HR-Case-Study.R

Souvik

2020-12-03

```
setwd("C:/Users/Souvik/Downloads/PPA")

#Loading the required packages
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(car)
```

```
## Loading required package: carData
```

```
library(ROCR)
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(caTools)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
library(rpart)
library(rpart.plot)
library(e1071)
library(rattle.data)
library(rattle)
```

```
## Loading required package: tibble
```

```
## Loading required package: bitops
```

```
## Rattle: A free graphical interface for data science with R.
## Version 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
##
## Attaching package: 'rattle'
```

```
## The following objects are masked from 'package:rattle.data':
##
##     locationsAUS, weather, weatherAUS
```

```
#Loading the datasets
HR <- read.csv("promotion_tr.csv", stringsAsFactors = TRUE)
summary(HR)
```

```
##   employee_id                department          region
##  Min.   :    1   Sales & Marketing:16840   region_2 :12343
##  1st Qu.:19670   Operations       :11348   region_22: 6428
##  Median :39226   Procurement      : 7138   region_7 : 4843
##  Mean   :39196   Technology       : 7138   region_15: 2808
##  3rd Qu.:58731   Analytics        : 5352   region_13: 2648
##  Max.   :78298   Finance          : 2536   region_26: 2260
##                  (Other)          : 4456   (Other)  :23478
##             education     gender    recruitment_channel no_of_trainings
##                  : 2409   f:16312   other   :30446      Min.   : 1.000
##  Bachelor's      :36669   m:38496   referred: 1142      1st Qu.: 1.000
##  Below Secondary :  805             sourcing:23220      Median : 1.000
##  Master's & above:14925                                 Mean   : 1.253
##                                                         3rd Qu.: 1.000
##                                                         Max.   :10.000
##
##       age       previous_year_rating length_of_service KPIs_met..80.
##  Min.   :20.0   Min.   :1.000        Min.   : 1.000    Min.   :0.000
##  1st Qu.:29.0   1st Qu.:3.000        1st Qu.: 3.000    1st Qu.:0.000
##  Median :33.0   Median :3.000        Median : 5.000    Median :0.000
##  Mean   :34.8   Mean   :3.329        Mean   : 5.866    Mean   :0.352
##  3rd Qu.:39.0   3rd Qu.:4.000        3rd Qu.: 7.000    3rd Qu.:1.000
##  Max.   :60.0   Max.   :5.000        Max.   :37.000    Max.   :1.000
##                 NA's   :4124
##   awards_won.      avg_training_score  is_promoted
##  Min.   :0.00000   Min.   :39.00      Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:51.00      1st Qu.:0.00000
##  Median :0.00000   Median :60.00      Median :0.00000
##  Mean   :0.02317   Mean   :63.39      Mean   :0.08517
##  3rd Qu.:0.00000   3rd Qu.:76.00      3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :99.00      Max.   :1.00000
##
```

```
HR$education <- as.factor(HR$education)
table(HR$education)
```

```
##
##                         Bachelor's  Below Secondary Master's & above
##             2409             36669              805             14925
```

```
table(HR$recruitment_channel)
```

```
##
##    other referred sourcing
##    30446     1142    23220
```

```
table(HR$is_promoted)
```

```
##
##     0     1
## 50140  4668
```

```
#Cleaning the dataset by removing and imputing rows

#mode imputation on column previous year rating
table(HR$previous_year_rating)
```

```
##
##     1     2     3     4     5
##  6223  4225 18618  9877 11741
```

```
HR$previous_year_rating[is.na(HR$previous_year_rating)] <- 3
table(HR$previous_year_rating)
```

```
##
##     1     2     3     4     5
##  6223  4225 22742  9877 11741
```

```
summary(HR$previous_year_rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.304   4.000   5.000
```

```
#NA imputing on empty cells on column education
HR[which(HR$education==""),]$education <-NA
summary(HR$education)
```

```
##                         Bachelor's  Below Secondary Master's & above
##                0             36669              805             14925
##             NA's
##             2409
```

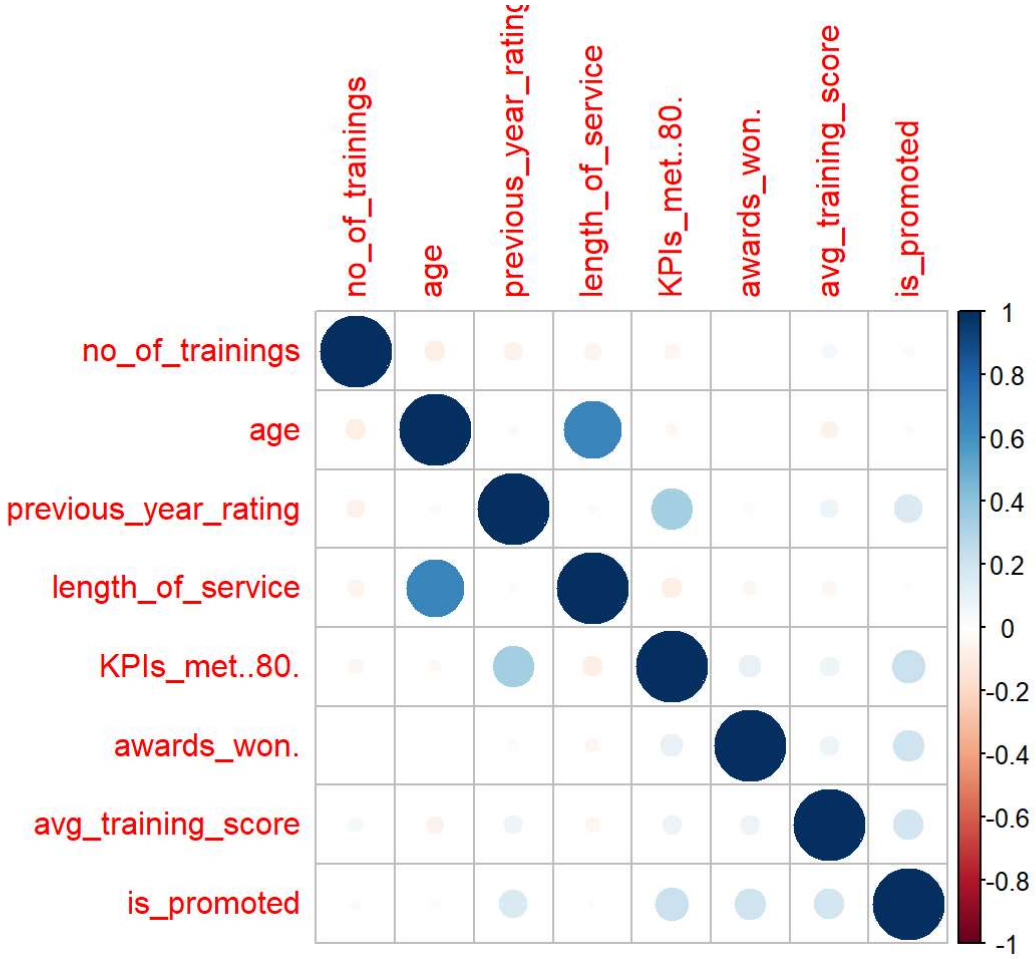
```
#removing all NA rows from the dataset
HR <- na.omit(HR)
summary(HR)
```

```
##   employee_id                department          region
##  Min.   :    1   Sales & Marketing:15265   region_2 :11497
##  1st Qu.:19652   Operations       :11122   region_22: 6108
##  Median :39207   Procurement      : 7066   region_7 : 4624
##  Mean   :39184   Technology       : 7039   region_15: 2617
##  3rd Qu.:58739   Analytics        : 5015   region_13: 2592
##  Max.   :78298   Finance          : 2500   region_26: 2160
##                  (Other)          : 4392   (Other)  :22801
##             education     gender    recruitment_channel no_of_trainings
##                  :    0   f:15921   other   :29061      Min.   : 1.000
##  Bachelor's      :36669   m:36478   referred: 1134      1st Qu.: 1.000
##  Below Secondary :  805             sourcing:22204      Median : 1.000
##  Master's & above:14925                                 Mean   : 1.256
##                                                         3rd Qu.: 1.000
##                                                         Max.   :10.000
##
##       age        previous_year_rating length_of_service KPIs_met..80.
##  Min.   :20.00   Min.   :1.000        Min.   : 1.000    Min.   :0.0000
##  1st Qu.:29.00   1st Qu.:3.000        1st Qu.: 3.000    1st Qu.:0.0000
##  Median :33.00   Median :3.000        Median : 5.000    Median :0.0000
##  Mean   :34.98   Mean   :3.313        Mean   : 5.933    Mean   :0.3568
##  3rd Qu.:39.00   3rd Qu.:4.000        3rd Qu.: 8.000    3rd Qu.:1.0000
##  Max.   :60.00   Max.   :5.000        Max.   :37.000    Max.   :1.0000
##
##   awards_won.      avg_training_score  is_promoted
##  Min.   :0.00000   Min.   :39.00      Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:51.00      1st Qu.:0.00000
##  Median :0.00000   Median :60.00      Median :0.00000
##  Mean   :0.02317   Mean   :63.63      Mean   :0.08676
##  3rd Qu.:0.00000   3rd Qu.:76.00      3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :99.00      Max.   :1.00000
##
```

