\_checking\_accountA12

1.325187

status\_of\_existing\_checking\_accountA13

1.136372

status\_of\_existing\_checking\_accountA14

1.336276

credit\_historyA34

1.049571

purposeA43

1.074941

purposeA46

1.052849

savings\_account\_bondsA65

1.040355

present\_employment\_sinceA74

1.046207

personal\_status\_and\_sexA93

1.121364

other\_debtors\_guarantorsA102

1.022818

other\_debtors\_guarantorsA103

1.037933

other\_installment\_plansA143

1.021659

> ##Predicting

>

> training$fit<-predict(model\_F, training, type= 'response')

> testing$fit<-predict(model\_F, testing, type= 'response')

>

> training$predscore<-ifelse(training$fit>0.5,1,0)

> testing$predscore<-ifelse(testing$fit>0.5,1,0)

> ##Model Diagnostics

> ##HL test

> ##install.packages("Hmisc")

> library(Hmisc)

> omers2(fitted(model\_F),training$default\_status)

Error: could not find function "omers2"

> par(mfrow=c(1,1))

> plot(residuals(model\_F,type="pearson"), main="Pearson Residual Plot")

>

> ##install.packages("ResourceSelection")

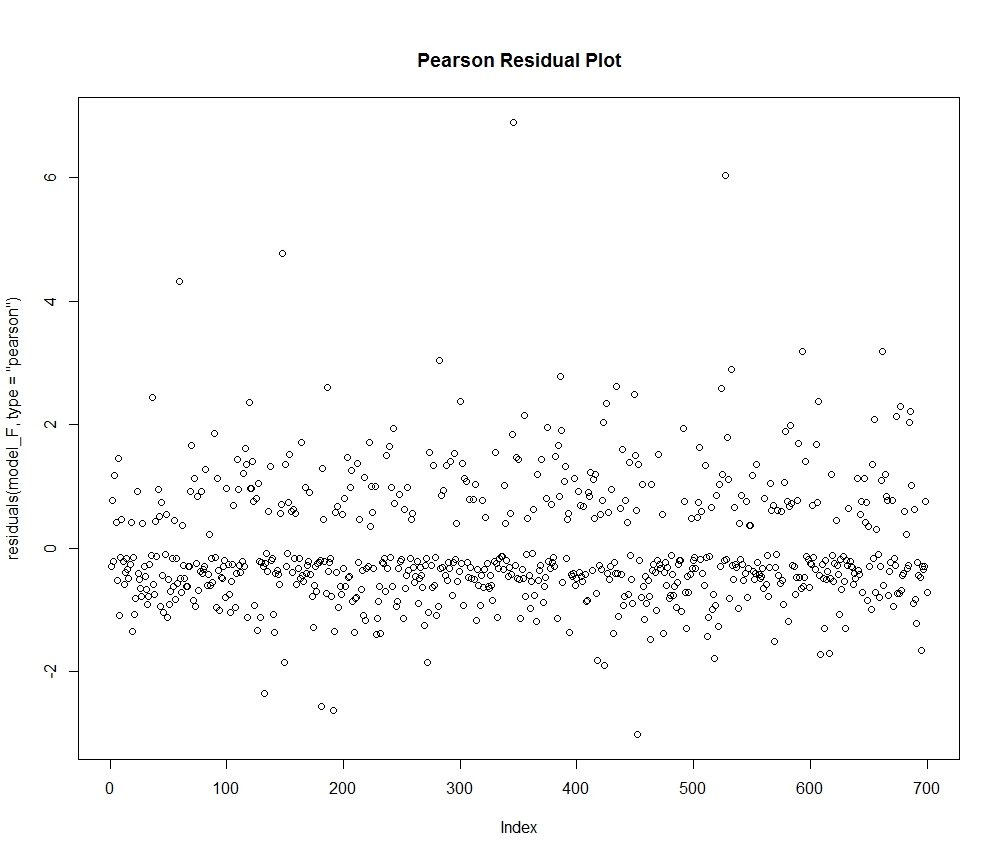
> library(ResourceSelection)

> hoslem.test(training$default\_status, fitted(model\_F))

Hosmer and Lemeshow goodness of fit (GOF) test

data: training$default\_status, fitted(model\_F)

X-squared = 9.5452, df = 8, p-value = 0.2984



> ##Crosstable

> ##install.packages("gmodels")

> library(gmodels)

> ##Training

> CrossTable(training$default\_status, training$predscore,prop.chisq = FALSE,

+ prop.c=FALSE, prop.r=FALSE, dnn=c('actual', 'predected'))

Cell Contents

|-------------------------|

| N |

| N / Table Total |

|-------------------------|

Total Observations in Table: 700

| predected

actual | 0 | 1 | Row Total |

-------------|-----------|-----------|-----------|

0 | 413 | 59 | 472 |

| 0.590 | 0.084 | |

-------------|-----------|-----------|-----------|

1 | 107 | 121 | 228 |

| 0.153 | 0.173 | |

-------------|-----------|-----------|-----------|

Column Total | 520 | 180 | 700 |

-------------|-----------|-----------|-----------|

> ##Testing

> CrossTable(testing$default\_status, testing$predscore,prop.chisq = FALSE,

+ prop.c=FALSE, prop.r=FALSE, dnn=c('actual', 'predected'))

