1. Output variable -> y

y -> Whether the client has subscribed a term deposit or not

Binomial ("yes" or "no")

Answer ->

**The Code with IMP highlight and explanation - >**

> #logistic regression assignment 1

>

> myMatrix <- read.table(file.choose(), sep=";",header=TRUE)

> mym <- myMatrix

> library('mlr')

Loading required package: ParamHelpers

'mlr' is in maintenance mode since July 2019. Future development efforts will go into its

successor 'mlr3' (<https://mlr3.mlr-org.com>).

> sum(is.na(myMatrix))

[1] 0

**>#Creating Dummy variables**

> mymm<-createDummyFeatures(mym, cols = c("job","marital","education","contact","month","poutcome"))

>

**> #making yes and no as 1 and 0**

> mymm$default<-(factor(as.numeric(mymm$default)-1))

> mymm$housing<-(factor(as.numeric(mymm$housing)-1))

> mymm$y<-(factor(as.numeric(mymm$y)-1))

> mymm$loan<-(factor(as.numeric(mymm$loan)-1))

>

**> #making the new dummy variable in factor form using loop**

> for(i in 12:49)

+ {

+ mymm[,i]<-factor(mymm[,i])

+ }

>

>

> summary(mymm)

age default balance housing loan day duration

Min. :18.00 0:44396 Min. : -8019 0:20081 0:37967 Min. : 1.00 Min. : 0.0

1st Qu.:33.00 1: 815 1st Qu.: 72 1:25130 1: 7244 1st Qu.: 8.00 1st Qu.: 103.0

Median :39.00 Median : 448 Median :16.00 Median : 180.0

Mean :40.94 Mean : 1362 Mean :15.81 Mean : 258.2

3rd Qu.:48.00 3rd Qu.: 1428 3rd Qu.:21.00 3rd Qu.: 319.0

Max. :95.00 Max. :102127 Max. :31.00 Max. :4918.0

campaign pdays previous y job.admin. job.blue.collar

Min. : 1.000 Min. : -1.0 Min. : 0.0000 0:39922 0:40040 0:35479

1st Qu.: 1.000 1st Qu.: -1.0 1st Qu.: 0.0000 1: 5289 1: 5171 1: 9732

Median : 2.000 Median : -1.0 Median : 0.0000

Mean : 2.764 Mean : 40.2 Mean : 0.5803

3rd Qu.: 3.000 3rd Qu.: -1.0 3rd Qu.: 0.0000

Max. :63.000 Max. :871.0 Max. :275.0000

job.entrepreneur job.housemaid job.management job.retired job.self.employed job.services job.student

0:43724 0:43971 0:35753 0:42947 0:43632 0:41057 0:44273

1: 1487 1: 1240 1: 9458 1: 2264 1: 1579 1: 4154 1: 938

job.technician job.unemployed job.unknown marital.divorced marital.married marital.single

0:37614 0:43908 0:44923 0:40004 0:17997 0:32421

1: 7597 1: 1303 1: 288 1: 5207 1:27214 1:12790

education.primary education.secondary education.tertiary education.unknown contact.cellular

0:38360 0:22009 0:31910 0:43354 0:15926

1: 6851 1:23202 1:13301 1: 1857 1:29285

contact.telephone contact.unknown month.apr month.aug month.dec month.feb month.jan month.jul

0:42305 0:32191 0:42279 0:38964 0:44997 0:42562 0:43808 0:38316

1: 2906 1:13020 1: 2932 1: 6247 1: 214 1: 2649 1: 1403 1: 6895

month.jun month.mar month.may month.nov month.oct month.sep poutcome.failure poutcome.other

0:39870 0:44734 0:31445 0:41241 0:44473 0:44632 0:40310 0:43371

1: 5341 1: 477 1:13766 1: 3970 1: 738 1: 579 1: 4901 1: 1840

poutcome.success poutcome.unknown

0:43700 0: 8252

1: 1511 1:36959

> str(mymm)

'data.frame': 45211 obs. of 49 variables:

$ age : int 58 44 33 47 33 35 28 42 58 43 ...