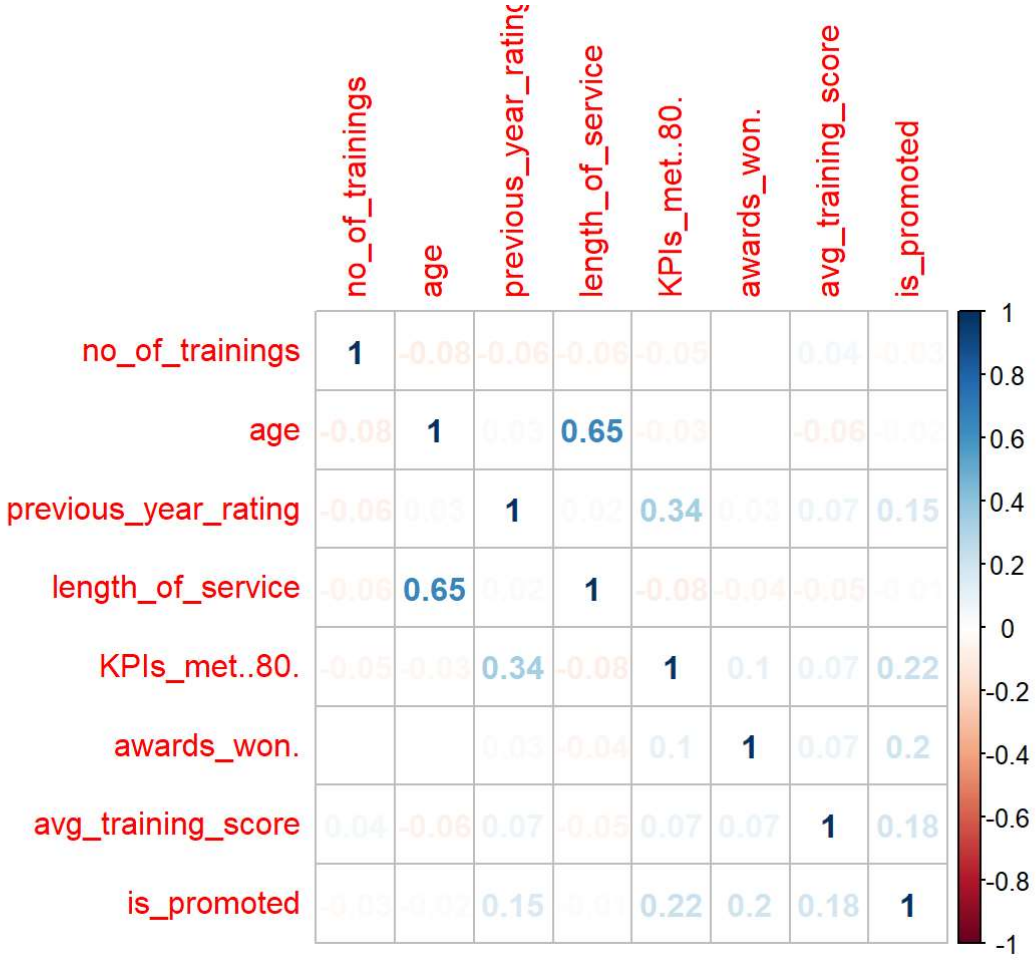
```
#checking correlation between the variables
CR <- cor(HR[c("no_of_trainings","age","previous_year_rating",
               "length_of_service","KPIs_met..80.","awards_won.",
               "avg_training_score","is_promoted")])
CR
```

```
##                       no_of_trainings          age previous_year_rating
## no_of_trainings           1.000000000 -0.082600969          -0.06266236
## age                      -0.082600969  1.000000000           0.02543949
## previous_year_rating     -0.062662360  0.025439493           1.00000000
## length_of_service        -0.057844984  0.652378704           0.02167519
## KPIs_met..80.            -0.047343765 -0.030966550           0.33851265
## awards_won.              -0.007658649 -0.008208638           0.02677259
## avg_training_score        0.041124495 -0.060652381           0.06673566
## is_promoted              -0.025388665 -0.018920186           0.15292145
##                       length_of_service KPIs_met..80.  awards_won.
## no_of_trainings             -0.05784498   -0.04734376 -0.007658649
## age                          0.65237870   -0.03096655 -0.008208638
## previous_year_rating         0.02167519    0.33851265  0.026772591
## length_of_service            1.00000000   -0.08210874 -0.040140083
## KPIs_met..80.               -0.08210874    1.00000000  0.095554930
## awards_won.                 -0.04014008    0.09555493  1.000000000
## avg_training_score          -0.04532875    0.07289898  0.072069289
## is_promoted                 -0.01216673    0.21993362  0.195451465
##                       avg_training_score is_promoted
## no_of_trainings              0.04112450 -0.02538867
## age                         -0.06065238 -0.01892019
## previous_year_rating         0.06673566  0.15292145
## length_of_service           -0.04532875 -0.01216673
## KPIs_met..80.                0.07289898  0.21993362
## awards_won.                  0.07206929  0.19545147
## avg_training_score           1.00000000  0.18048937
## is_promoted                  0.18048937  1.00000000
```

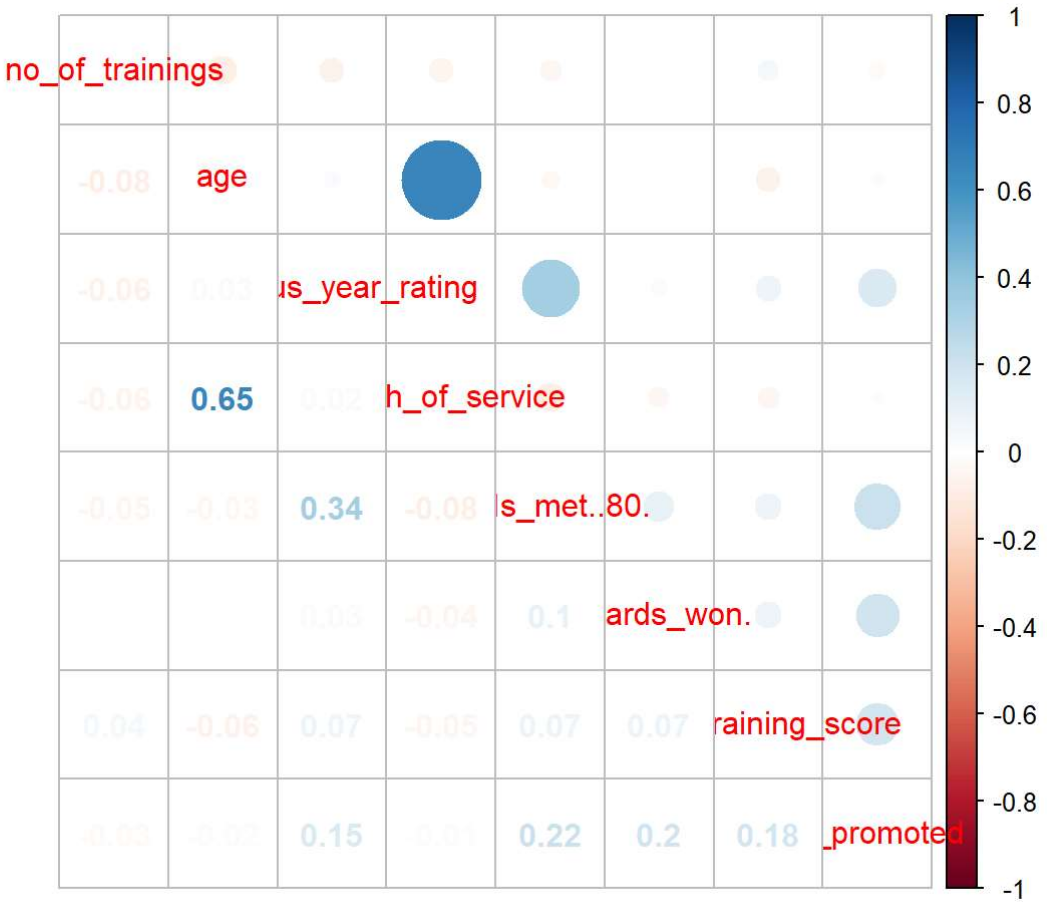
```
corrplot(CR, type = "full")
```



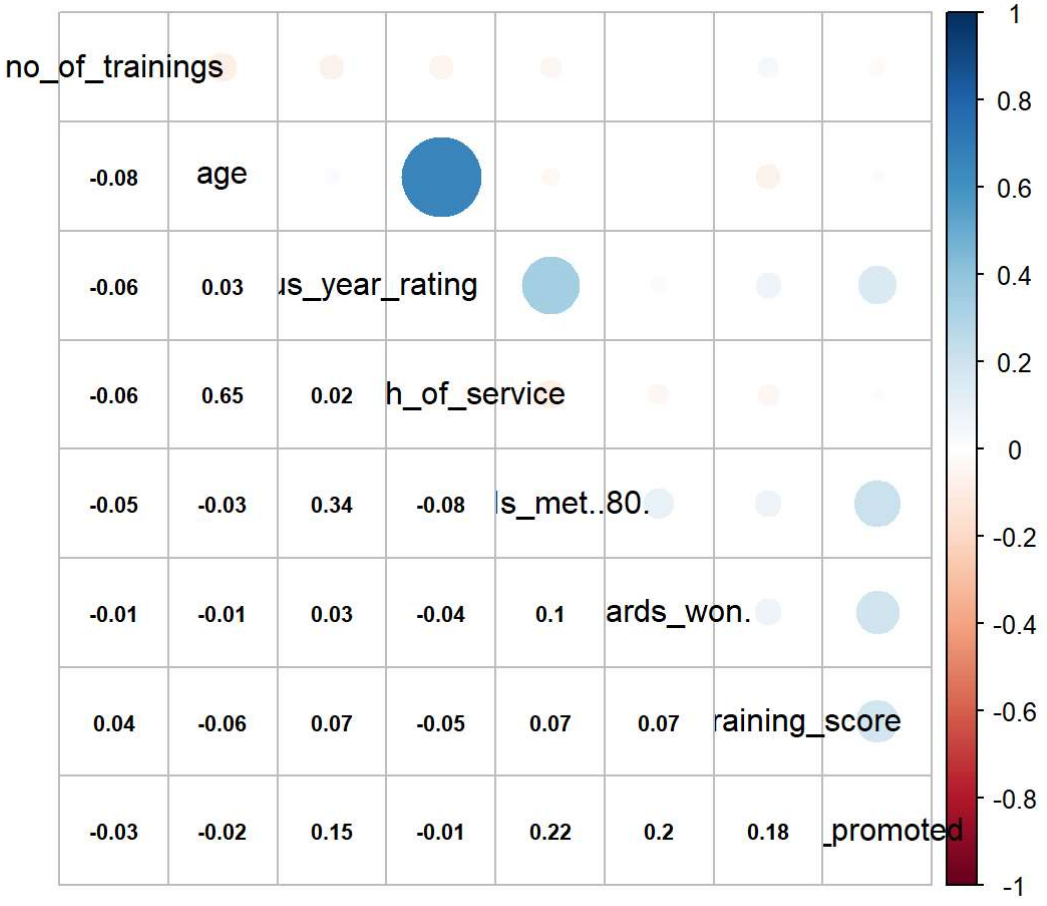
```
corrplot(CR,method = "number")
```




```
corrplot.mixed(CR)
```




```
corrplot.mixed(CR,  tl.col = "black", lower.col = "black",number.cex = .7,tl.cex=1 )
```