Cell Contents

|-------------------------|

| N |

| N / Table Total |

|-------------------------|

Total Observations in Table: 300

| predected

actual | 0 | 1 | Row Total |

-------------|-----------|-----------|-----------|

0 | 194 | 34 | 228 |

| 0.647 | 0.113 | |

-------------|-----------|-----------|-----------|

1 | 36 | 36 | 72 |

| 0.120 | 0.120 | |

-------------|-----------|-----------|-----------|

Column Total | 230 | 70 | 300 |

-------------|-----------|-----------|-----------|

> library(ROCR)

> ##ROC Curve

> ##install.packages("ROCR")

> library(ROCR)

> ##Training

> predTraining<-prediction(training$default\_status, training$predscore)

> perfTraining<-performance(predTraining,"tpr", "fpr")

> plot(perfTraining, main="ROC Curve", col = 2,lwd = 2)

> abline(a = 0,b = 1,lwd = 2,lty = 3,col = "black")

> aucTraining<- performance(predTraining,"auc")

> aucTraining

An object of class "performance"

Slot "x.name":

[1] "None"

Slot "y.name":

[1] "Area under the ROC curve"

Slot "alpha.name":

[1] "none"

Slot "x.values":

list()

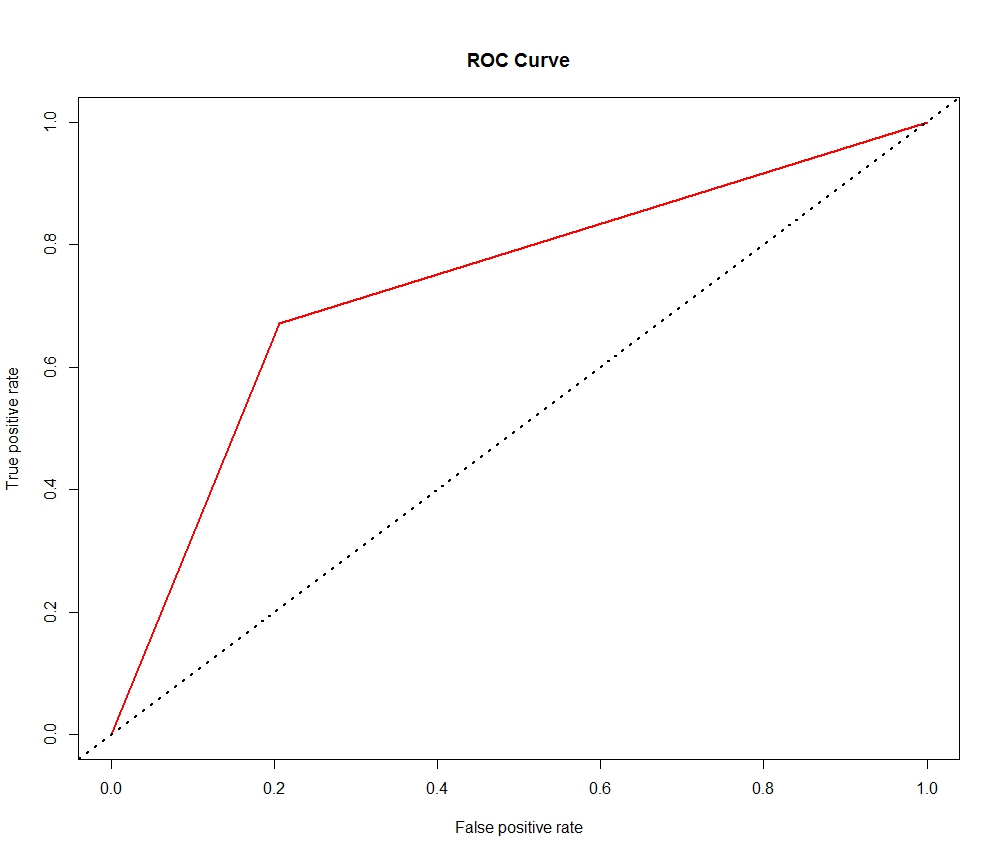
Slot "y.values":

[[1]]

[1] 0.7332265

Slot "alpha.values":

list()



> ##Testing

> predTesting<-prediction(testing$default\_status, testing$predscore)

> perfTesting<-performance(predTesting,"tpr", "fpr")

> plot(perfTesting, main="ROC Curve", col = 2,lwd = 2)

> abline(a = 0,b = 1,lwd = 2,lty = 3,col = "black")

> aucTesting<- performance(predTesting,"auc")

> aucTesting

An object of class "performance"

Slot "x.name":

[1] "None"

Slot "y.name":

[1] "Area under the ROC curve"

Slot "alpha.name":

[1] "none"

Slot "x.values":

list()

Slot "y.values":

[[1]]

[1] 0.678882

Slot "alpha.values":

list()