$ default : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...

$ balance : int 2143 29 2 1506 1 231 447 2 121 593 ...

$ housing : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 2 2 ...

$ loan : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 1 1 ...

$ day : int 5 5 5 5 5 5 5 5 5 5 ...

$ duration : int 261 151 76 92 198 139 217 380 50 55 ...

$ campaign : int 1 1 1 1 1 1 1 1 1 1 ...

$ pdays : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...

$ previous : int 0 0 0 0 0 0 0 0 0 0 ...

$ y : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ job.admin. : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ job.blue.collar : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...

$ job.entrepreneur : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 2 1 1 ...

$ job.housemaid : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ job.management : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 2 1 1 1 ...

$ job.retired : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...

$ job.self.employed : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ job.services : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ job.student : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ job.technician : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 2 ...

$ job.unemployed : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ job.unknown : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...

$ marital.divorced : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...

$ marital.married : Factor w/ 2 levels "0","1": 2 1 2 2 1 2 1 1 2 1 ...

$ marital.single : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 2 1 1 2 ...

$ education.primary : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...

$ education.secondary: Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 1 1 2 ...

$ education.tertiary : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 2 2 1 1 ...

$ education.unknown : Factor w/ 2 levels "0","1": 1 1 1 2 2 1 1 1 1 1 ...

$ contact.cellular : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ contact.telephone : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ contact.unknown : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...

$ month.apr : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.aug : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.dec : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.feb : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.jan : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.jul : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.jun : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.mar : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.may : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...

$ month.nov : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.oct : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ month.sep : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ poutcome.failure : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ poutcome.other : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ poutcome.success : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ poutcome.unknown : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...

> attach(mymm)

> logit <- glm(y~.,family="binomial",data=mymm)

> summary(logit)

Call:

glm(formula = y ~ ., family = "binomial", data = mymm)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.7286 -0.3744 -0.2530 -0.1502 3.4288

Coefficients: (6 not defined because of singularities)