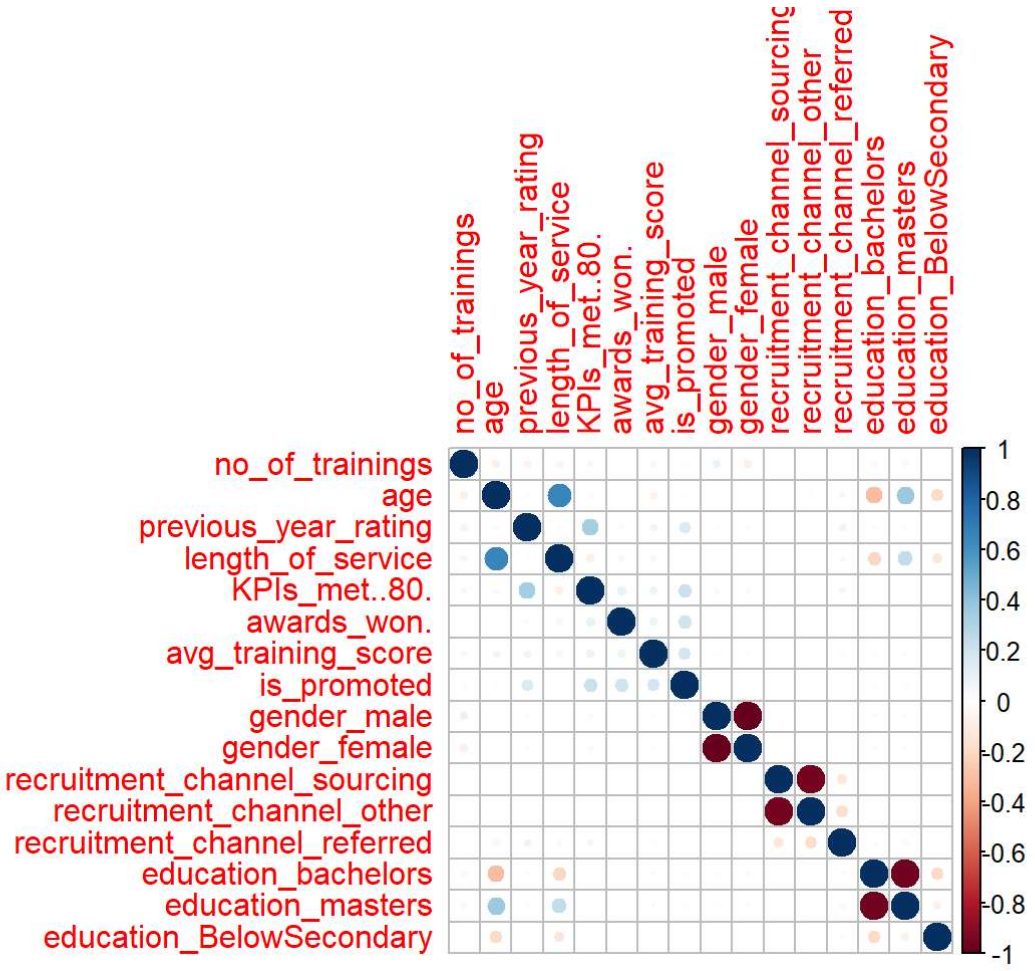
```
#To create dummy variables
HR$gender_male<- ifelse(HR$gender == "m", 1,0)
HR$gender_female<- ifelse(HR$gender == "f", 1,0)

HR$recruitment_channel_sourcing<- ifelse(HR$recruitment_channel == "sourcing", 1,0)
HR$recruitment_channel_other<- ifelse(HR$recruitment_channel == "other", 1,0)
HR$recruitment_channel_referred<- ifelse(HR$recruitment_channel == "referred", 1,0)

HR$education_bachelors<- ifelse(HR$education == "Bachelor's", 1,0)
HR$education_masters<- ifelse(HR$education == "Master's & above", 1,0)
HR$education_BelowSecondary<- ifelse(HR$education == "Below Secondary", 1,0)


CR1 <- cor(HR[c("no_of_trainings","age","previous_year_rating",
               "length_of_service","KPIs_met..80.","awards_won.",
               "avg_training_score","is_promoted","gender_male","gender_female",
               "recruitment_channel_sourcing","recruitment_channel_other",
               "recruitment_channel_referred","education_bachelors","education_masters",
               "education_BelowSecondary")])
CR1
```

```
##                              no_of_trainings          age previous_year_rating
## no_of_trainings                 1.000000000 -0.082600969         -0.062662360
## age                            -0.082600969  1.000000000          0.025439493
## previous_year_rating           -0.062662360  0.025439493          1.000000000
## length_of_service              -0.057844984  0.652378704          0.021675186
## KPIs_met..80.                  -0.047343765 -0.030966550          0.338512652
## awards_won.                    -0.007658649 -0.008208638          0.026772591
## avg_training_score              0.041124495 -0.060652381          0.066735661
## is_promoted                    -0.025388665 -0.018920186          0.152921451
## gender_male                     0.086653908 -0.003206708         -0.022217559
## gender_female                  -0.086653908  0.003206708          0.022217559
## recruitment_channel_sourcing   -0.009306297 -0.003725711         -0.004501268
## recruitment_channel_other       0.013542868  0.017017528         -0.014720747
## recruitment_channel_referred   -0.014653279 -0.045473009          0.065566792
## education_bachelors             0.036525399 -0.307610597         -0.022631812
## education_masters              -0.038178377  0.362175613          0.023898981
## education_BelowSecondary        0.003984614 -0.182706687         -0.003360493
##                              length_of_service KPIs_met..80.   awards_won.
## no_of_trainings                    -0.057844984  -0.047343765 -0.0076586487
## age                                 0.652378704  -0.030966550 -0.0082086384
## previous_year_rating                0.021675186   0.338512652  0.0267725907
## length_of_service                   1.000000000  -0.082108741 -0.0401400834
## KPIs_met..80.                      -0.082108741   1.000000000  0.0955549296
## awards_won.                        -0.040140083   0.095554930  1.0000000000
## avg_training_score                 -0.045328751   0.072898978  0.0720692892
## is_promoted                        -0.012166730   0.219933615  0.1954514650
## gender_male                        -0.011311920  -0.035892880  0.0027207314
## gender_female                       0.011311920   0.035892880 -0.0027207314
## recruitment_channel_sourcing        0.002921794  -0.007046226 -0.0067852896
## recruitment_channel_other           0.006590607  -0.006811937  0.0057949730
## recruitment_channel_referred       -0.032433414   0.047195914  0.0032493104
## education_bachelors                -0.209939779  -0.007346583  0.0020588354
## education_masters                   0.246854792   0.004391178 -0.0007836579
## education_BelowSecondary           -0.123502688   0.011263916 -0.0047967687
##                              avg_training_score   is_promoted  gender_male
## no_of_trainings                     0.041124495 -2.538867e-02  0.086653908
## age                                -0.060652381 -1.892019e-02 -0.003206708
## previous_year_rating                0.066735661  1.529215e-01 -0.022217559
## length_of_service                  -0.045328751 -1.216673e-02 -0.011311920
## KPIs_met..80.                       0.072898978  2.199336e-01 -0.035892880
## awards_won.                         0.072069289  1.954515e-01  0.002720731
## avg_training_score                  1.000000000  1.804894e-01 -0.018414671
## is_promoted                         0.180489370  1.000000e+00 -0.010575200
## gender_male                        -0.018414671 -1.057520e-02  1.000000000
## gender_female                       0.018414671  1.057520e-02 -1.000000000
## recruitment_channel_sourcing       -0.007302792 -4.951526e-05  0.003988762
## recruitment_channel_other          -0.001505156 -5.355196e-03 -0.006516994
## recruitment_channel_referred        0.029941240  1.845949e-02  0.008713816
## education_bachelors                -0.023555948 -2.563624e-02  0.026034704
## education_masters                   0.020598515  2.646089e-02 -0.022356179
## education_BelowSecondary            0.012194974 -1.565407e-03 -0.014982373
##                              gender_female recruitment_channel_sourcing
## no_of_trainings                -0.086653908                -9.306297e-03
## age                             0.003206708                -3.725711e-03
## previous_year_rating            0.022217559                -4.501268e-03
## length_of_service               0.011311920                 2.921794e-03
## KPIs_met..80.                   0.035892880                -7.046226e-03
```

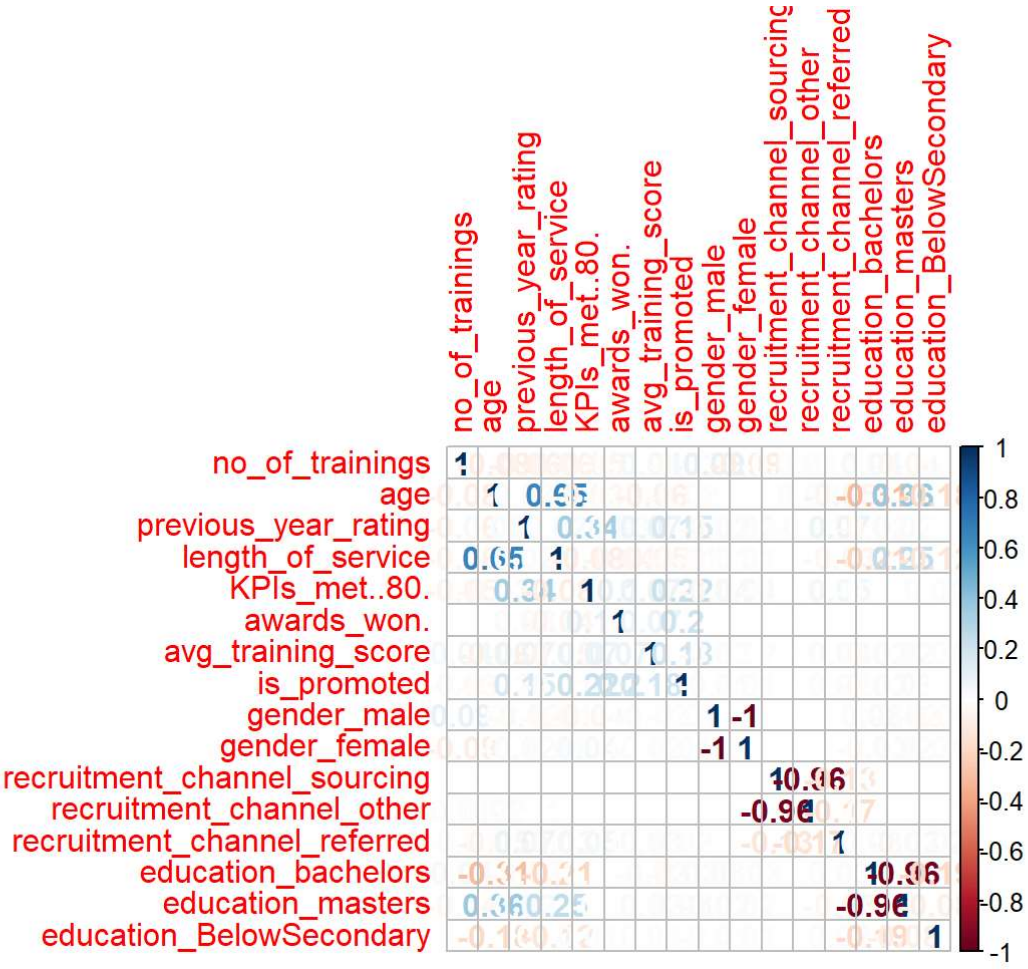
```
## awards_won.                   -0.002720731               -6.785290e-03
## avg_training_score             0.018414671               -7.302792e-03
## is_promoted                    0.010575200               -4.951526e-05
## gender_male                   -1.000000000                3.988762e-03
## gender_female                  1.000000000               -3.988762e-03
## recruitment_channel_sourcing  -0.003988762                1.000000e+00
## recruitment_channel_other      0.006516994               -9.569110e-01
## recruitment_channel_referred  -0.008713816               -1.275393e-01
## education_bachelors           -0.026034704               -2.894690e-04
## education_masters              0.022356179               -2.949599e-04
## education_BelowSecondary       0.014982373                2.161143e-03
##                              recruitment_channel_other
## no_of_trainings                          0.013542868
## age                                      0.017017528
## previous_year_rating                    -0.014720747
## length_of_service                        0.006590607
## KPIs_met..80.                           -0.006811937
## awards_won.                              0.005794973
## avg_training_score                      -0.001505156
## is_promoted                             -0.005355196
## gender_male                             -0.006516994
## gender_female                            0.006516994
## recruitment_channel_sourcing            -0.956911011
## recruitment_channel_other                1.000000000
## recruitment_channel_referred            -0.165966145
## education_bachelors                     -0.007203730
## education_masters                        0.009736963
## education_BelowSecondary                -0.008885607
##                              recruitment_channel_referred education_bachelors
## no_of_trainings                           -0.014653279         0.036525399
## age                                       -0.045473009        -0.307610597
## previous_year_rating                       0.065566792        -0.022631812
## length_of_service                         -0.032433414        -0.209939779
## KPIs_met..80.                              0.047195914        -0.007346583
## awards_won.                                0.003249310         0.002058835
## avg_training_score                         0.029941240        -0.023555948
## is_promoted                                0.018459490        -0.025636244
## gender_male                                0.008713816         0.026034704
## gender_female                             -0.008713816        -0.026034704
## recruitment_channel_sourcing              -0.127539332        -0.000289469
## recruitment_channel_other                 -0.165966145        -0.007203730
## recruitment_channel_referred               1.000000000         0.025588268
## education_bachelors                        0.025588268         1.000000000
## education_masters                         -0.032256127        -0.963556789
## education_BelowSecondary                   0.023010677        -0.190714412
##                              education_masters education_BelowSecondary
## no_of_trainings                -0.0381783771              0.003984614
## age                             0.3621756134             -0.182706687
## previous_year_rating            0.0238989806             -0.003360493
## length_of_service               0.2468547915             -0.123502688
## KPIs_met..80.                   0.0043911781              0.011263916
## awards_won.                    -0.0007836579             -0.004796769
## avg_training_score              0.0205985153              0.012194974
## is_promoted                     0.0264608939             -0.001565407
## gender_male                    -0.0223561791             -0.014982373
## gender_female                   0.0223561791              0.014982373
## recruitment_channel_sourcing   -0.0002949599              0.002161143
## recruitment_channel_other       0.0097369632             -0.008885607
```

```
## recruitment_channel_referred     -0.0322561273              0.023010677
## education_bachelors              -0.9635567888             -0.190714412
## education_masters                 1.0000000000             -0.078829811
## education_BelowSecondary         -0.0788298108              1.000000000
```