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

(Intercept) -3.347e+00 2.798e-01 -11.963 < 2e-16 \*\*\*

age 1.127e-04 2.205e-03 0.051 0.959233

default1 -1.668e-02 1.628e-01 -0.102 0.918407

balance 1.283e-05 5.148e-06 2.493 0.012651 \*

housing1 -6.754e-01 4.387e-02 -15.395 < 2e-16 \*\*\*

loan1 -4.254e-01 5.999e-02 -7.091 1.33e-12 \*\*\*

day 9.969e-03 2.497e-03 3.993 6.53e-05 \*\*\*

duration 4.194e-03 6.453e-05 64.986 < 2e-16 \*\*\*

campaign -9.078e-02 1.014e-02 -8.955 < 2e-16 \*\*\*

pdays -1.027e-04 3.061e-04 -0.335 0.737268

previous 1.015e-02 6.503e-03 1.561 0.118476

job.admin.1 3.133e-01 2.335e-01 1.342 0.179656

job.blue.collar1 3.392e-03 2.328e-01 0.015 0.988376

job.entrepreneur1 -4.384e-02 2.533e-01 -0.173 0.862619

job.housemaid1 -1.907e-01 2.577e-01 -0.740 0.459187

job.management1 1.480e-01 2.317e-01 0.639 0.523110

job.retired1 5.656e-01 2.373e-01 2.384 0.017138 \*

job.self.employed1 1.493e-02 2.471e-01 0.060 0.951822

job.services1 8.947e-02 2.374e-01 0.377 0.706261

job.student1 6.954e-01 2.452e-01 2.836 0.004570 \*\*

job.technician1 1.372e-01 2.317e-01 0.592 0.553700

job.unemployed1 1.366e-01 2.469e-01 0.553 0.580282

job.unknown1 NA NA NA NA

marital.divorced1 -9.250e-02 6.726e-02 -1.375 0.169066

marital.married1 -2.720e-01 4.594e-02 -5.919 3.23e-09 \*\*\*

marital.single1 NA NA NA NA

education.primary1 -2.505e-01 1.039e-01 -2.411 0.015915 \*

education.secondary1 -6.695e-02 9.124e-02 -0.734 0.463085

education.tertiary1 1.285e-01 9.586e-02 1.340 0.180204

education.unknown1 NA NA NA NA

contact.cellular1 1.623e+00 7.317e-02 22.184 < 2e-16 \*\*\*

contact.telephone1 1.460e+00 1.006e-01 14.508 < 2e-16 \*\*\*

contact.unknown1 NA NA NA NA

month.apr1 -8.741e-01 1.195e-01 -7.314 2.58e-13 \*\*\*

month.aug1 -1.568e+00 1.156e-01 -13.568 < 2e-16 \*\*\*

month.dec1 -1.829e-01 1.950e-01 -0.938 0.348130

month.feb1 -1.021e+00 1.213e-01 -8.419 < 2e-16 \*\*\*

month.jan1 -2.136e+00 1.522e-01 -14.034 < 2e-16 \*\*\*

month.jul1 -1.705e+00 1.187e-01 -14.363 < 2e-16 \*\*\*

month.jun1 -4.204e-01 1.238e-01 -3.395 0.000686 \*\*\*

month.mar1 7.158e-01 1.464e-01 4.889 1.01e-06 \*\*\*

month.may1 -1.273e+00 1.150e-01 -11.075 < 2e-16 \*\*\*

month.nov1 -1.747e+00 1.227e-01 -14.236 < 2e-16 \*\*\*

month.oct1 7.379e-03 1.376e-01 0.054 0.957222

month.sep1 NA NA NA NA

poutcome.failure1 9.179e-02 9.347e-02 0.982 0.326093

poutcome.other1 2.953e-01 1.068e-01 2.766 0.005677 \*\*

poutcome.success1 2.383e+00 8.625e-02 27.627 < 2e-16 \*\*\*

poutcome.unknown1 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: 32631** on 45210 degrees of freedom

**Residual deviance: 21562** on 45168 degrees of freedom

**AIC: 21648**

Number of Fisher Scoring iterations: 6

>

> #null deviace must be higher than Residual deviance

**> #dropping the non significant collumns to improve the model**

>

> logit1 <- glm(y~.-default-pdays-previous-job.admin.-job.unknown-marital.divorced-job.blue.collar-job.entrepreneur-job.housemaid-job.management-job.self.employed-job.services-job.technician-job.unemployed-marital.single-education.secondary-education.tertiary-education.unknown-contact.unknown-month.dec-month.oct-month.sep-poutcome.failure-poutcome.unknown,family = "binomial",data=mymm )

> summary(logit1)

Call:

glm(formula = y ~ . - default - pdays - previous - job.admin. -

job.unknown - marital.divorced - job.blue.collar - job.entrepreneur -

job.housemaid - job.management - job.self.employed - job.services -

job.technician - job.unemployed - marital.single - education.secondary -

education.tertiary - education.unknown - contact.unknown -

month.dec - month.oct - month.sep - poutcome.failure - poutcome.unknown,

family = "binomial", data = mymm)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.6831 -0.3750 -0.2554 -0.1513 3.4236

Coefficients:

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

(Intercept) -3.176e+00 1.379e-01 -23.029 < 2e-16 \*\*\*

age -1.497e-03 2.034e-03 -0.736 0.461952

balance 1.500e-05 5.103e-06 2.940 0.003278 \*\*

housing1 -6.826e-01 4.311e-02 -15.834 < 2e-16 \*\*\*

loan1 -4.388e-01 5.963e-02 -7.358 1.86e-13 \*\*\*

day 1.029e-02 2.478e-03 4.150 3.32e-05 \*\*\*

duration 4.170e-03 6.414e-05 65.012 < 2e-16 \*\*\*

campaign -9.085e-02 1.010e-02 -8.996 < 2e-16 \*\*\*

job.retired1 4.669e-01 8.206e-02 5.690 1.27e-08 \*\*\*

job.student1 5.327e-01 9.687e-02 5.499 3.82e-08 \*\*\*

marital.married1 -2.619e-01 3.909e-02 -6.701 2.08e-11 \*\*\*

education.primary1 -3.395e-01 5.878e-02 -5.777 7.61e-09 \*\*\*

contact.cellular1 1.678e+00 7.158e-02 23.448 < 2e-16 \*\*\*

contact.telephone1 1.502e+00 9.950e-02 15.093 < 2e-16 \*\*\*

month.apr1 -8.900e-01 8.649e-02 -10.291 < 2e-16 \*\*\*

month.aug1 -1.573e+00 8.105e-02 -19.403 < 2e-16 \*\*\*

month.feb1 -1.025e+00 9.138e-02 -11.218 < 2e-16 \*\*\*

month.jan1 -2.150e+00 1.264e-01 -17.011 < 2e-16 \*\*\*

month.jul1 -1.750e+00 8.410e-02 -20.814 < 2e-16 \*\*\*

month.jun1 -4.214e-01 9.485e-02 -4.442 8.90e-06 \*\*\*

month.mar1 7.494e-01 1.218e-01 6.152 7.67e-10 \*\*\*

month.may1 -1.291e+00 8.138e-02 -15.867 < 2e-16 \*\*\*

month.nov1 -1.762e+00 9.041e-02 -19.495 < 2e-16 \*\*\*

poutcome.other1 2.895e-01 7.797e-02 3.713 0.000205 \*\*\*

poutcome.success1 2.393e+00 6.554e-02 36.515 < 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: 32631** on 45210 degrees of freedom

**Residual deviance: 21624** on 45186 degrees of freedom

**AIC: 21674**

Number of Fisher Scoring iterations: 6

**> #Improvised model-> Logit1 is not performing better than logit model as given by the AIC value. So we will go with the logit model itself**

> #model with least AIC is better model

> prob <- predict(logit,type=c("response"),mymm)

Warning message:

In predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :

prediction from a rank-deficient fit may be misleading

> prob

1 2 3 4 5 6 7 8

0.014727947 0.009685532 0.002953187 0.005517058 0.021315020 0.008646341 0.010289748 0.022699351

9 10 11 12 13 14 15 16

0.006210440 0.006542593 0.014155953 0.010909609 0.033448459 0.005676271 0.007786911 0.021795673

17 18 19 20 21 22 23 24

0.009885215 0.003372067 0.012528806 0.004699814 0.006786813 0.009682629 0.004802371 0.015550384

25 26 27 28 29 30 31 32

0.006999632 0.009561499 0.021650995 0.003667875 0.013726669 0.014421968 0.013033251 0.012656537

33 34 35 36 37 38 39 40

0.007335093 0.008859633 0.022303922 0.012915099 0.015698771 0.893794045 0.059966143 0.010865414 ………………………………………………..

……………………………….

[ reached getOption("max.print") -- omitted 44211 entries ]

>

**> #confusion mat**

> conf<-table(prob**>0.5**,mymm$y)

> table(mymm$y)

0 1

39922 5289

> conf

0 1

FALSE 38940 3456

TRUE 982 1833

>

>

> #accuracy

> Accuracy<-sum(diag(conf)/sum(conf))

**> Accuracy**

**[1] 0.901838**

**>#Accuracy is 90.18%**

> #ROCR

> #install.packages("ROCR")

> library('ROCR')

> rocrpred<-prediction(prob,mymm$y)

> rocrperf<-performance(rocrpred,'tpr','fpr')

>

> str(rocrperf)

Formal class 'performance' [package "ROCR"] with 6 slots

..@ x.name : chr "False positive rate"

..@ y.name : chr "True positive rate"

..@ alpha.name : chr "Cutoff"

..@ x.values :List of 1

.. ..$ : num [1:45212] 0.00 2.50e-05 2.50e-05 5.01e-05 7.51e-05 ...