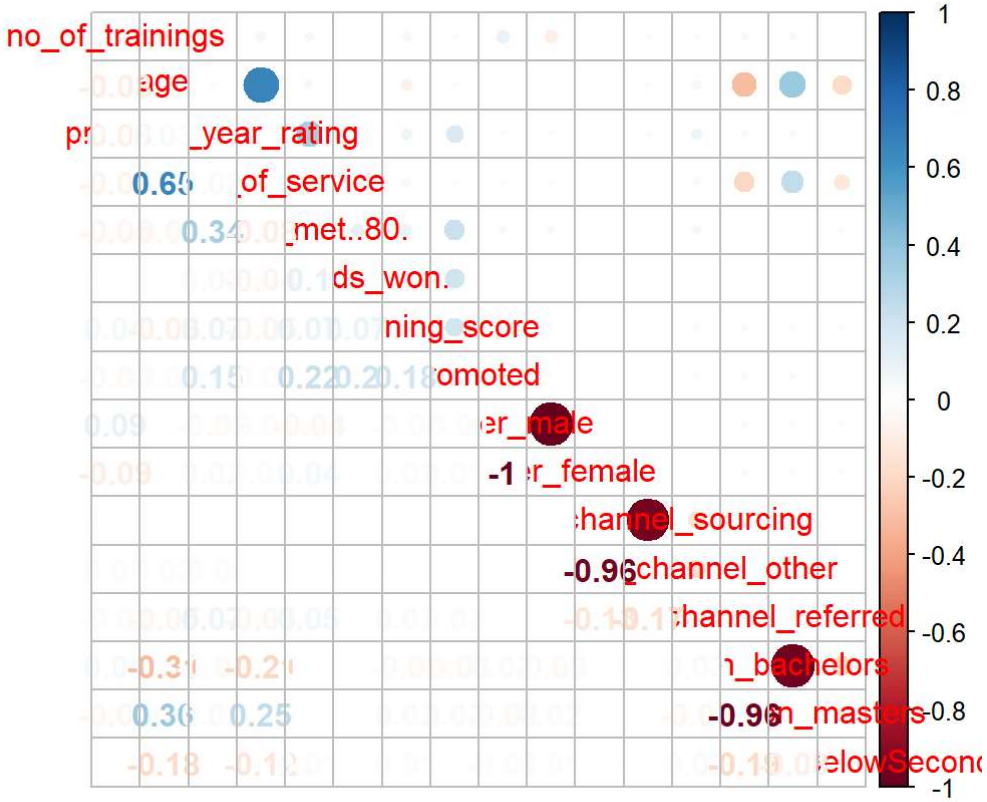
```
corrplot(CR1, type = "full")
```



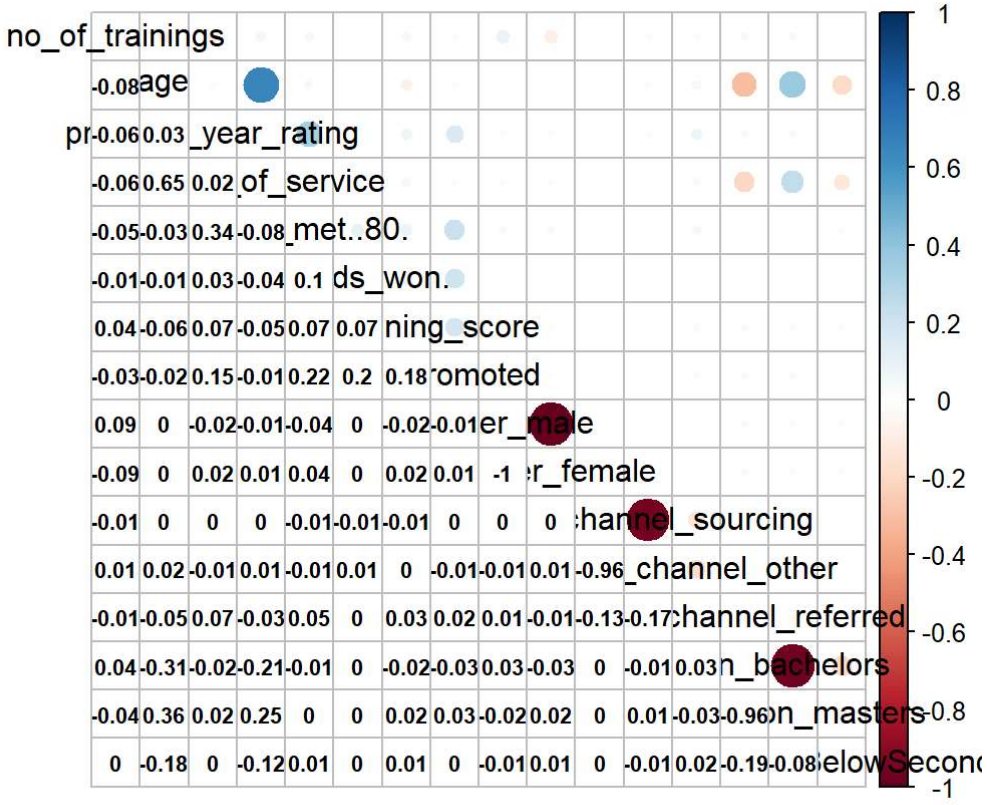

```
corrplot(CR1,method = "number")
```




```
corrplot.mixed(CR1)
```



```
corrplot.mixed(CR1,  tl.col = "black", lower.col = "black",number.cex = .7,tl.cex=1 )
```




```
#Splitting of the dataset into training and testing

split <- sample.split(HR$is_promoted, SplitRatio = 0.7)
HR_TR <- subset(HR, split == "TRUE")
HR_TS <- subset(HR, split == "FALSE")


#Logistic regression models
#adding variables at first
model1 <- glm(is_promoted ~ no_of_trainings, data=HR_TR, family = binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings, family = binomial,
##     data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4354  -0.4354  -0.4354  -0.3956   2.5107
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.10854    0.04666 -45.187  < 2e-16 ***
## no_of_trainings -0.19989    0.03563  -5.611 2.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21638  on 36678  degrees of freedom
## Residual deviance: 21602  on 36677  degrees of freedom
## AIC: 21606
##
## Number of Fisher Scoring iterations: 5
```

```
model2 <- glm(is_promoted ~ no_of_trainings+age, data=HR_TR, family = binomial)
summary(model2)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age, family = binomial,
##     data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4700  -0.4444  -0.4310  -0.4011   2.5572
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.72139    0.10171 -16.924  < 2e-16 ***
## no_of_trainings -0.21188    0.03580  -5.919 3.25e-09 ***
## age             -0.01071    0.00252  -4.252 2.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21638  on 36678  degrees of freedom
## Residual deviance: 21584  on 36676  degrees of freedom
## AIC: 21590
##
## Number of Fisher Scoring iterations: 5
```

```
model3 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating, data=HR_TR,
              family = binomial)
summary(model3)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating,
##     family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6624  -0.4791  -0.3793  -0.2970   2.8346
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.476268   0.122667 -28.339  < 2e-16 ***
## no_of_trainings      -0.160853   0.036359  -4.424 9.69e-06 ***
## age                  -0.014174   0.002602  -5.448 5.10e-08 ***
## previous_year_rating  0.503061   0.017500  28.746  < 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: 21638  on 36678  degrees of freedom
## Residual deviance: 20661  on 36675  degrees of freedom
## AIC: 20669
##
## Number of Fisher Scoring iterations: 5
```

```
model4 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service,
              data=HR_TR, family = binomial)
summary(model4)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating +
##     length_of_service, family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6599  -0.4785  -0.3794  -0.2971   2.8392
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.517758   0.129684 -27.126  < 2e-16 ***
## no_of_trainings      -0.160931   0.036355  -4.427 9.57e-06 ***
## age                  -0.012038   0.003392  -3.549 0.000387 ***
## previous_year_rating  0.503553   0.017514  28.751  < 2e-16 ***
## length_of_service    -0.005913   0.006090  -0.971 0.331570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21638  on 36678  degrees of freedom
## Residual deviance: 20660  on 36674  degrees of freedom
## AIC: 20670
##
## Number of Fisher Scoring iterations: 5
```

```
model5 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
              +KPIs_met..80.,
              data=HR_TR, family = binomial)
summary(model5)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating +
##     length_of_service + KPIs_met..80., family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7673  -0.5127  -0.2951  -0.2510   2.8404
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.516218   0.131391 -26.761  < 2e-16 ***
## no_of_trainings      -0.137801   0.036882  -3.736 0.000187 ***
## age                  -0.014794   0.003453  -4.285 1.83e-05 ***
## previous_year_rating  0.306131   0.018374  16.661  < 2e-16 ***
## length_of_service     0.010766   0.006178   1.743 0.081411 .
## KPIs_met..80.         1.325704   0.042853  30.936  < 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: 21638  on 36678  degrees of freedom
## Residual deviance: 19623  on 36673  degrees of freedom
## AIC: 19635
##
## Number of Fisher Scoring iterations: 6
```

```
model6 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
              +KPIs_met..80.+awards_won.,
              data=HR_TR, family = binomial)
summary(model6)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won., family = binomial,
##     data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5387  -0.4816  -0.2849  -0.2395   2.8855
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.599499   0.134241 -26.814  < 2e-16 ***
## no_of_trainings      -0.131097   0.037451  -3.501 0.000464 ***
## age                  -0.018837   0.003544  -5.315 1.06e-07 ***
## previous_year_rating  0.329142   0.018661  17.638  < 2e-16 ***
## length_of_service     0.021825   0.006283   3.474 0.000513 ***
## KPIs_met..80.         1.257056   0.043242  29.070  < 2e-16 ***
## awards_won.           2.051533   0.076908  26.675  < 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: 21638  on 36678  degrees of freedom
## Residual deviance: 19000  on 36672  degrees of freedom
## AIC: 19014
##
## Number of Fisher Scoring iterations: 6
```

```
model7 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
              +KPIs_met..80.+awards_won.+avg_training_score,
              data=HR_TR, family = binomial)
summary(model7)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won. + avg_training_score,
##     family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8795  -0.4327  -0.2797  -0.1909   2.9198
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -6.596934   0.174352 -37.837  < 2e-16 ***
## no_of_trainings      -0.237970   0.039490  -6.026 1.68e-09 ***
## age                  -0.014029   0.003610  -3.887 0.000102 ***
## previous_year_rating  0.321109   0.019239  16.691  < 2e-16 ***
## length_of_service     0.023563   0.006388   3.688 0.000226 ***
## KPIs_met..80.         1.249972   0.044045  28.379  < 2e-16 ***
## awards_won.           1.914970   0.079879  23.973  < 2e-16 ***
## avg_training_score    0.044526   0.001498  29.732  < 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: 21638  on 36678  degrees of freedom
## Residual deviance: 18077  on 36671  degrees of freedom
## AIC: 18093
##
## Number of Fisher Scoring iterations: 6
```

```
model8 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
              +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
              +education_masters+education_BelowSecondary,
              data=HR_TR, family = binomial)
summary(model8)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won. + avg_training_score +
##     education_bachelors + education_masters + education_BelowSecondary,
##     family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8572  -0.4330  -0.2793  -0.1901   2.9393
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.506095   0.225365 -28.869  < 2e-16 ***
## no_of_trainings          -0.235047   0.039517  -5.948 2.71e-09 ***
## age                      -0.019653   0.003858  -5.094 3.50e-07 ***
## previous_year_rating      0.320243   0.019253  16.634  < 2e-16 ***
## length_of_service         0.023374   0.006422   3.640 0.000273 ***
## KPIs_met..80.             1.249610   0.044076  28.351  < 2e-16 ***
## awards_won.               1.918465   0.079914  24.007  < 2e-16 ***
## avg_training_score        0.044252   0.001499  29.522  < 2e-16 ***
## education_bachelors       0.056558   0.162806   0.347 0.728292
## education_masters         0.280123   0.168742   1.660 0.096901 .
## education_BelowSecondary        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: 21638  on 36678  degrees of freedom
## Residual deviance: 18054  on 36669  degrees of freedom
## AIC: 18074
##
## Number of Fisher Scoring iterations: 6
```

```
model9 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
              +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
              +education_masters+education_BelowSecondary+recruitment_channel_sourcing
              +recruitment_channel_other+recruitment_channel_referred,
              data=HR_TR, family = binomial)
summary(model9)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won. + avg_training_score +
##     education_bachelors + education_masters + education_BelowSecondary +
##     recruitment_channel_sourcing + recruitment_channel_other +
##     recruitment_channel_referred, family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8498  -0.4327  -0.2790  -0.1900   2.9430
##
## Coefficients: (2 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.340222   0.249135 -25.449  < 2e-16 ***
## no_of_trainings              -0.234208   0.039523  -5.926 3.11e-09 ***
## age                          -0.019493   0.003859  -5.051 4.39e-07 ***
## previous_year_rating          0.319132   0.019262  16.568  < 2e-16 ***
## length_of_service             0.023459   0.006422   3.653 0.000259 ***
## KPIs_met..80.                 1.248670   0.044075  28.331  < 2e-16 ***
## awards_won.                   1.920807   0.079944  24.027  < 2e-16 ***
## avg_training_score            0.044217   0.001499  29.496  < 2e-16 ***
## education_bachelors           0.061774   0.162913   0.379 0.704552
## education_masters             0.287462   0.168880   1.702 0.088724 .
## education_BelowSecondary            NA         NA      NA       NA
## recruitment_channel_sourcing -0.164540   0.114855  -1.433 0.151975
## recruitment_channel_other    -0.186543   0.114000  -1.636 0.101769
## recruitment_channel_referred        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: 21638  on 36678  degrees of freedom
## Residual deviance: 18051  on 36667  degrees of freedom
## AIC: 18075
##
## Number of Fisher Scoring iterations: 6
```

```
model10 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
               +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
               +education_masters+education_BelowSecondary+recruitment_channel_sourcing
               +recruitment_channel_other+recruitment_channel_referred
               +gender_male+gender_female,
               data=HR_TR, family = binomial)
summary(model10)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won. + avg_training_score +
##     education_bachelors + education_masters + education_BelowSecondary +
##     recruitment_channel_sourcing + recruitment_channel_other +
##     recruitment_channel_referred + gender_male + gender_female,
##     family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8431  -0.4325  -0.2791  -0.1897   2.9488
##
## Coefficients: (3 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.316697   0.249960 -25.271  < 2e-16 ***
## no_of_trainings              -0.230784   0.039613  -5.826 5.68e-09 ***
## age                          -0.019470   0.003858  -5.047 4.49e-07 ***
## previous_year_rating          0.319266   0.019266  16.572  < 2e-16 ***
## length_of_service             0.023286   0.006425   3.624  0.00029 ***
## KPIs_met..80.                 1.247036   0.044097  28.279  < 2e-16 ***
## awards_won.                   1.920650   0.079970  24.017  < 2e-16 ***
## avg_training_score            0.044317   0.001503  29.476  < 2e-16 ***
## education_bachelors           0.064644   0.162962   0.397  0.69160
## education_masters             0.289758   0.168921   1.715  0.08628 .
## education_BelowSecondary            NA         NA      NA       NA
## recruitment_channel_sourcing -0.166711   0.114858  -1.451  0.14665
## recruitment_channel_other    -0.189155   0.114010  -1.659  0.09709 .
## recruitment_channel_referred        NA         NA      NA       NA
## gender_male                  -0.050322   0.042778  -1.176  0.23945
## gender_female                       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: 21638  on 36678  degrees of freedom
## Residual deviance: 18050  on 36666  degrees of freedom
## AIC: 18076
##
## Number of Fisher Scoring iterations: 6
```

```
#removing variables

model11 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
               +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