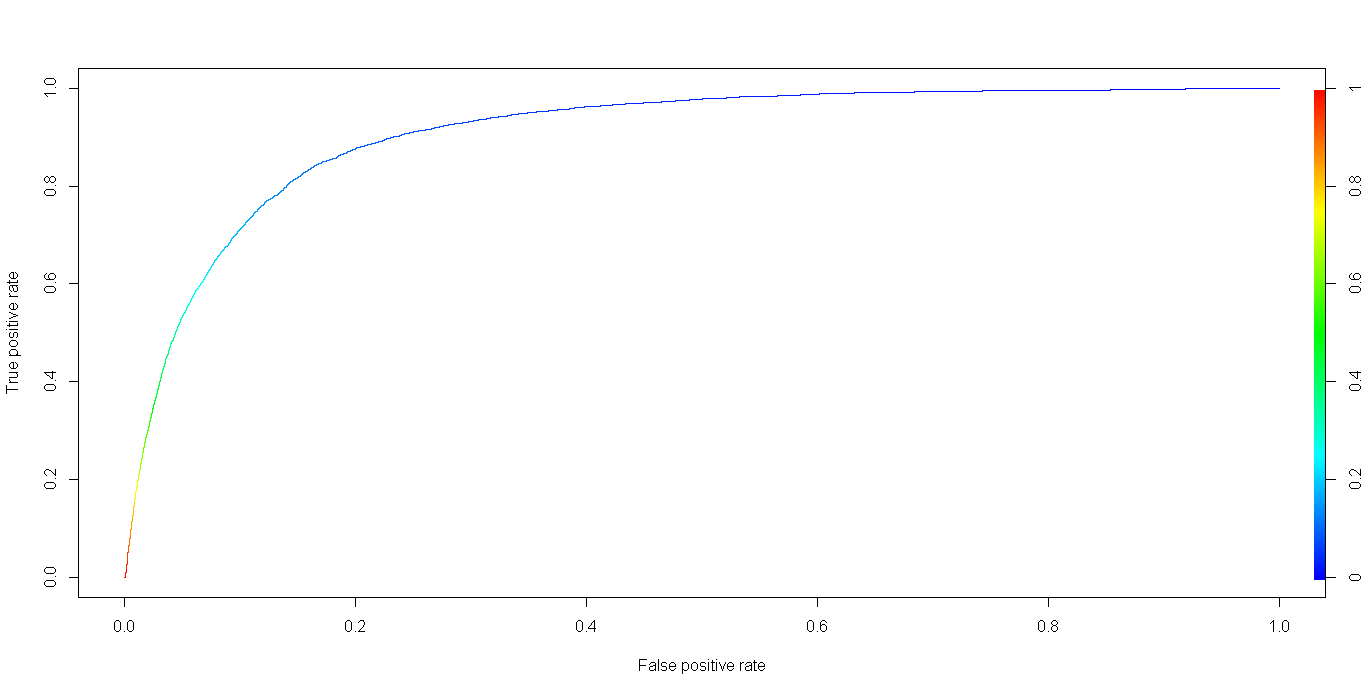
..@ y.values :List of 1

.. ..$ : num [1:45212] 0 0 0.000189 0.000189 0.000189 ...

..@ alpha.values:List of 1

.. ..$ : num [1:45212] Inf 1 1 1 1 ...

> plot(rocrperf,colorize=T,text.adj=c(-0.2,1.7))

> 

> rocr\_cutoff <- data.frame(cut\_off = rocrperf@alpha.values[[1]],fpr=rocrperf@x.values,tpr=rocrperf@y.values)

> colnames(rocr\_cutoff) <- c("cut\_off","FPR","TPR")

> rocr\_cutoff<-round(rocr\_cutoff,2)

>

**> #from the table we can see best accuracy is for cut off of 0.39**

> conf<-table(**prob>0.39**,mymm$y)

>

> Accuracy<-sum(diag(conf)/sum(conf))

**> Accuracy**

**[1] 0.9037624**

**#Thus improved accuracy is 90.37%**

**Problem 2 –**

classify that a person had an affair or not

Data description :A data frame containing 601 observations on 9 variables.affairs : numeric. How often engaged in extramarital sexual intercourse during the past yeargender : factor indicating gender.age : numeric variable coding age in years: 17.5 = under 20, 22 = 20–24, 27 = 25–29, 32 = 30–34, 37 = 35–39, 42 = 40–44, 47 = 45–49, 52 = 50–54, 57 = 55 or over.yearsmarried : numeric variable coding number of years married: 0.125 = 3 months or less, 0.417 = 4–6 months, 0.75 = 6 months–1 year, 1.5 = 1–2 years, 4 = 3–5 years, 7 = 6–8 years, 10 = 9–11 years, 15 = 12 or more years.

children : factor. Are there children in the marriage?religiousness : numeric variable coding religiousness: 1 = anti, 2 = not at all, 3 = slightly, 4 = somewhat, 5 = very.education : numeric variable coding level of education: 9 = grade school, 12 = high school graduate, 14 = some college, 16 = college graduate, 17 = some graduate work, 18 = master's degree, 20 = Ph.D., M.D., or other advanced degree.occupation : numeric variable coding occupation according to Hollingshead classification (reverse numbering).rating : numeric variable coding self rating of marriage: 1 = very unhappy, 2 = somewhat unhappy, 3 = average, 4 = happier than average, 5 = very happy.

Sol - >

**The Code with IMP highlight and explanation - >**

> myMatrix<-read.csv(file.choose())

> mymm<-myMatrix

>

**> #converting affairs in into 1 or 0 where 1 is yes affair and no affair resp.**

> mymm$affairs<- ifelse(mymm$affairs>0,1,0)

>

> mymm$gender<-(factor(as.numeric(mymm$gender)-1))

> mymm$children<-(factor(as.numeric(mymm$gender)-1))

> mymm$affairs<-(factor(as.numeric(mymm$affairs)))

>

> sum(is.na(mymm))

[1] 0

>

> summary(mymm)

affairs gender age yearsmarried children religiousness education

0:451 0:315 Min. :17.50 Min. : 0.125 0:315 Min. :1.000 Min. : 9.00

1:150 1:286 1st Qu.:27.00 1st Qu.: 4.000 1:286 1st Qu.:2.000 1st Qu.:14.00

Median :32.00 Median : 7.000 Median :3.000 Median :16.00

Mean :32.49 Mean : 8.178 Mean :3.116 Mean :16.17

3rd Qu.:37.00 3rd Qu.:15.000 3rd Qu.:4.000 3rd Qu.:18.00

Max. :57.00 Max. :15.000 Max. :5.000 Max. :20.00

occupation rating

Min. :1.000 Min. :1.000

1st Qu.:3.000 1st Qu.:3.000

Median :5.000 Median :4.000

Mean :4.195 Mean :3.932

3rd Qu.:6.000 3rd Qu.:5.000

Max. :7.000 Max. :5.000

> str(mymm)

'data.frame': 601 obs. of 9 variables:

$ affairs : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ gender : Factor w/ 2 levels "0","1": 2 1 1 2 2 1 1 2 1 2 ...

$ age : num 37 27 32 57 22 32 22 57 32 22 ...

$ yearsmarried : num 10 4 15 15 0.75 1.5 0.75 15 15 1.5 ...

$ children : Factor w/ 2 levels "0","1": 2 1 1 2 2 1 1 2 1 2 ...

$ religiousness: int 3 4 1 5 2 2 2 2 4 4 ...

$ education : int 18 14 12 18 17 17 12 14 16 14 ...

$ occupation : int 7 6 1 6 6 5 1 4 1 4 ...

$ rating : int 4 4 4 5 3 5 3 4 2 5 ...