               +education_masters+education_BelowSecondary
               +gender_male+gender_female,
               data=HR_TR, family = binomial)
summary(model11)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won. + avg_training_score +
##     education_bachelors + education_masters + education_BelowSecondary +
##     gender_male + gender_female, family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8509  -0.4321  -0.2793  -0.1898   2.9449
##
## Coefficients: (2 not defined because of singularities)
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.485438   0.226123 -28.681  < 2e-16 ***
## no_of_trainings          -0.231734   0.039606  -5.851 4.89e-09 ***
## age                      -0.019631   0.003857  -5.090 3.58e-07 ***
## previous_year_rating      0.320386   0.019256  16.638  < 2e-16 ***
## length_of_service         0.023204   0.006425   3.612 0.000304 ***
## KPIs_met..80.             1.248023   0.044099  28.301  < 2e-16 ***
## awards_won.               1.918279   0.079939  23.997  < 2e-16 ***
## avg_training_score        0.044351   0.001503  29.502  < 2e-16 ***
## education_bachelors       0.059224   0.162851   0.364 0.716106
## education_masters         0.282208   0.168779   1.672 0.094513 .
## education_BelowSecondary        NA         NA      NA       NA
## gender_male              -0.048902   0.042767  -1.143 0.252847
## gender_female                   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: 21638  on 36678  degrees of freedom
## Residual deviance: 18053  on 36668  degrees of freedom
## AIC: 18075
##
## Number of Fisher Scoring iterations: 6
```

```
model12 <- glm(is_promoted ~no_of_trainings+previous_year_rating+length_of_service
               +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
               +education_masters+education_BelowSecondary
               +gender_male+gender_female,
               data=HR_TR, family = binomial)
summary(model12)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won. + avg_training_score +
##     education_bachelors + education_masters + education_BelowSecondary +
##     gender_male + gender_female, family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8447  -0.4327  -0.2799  -0.1894   2.9520
##
## Coefficients: (2 not defined because of singularities)
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.957957   0.206914 -33.627  < 2e-16 ***
## no_of_trainings          -0.221718   0.039479  -5.616 1.95e-08 ***
## previous_year_rating      0.319662   0.019223  16.629  < 2e-16 ***
## length_of_service         0.003324   0.005078   0.655    0.513
## KPIs_met..80.             1.242547   0.044046  28.210  < 2e-16 ***
## awards_won.               1.894892   0.079765  23.756  < 2e-16 ***
## avg_training_score        0.044846   0.001502  29.850  < 2e-16 ***
## education_bachelors      -0.052340   0.161395  -0.324    0.746
## education_masters         0.097977   0.164891   0.594    0.552
## education_BelowSecondary        NA         NA      NA       NA
## gender_male              -0.049705   0.042740  -1.163    0.245
## gender_female                   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: 21638  on 36678  degrees of freedom
## Residual deviance: 18080  on 36669  degrees of freedom
## AIC: 18100
##
## Number of Fisher Scoring iterations: 6
```

```
model13 <- glm(is_promoted ~age+previous_year_rating+length_of_service
               +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
               +education_masters+education_BelowSecondary
               +gender_male+gender_female,
               data=HR_TR, family = binomial)
summary(model13)
```

```
##
## Call:
## glm(formula = is_promoted ~ age + previous_year_rating + length_of_service +
##     KPIs_met..80. + awards_won. + avg_training_score + education_bachelors +
##     education_masters + education_BelowSecondary + gender_male +
##     gender_female, family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8190  -0.4316  -0.2796  -0.1910   2.9600
##
## Coefficients: (2 not defined because of singularities)
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.747420   0.221864 -30.412  < 2e-16 ***
## age                      -0.018452   0.003847  -4.796 1.62e-06 ***
## previous_year_rating      0.321473   0.019253  16.698  < 2e-16 ***
## length_of_service         0.023444   0.006416   3.654 0.000258 ***
## KPIs_met..80.             1.251705   0.044076  28.399  < 2e-16 ***
## awards_won.               1.926231   0.079794  24.140  < 2e-16 ***
## avg_training_score        0.043539   0.001496  29.102  < 2e-16 ***
## education_bachelors       0.056796   0.162750   0.349 0.727109
## education_masters         0.284588   0.168682   1.687 0.091578 .
## education_BelowSecondary        NA         NA      NA       NA
## gender_male              -0.069454   0.042634  -1.629 0.103297
## gender_female                   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: 21638  on 36678  degrees of freedom
## Residual deviance: 18091  on 36669  degrees of freedom
## AIC: 18111
##
## Number of Fisher Scoring iterations: 6
```

```
model14 <- glm(is_promoted ~no_of_trainings+previous_year_rating+length_of_service
               +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
               +education_masters+education_BelowSecondary,
               data=HR_TR, family = binomial)
summary(model14)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won. + avg_training_score +
##     education_bachelors + education_masters + education_BelowSecondary,
##     family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8512  -0.4332  -0.2797  -0.1898   2.9464
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.979562   0.206023 -33.878  < 2e-16 ***
## no_of_trainings          -0.225054   0.039393  -5.713 1.11e-08 ***
## previous_year_rating      0.319519   0.019219  16.625  < 2e-16 ***
## length_of_service         0.003478   0.005074   0.685    0.493
## KPIs_met..80.             1.244177   0.044024  28.262  < 2e-16 ***
## awards_won.               1.895080   0.079740  23.766  < 2e-16 ***
## avg_training_score        0.044746   0.001498  29.871  < 2e-16 ***
## education_bachelors      -0.055201   0.161345  -0.342    0.732
## education_masters         0.095611   0.164846   0.580    0.562
## education_BelowSecondary        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: 21638  on 36678  degrees of freedom
## Residual deviance: 18081  on 36670  degrees of freedom
## AIC: 18099
##
## Number of Fisher Scoring iterations: 6
```

```
model15 <- glm(is_promoted ~no_of_trainings+previous_year_rating+length_of_service
               +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
               +education_masters,
               data=HR_TR, family = binomial)
summary(model15)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won. + avg_training_score +
##     education_bachelors + education_masters, family = binomial,
##     data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8512  -0.4332  -0.2797  -0.1898   2.9464
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -6.979562   0.206023 -33.878  < 2e-16 ***
## no_of_trainings      -0.225054   0.039393  -5.713 1.11e-08 ***
## previous_year_rating  0.319519   0.019219  16.625  < 2e-16 ***
## length_of_service     0.003478   0.005074   0.685    0.493
## KPIs_met..80.         1.244177   0.044024  28.262  < 2e-16 ***
## awards_won.           1.895080   0.079740  23.766  < 2e-16 ***
## avg_training_score    0.044746   0.001498  29.871  < 2e-16 ***
## education_bachelors  -0.055201   0.161345  -0.342    0.732
## education_masters     0.095611   0.164846   0.580    0.562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21638  on 36678  degrees of freedom
## Residual deviance: 18081  on 36670  degrees of freedom
## AIC: 18099
##
## Number of Fisher Scoring iterations: 6
```

```
model16 <- glm(is_promoted ~no_of_trainings+previous_year_rating
               +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
               +education_masters,
               data=HR_TR, family = binomial)
summary(model16)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + previous_year_rating +
##     KPIs_met..80. + awards_won. + avg_training_score + education_bachelors +
##     education_masters, family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8528  -0.4341  -0.2794  -0.1906   2.9463
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -6.969271   0.205423 -33.927  < 2e-16 ***
## no_of_trainings      -0.226263   0.039356  -5.749 8.97e-09 ***
## previous_year_rating  0.320620   0.019156  16.738  < 2e-16 ***
## KPIs_met..80.         1.241429   0.043827  28.325  < 2e-16 ***
## awards_won.           1.891697   0.079570  23.774  < 2e-16 ***
## avg_training_score    0.044684   0.001495  29.895  < 2e-16 ***
## education_bachelors  -0.043918   0.160467  -0.274    0.784
## education_masters     0.114156   0.162564   0.702    0.483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21638  on 36678  degrees of freedom
## Residual deviance: 18082  on 36671  degrees of freedom
## AIC: 18098
##
## Number of Fisher Scoring iterations: 6
```

```
model17 <- glm(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
               +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
               +education_masters+education_BelowSecondary+recruitment_channel_sourcing
               +recruitment_channel_other+recruitment_channel_referred,
               data=HR_TR, family = binomial)
summary(model17)
```

```
##
## Call:
## glm(formula = is_promoted ~ no_of_trainings + age + previous_year_rating +
##     length_of_service + KPIs_met..80. + awards_won. + avg_training_score +
##     education_bachelors + education_masters + education_BelowSecondary +
##     recruitment_channel_sourcing + recruitment_channel_other +
##     recruitment_channel_referred, family = binomial, data = HR_TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8498  -0.4327  -0.2790  -0.1900   2.9430
##
## Coefficients: (2 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.340222   0.249135 -25.449  < 2e-16 ***
## no_of_trainings              -0.234208   0.039523  -5.926 3.11e-09 ***
## age                          -0.019493   0.003859  -5.051 4.39e-07 ***
## previous_year_rating          0.319132   0.019262  16.568  < 2e-16 ***
## length_of_service             0.023459   0.006422   3.653 0.000259 ***
## KPIs_met..80.                 1.248670   0.044075  28.331  < 2e-16 ***
## awards_won.                   1.920807   0.079944  24.027  < 2e-16 ***
## avg_training_score            0.044217   0.001499  29.496  < 2e-16 ***
## education_bachelors           0.061774   0.162913   0.379 0.704552
## education_masters             0.287462   0.168880   1.702 0.088724 .
## education_BelowSecondary            NA         NA      NA       NA
## recruitment_channel_sourcing -0.164540   0.114855  -1.433 0.151975
## recruitment_channel_other    -0.186543   0.114000  -1.636 0.101769
## recruitment_channel_referred        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: 21638  on 36678  degrees of freedom
## Residual deviance: 18051  on 36667  degrees of freedom
## AIC: 18075
##
## Number of Fisher Scoring iterations: 6
```

```
#so far model 11
#Logistic regression prediction
#prediction and confusion matrix
res <- predict(model11,HR_TS, type='response')
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
head(res)
```

```
##          2          6          8         10         12         13
## 0.05021090 0.06575706 0.03778852 0.14790082 0.09903903 0.07380697
```

```
head(HR_TS$is_promoted)
```

```
## [1] 0 0 0 0 1 0
```

```
table(Actualvalue=HR_TS$is_promoted, PredictedValue= res>0.2)
```

```
##            PredictedValue
## Actualvalue FALSE  TRUE
##           0 13131  1225
##           1   903   461
```

```
table(Actualvalue=HR_TS$is_promoted, PredictedValue= res>0.3)
```

```
##            PredictedValue
## Actualvalue FALSE  TRUE
##           0 13867   489
##           1  1084   280
```

```
table(Actualvalue=HR_TS$is_promoted, PredictedValue= res>0.4)
```

```
##            PredictedValue
## Actualvalue FALSE  TRUE
##           0 14241   115
##           1  1213   151
```

```
table(Actualvalue=HR_TS$is_promoted, PredictedValue= res>0.45)
```

```
##            PredictedValue
## Actualvalue FALSE  TRUE
##           0 14278    78
##           1  1243   121
```

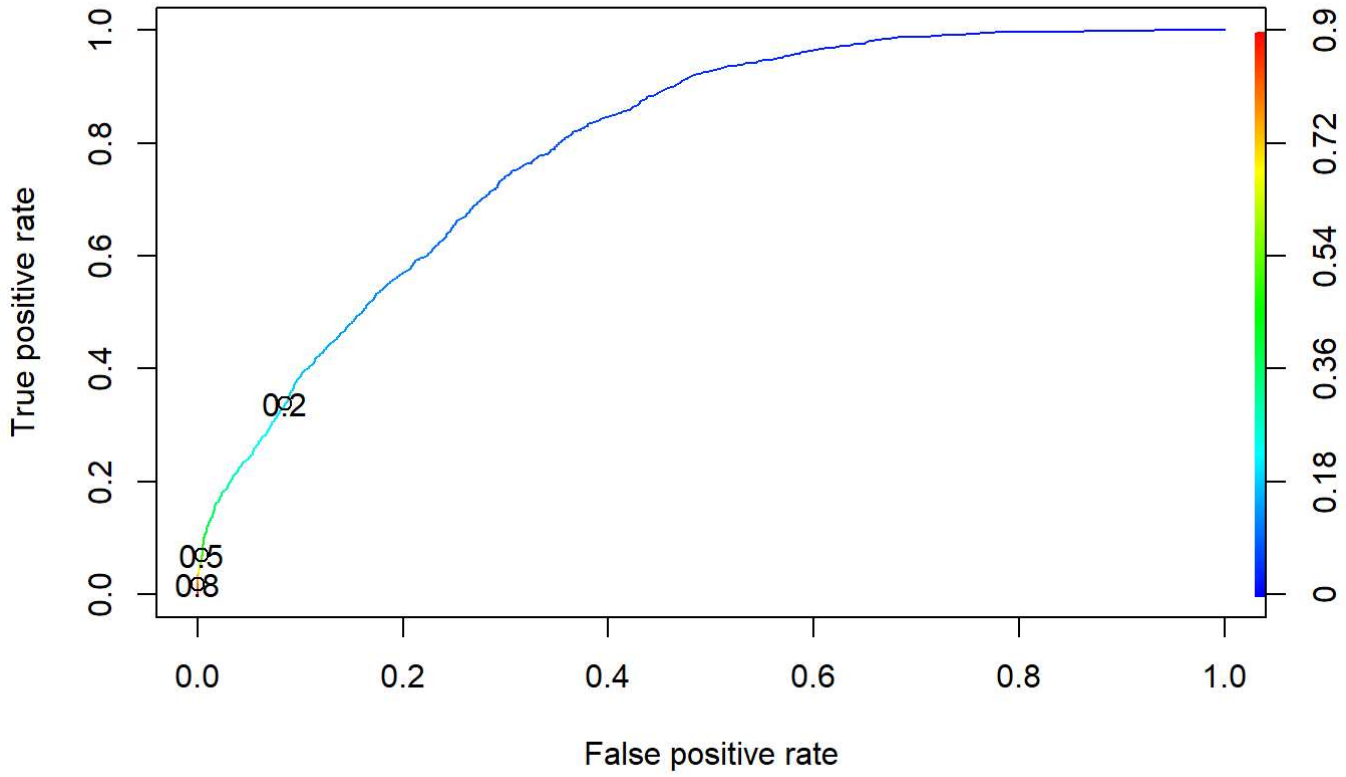
```
table(Actualvalue=HR_TS$is_promoted, PredictedValue= res>0.5)
```

```
##            PredictedValue
## Actualvalue FALSE  TRUE
##           0 14293    63
##           1  1270    94
```

```
table(Actualvalue=HR_TS$is_promoted, PredictedValue= res>0.6)
```

```
##            PredictedValue
## Actualvalue FALSE  TRUE
##           0 14321    35
##           1  1293    71
```

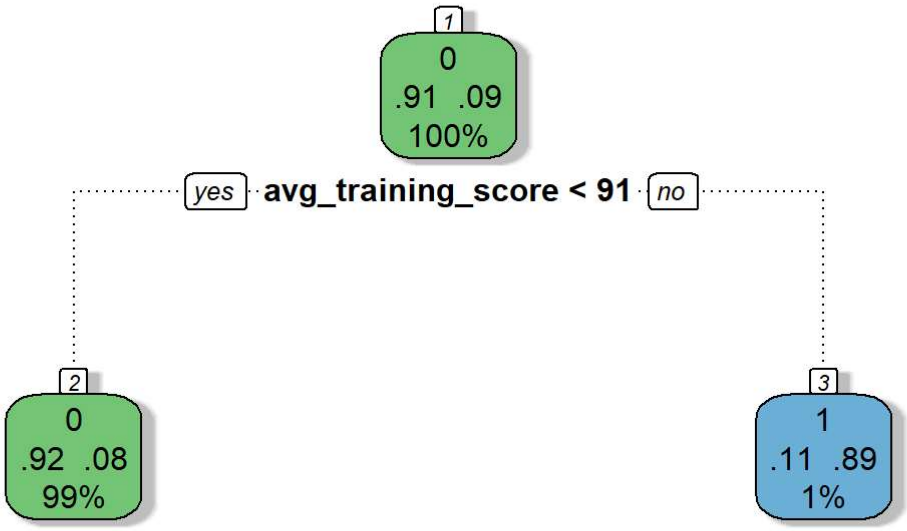
```
#highest accuracy 91.6%

#ROCR Curve

ROCRpred <- prediction(res,HR_TS$is_promoted)
ROCRpref <- performance(ROCRpred, "tpr","fpr")
plot(ROCRpref,colorize = TRUE, print.cutoffs.at=seq(0.2,by=0.3))
```




```
#Decision tree

#without any splitting criteria
model20 <- rpart(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
                 +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
                 +education_masters+education_BelowSecondary
                 +gender_male+gender_female,
                 data=HR_TR,
                 method = "class")
model20
```

```
## n= 36679
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 36679 3182 0 (0.91324736 0.08675264)
##   2) avg_training_score< 90.5 36295 2842 0 (0.92169720 0.07830280) *
##   3) avg_training_score>=90.5 384   44 1 (0.11458333 0.88541667) *
```

```
fancyRpartPlot(model20)
```