> attach(mymm)

> logit <- glm(affairs~.,family="binomial",data=mymm)

> summary(logit)

Call:

glm(formula = affairs ~ ., family = "binomial", data = mymm)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5752 -0.7467 -0.5666 -0.2757 2.4207

Coefficients: (1 not defined because of singularities)

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

(Intercept) 1.58306 0.87349 1.812 0.069933 .

gender1 0.31432 0.23689 1.327 0.184554

age -0.04507 0.01819 -2.478 0.013226 \*

yearsmarried 0.11186 0.02981 3.752 0.000175 \*\*\*

children1 NA NA NA NA

religiousness -0.32317 0.08963 -3.606 0.000311 \*\*\*

education 0.02384 0.05043 0.473 0.636379

occupation 0.01624 0.07104 0.229 0.819185

rating -0.47633 0.09071 -5.251 1.51e-07 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

**Null deviance: 675.38** on 600 degrees of freedom

**Residual deviance: 611.40** on 593 degrees of freedom

**AIC: 627.4**

Number of Fisher Scoring iterations: 4

>

> logit1 <- glm(affairs~.-gender-children-education-occupation ,family="binomial",data=mymm)

> summary(logit1)

Call:

glm(formula = affairs ~ . - gender - children - education - occupation,

family = "binomial", data = mymm)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6278 -0.7550 -0.5701 -0.2624 2.3998

Coefficients:

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

(Intercept) 1.93083 0.61032 3.164 0.001558 \*\*

age -0.03527 0.01736 -2.032 0.042127 \*

yearsmarried 0.10062 0.02921 3.445 0.000571 \*\*\*

religiousness -0.32902 0.08945 -3.678 0.000235 \*\*\*

rating -0.46136 0.08884 -5.193 2.06e-07 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

**Null deviance: 675.38** on 600 degrees of freedom

**Residual deviance: 615.36** on 596 degrees of freedom

**AIC: 625.36**

Number of Fisher Scoring iterations: 4

>

**> #model with least AIC is better model**

**> #here improvised model gave better result so we will go with model logit1**

> prob <- predict(logit1,type=c("response"),mymm)

> prob

1 2 3 4 5 6 7 8 9

0.23138786 0.14423644 0.53420297 0.07431584 0.30750719 0.11797305 0.30750719 0.25468568 0.51816131

10 11 12 13 14 15 16 17 18

0.08971637 0.63514871 0.14423644 0.15340560 0.23188574 0.14423644 0.37737826 0.52323522 0.16767650

19 20 21 22 23 24 25 26 27

0.15988886 0.27286846 0.15988886 0.15988886 0.23562737 0.18459857 0.18615297 0.15988886 0.12564418

28 29 30 31 32 33 34 35 36

0.17773330 0.12865789 0.18976831 0.32269908 0.12046024 0.10516851 0.17773330 0.27491554 0.19870943

37 38 39 40 41 42 43 44 45

0.29447814 0.11249094 0.20915418 0.11269076 0.10493599 0.19417130 0.09164276 0.56502756 0.17906574

46 47 48 49 50 51 52 53 54

0.14975522 0.12564418 0.36709280 0.07631804 0.36709280 0.06171111 0.26953563 0.07103321 0.23562737

55 56 57 58 59 60 ……………………………………………………..

>

> #confusion mat

> conf<-table(prob>0.5,mymm$affairs)

> table(mymm$affairs)

0 1

451 150

> conf

0 1

FALSE 432 128

TRUE 19 22

>

> #accuracy

> Accuracy<-sum(diag(conf)/sum(conf))

**> Accuracy**

**[1] 0.7554077**

>

> #ROCR

> #install.packages("ROCR")

> library('ROCR')

> rocrpred<-prediction(prob,mymm$affairs)

> rocrperf<-performance(rocrpred,'tpr','fpr')

>

> str(rocrperf)

Formal class 'performance' [package "ROCR"] with 6 slots

..@ x.name : chr "False positive rate"

..@ y.name : chr "True positive rate"

..@ alpha.name : chr "Cutoff"

..@ x.values :List of 1

.. ..$ : num [1:277] 0 0 0.00443 0.00443 0.00443 ...

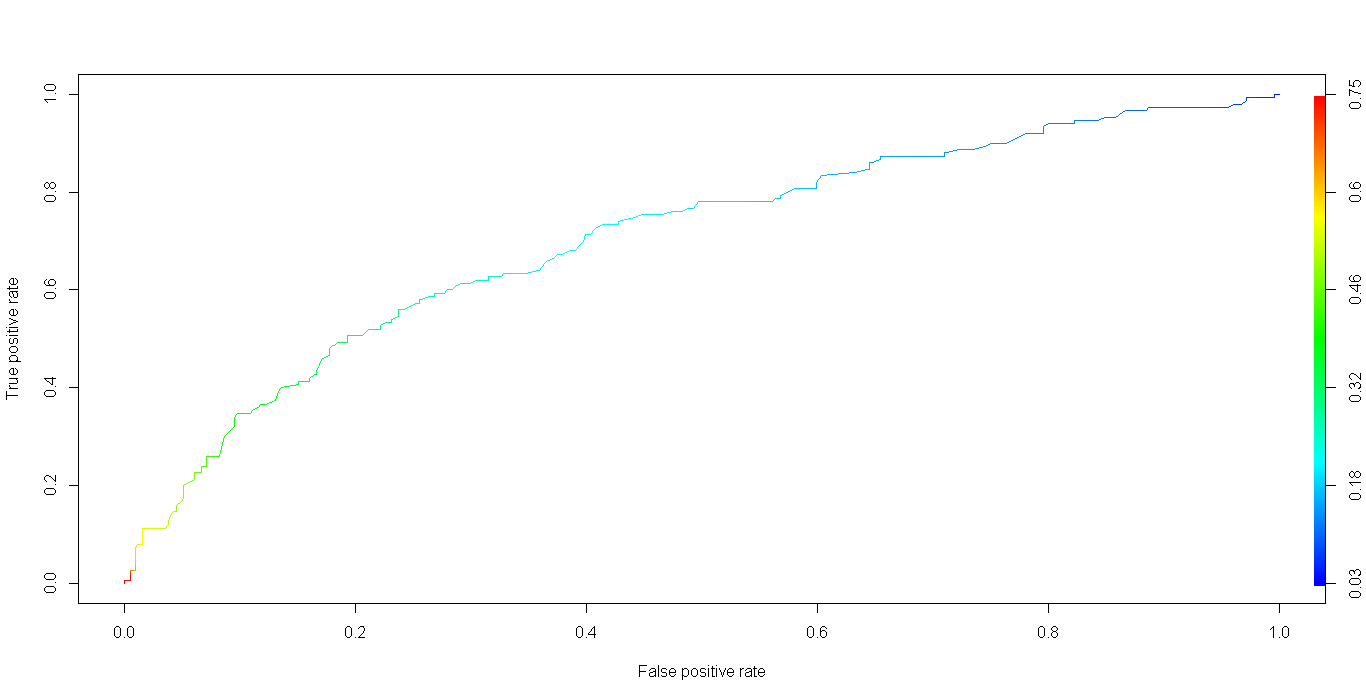
..@ y.values :List of 1

.. ..$ : num [1:277] 0 0.00667 0.00667 0.02 0.02667 ...

..@ alpha.values:List of 1

.. ..$ : num [1:277] Inf 0.743 0.734 0.703 0.644 ...

> plot(rocrperf,colorize=T,text.adj=c(-0.2,1.7))

> 

> rocr\_cutoff <- data.frame(cut\_off = rocrperf@alpha.values[[1]],fpr=rocrperf@x.values,tpr=rocrperf@y.values)

> colnames(rocr\_cutoff) <- c("cut\_off","FPR","TPR")

> rocr\_cutoff<-round(rocr\_cutoff,2)

>

**> #from the table we can see best accuracy is for cut off of 0.39**

> conf<-table(prob>0.39,mymm$affairs)

>

> Accuracy<-sum(diag(conf)/sum(conf))

**> Accuracy**

**[1] 0.7637271**

**Thus Improved Accuracy is 76.37%**