Rattle 2020-Dec-03 05:51:34 Souvik

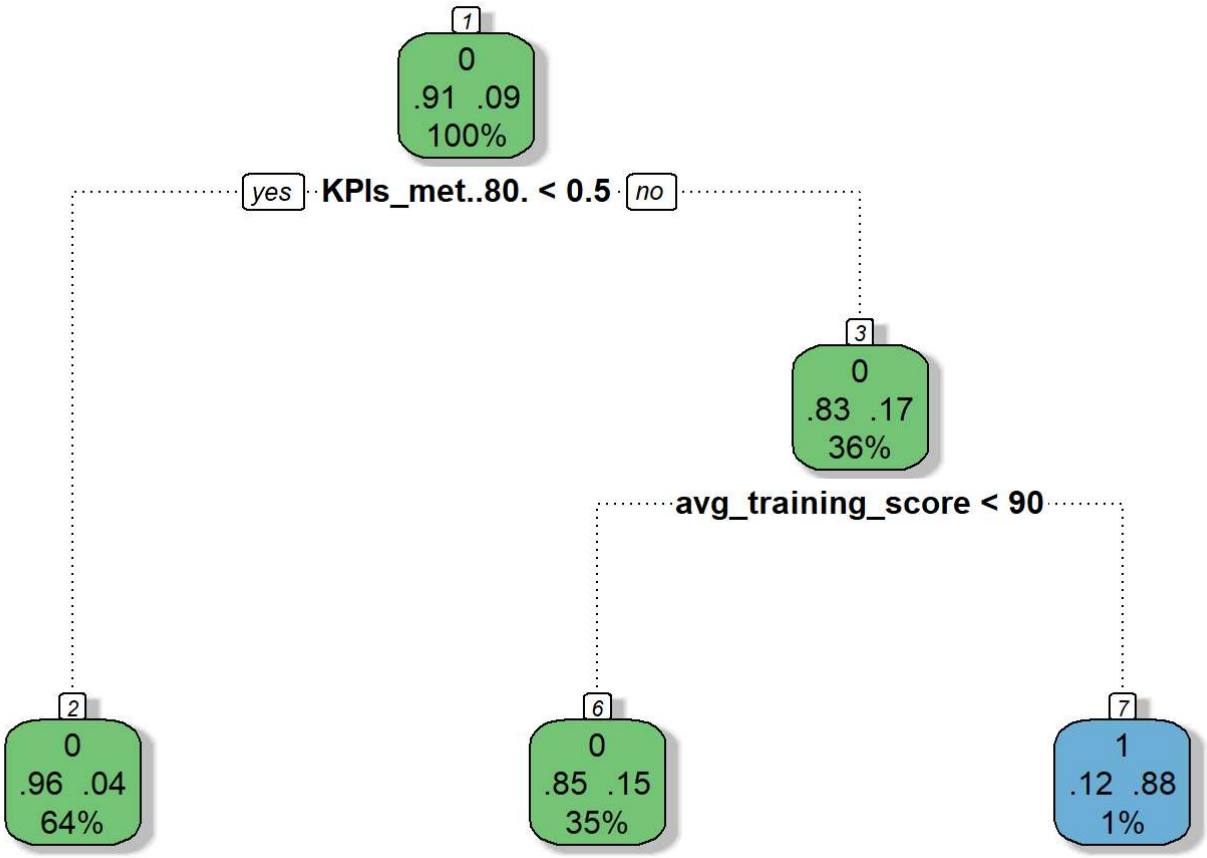
```
pred2 <- predict(model20, newdata = HR_TS, type = "class")
confusionMatrix(table(pred2,HR_TS$is_promoted))
```

```
## Confusion Matrix and Statistics
##
##
## pred2     0     1
##     0 14334  1210
##     1    22   154
##
##                Accuracy : 0.9216
##                  95% CI : (0.9173, 0.9258)
##     No Information Rate : 0.9132
##     P-Value [Acc > NIR] : 7.891e-05
##
##                   Kappa : 0.1838
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9985
##             Specificity : 0.1129
##          Pos Pred Value : 0.9222
##          Neg Pred Value : 0.8750
##              Prevalence : 0.9132
##          Detection Rate : 0.9118
##    Detection Prevalence : 0.9888
##       Balanced Accuracy : 0.5557
##
##        'Positive' Class : 0
##
```

```
#information gain
model18 <- rpart(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
               +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
               +education_masters+education_BelowSecondary
               +gender_male+gender_female,
               data=HR_TR,
               method = "class",
               parms = list(split = "information"))
model18
```

```
## n= 36679
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 36679 3182 0 (0.91324736 0.08675264)
##   2) KPIs_met..80.< 0.5 23562  958 0 (0.95934131 0.04065869) *
##   3) KPIs_met..80.>=0.5 13117 2224 0 (0.83044904 0.16955096)
##     6) avg_training_score< 89.5 12829 1971 0 (0.84636371 0.15363629) *
##     7) avg_training_score>=89.5 288   35 1 (0.12152778 0.87847222) *
```

```
fancyRpartPlot(model18)
```



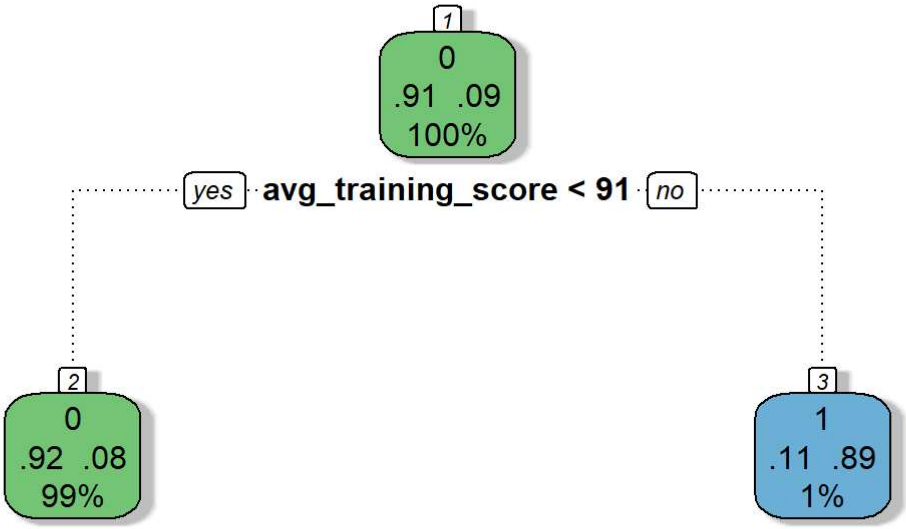

```
pred <- predict(model18, newdata = HR_TS, type = "class")
confusionMatrix(table(pred,HR_TS$is_promoted))
```

```
## Confusion Matrix and Statistics
##
##
## pred     0     1
##    0 14334  1244
##    1    22   120
##
##                Accuracy : 0.9195
##                  95% CI : (0.9151, 0.9237)
##     No Information Rate : 0.9132
##     P-Value [Acc > NIR] : 0.002641
##
##                   Kappa : 0.1454
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.99847
##             Specificity : 0.08798
##          Pos Pred Value : 0.92014
##          Neg Pred Value : 0.84507
##              Prevalence : 0.91323
##          Detection Rate : 0.91183
##    Detection Prevalence : 0.99097
##       Balanced Accuracy : 0.54322
##
##        'Positive' Class : 0
##
```

```
#gini index
model19 <- rpart(is_promoted ~ no_of_trainings+age+previous_year_rating+length_of_service
                 +KPIs_met..80.+awards_won.+avg_training_score+education_bachelors
                 +education_masters+education_BelowSecondary
                 +gender_male+gender_female,
                 data=HR_TR,
                 method = "class",
                 parms = list(split = "gini"))
model19
```

```
## n= 36679
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 36679 3182 0 (0.91324736 0.08675264)
##   2) avg_training_score< 90.5 36295 2842 0 (0.92169720 0.07830280) *
##   3) avg_training_score>=90.5 384   44 1 (0.11458333 0.88541667) *
```

```
fancyRpartPlot(model19)
```



Rattle 2020-Dec-03 05:51:38 Souvik

```
pred0 <- predict(model19, newdata = HR_TS, type = "class")
confusionMatrix(table(pred0,HR_TS$is_promoted))
```

```
## Confusion Matrix and Statistics
##
##
## pred0     0     1
##     0 14334  1210
##     1    22   154
##
##                Accuracy : 0.9216
##                  95% CI : (0.9173, 0.9258)
##     No Information Rate : 0.9132
##     P-Value [Acc > NIR] : 7.891e-05
##
##                   Kappa : 0.1838
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9985
##             Specificity : 0.1129
##          Pos Pred Value : 0.9222
##          Neg Pred Value : 0.8750
##              Prevalence : 0.9132
##          Detection Rate : 0.9118
##    Detection Prevalence : 0.9888
##       Balanced Accuracy : 0.5557
##
##        'Positive' Class : 0
##
```

```
##Running our finals models to predict the testing data

TrainHR <- read.csv("promotion_ts.csv", stringsAsFactors = TRUE)
summary(TrainHR)
```

```
##   employee_id                department         region
##  Min.   :    3   Sales & Marketing:7315   region_2 : 5299
##  1st Qu.:19370   Operations       :4764   region_22: 2739
##  Median :38964   Procurement      :3020   region_7 : 1982
##  Mean   :39041   Technology       :3011   region_13: 1167
##  3rd Qu.:58690   Analytics        :2319   region_15: 1130
##  Max.   :78295   Finance          :1091   region_26: 1011
##                  (Other)          :1970   (Other)  :10162
##             education     gender    recruitment_channel no_of_trainings
##                  : 1034   f: 6894   other   :13078      Min.   :1.000
##  Bachelor's      :15578   m:16596   referred:  451      1st Qu.:1.000
##  Below Secondary :  374             sourcing: 9961      Median :1.000
##  Master's & above: 6504                                 Mean   :1.254
##                                                         3rd Qu.:1.000
##                                                         Max.   :9.000
##
##       age        previous_year_rating length_of_service KPIs_met..80.
##  Min.   :20.00   Min.   :1.000        Min.   : 1.00     Min.   :0.0000
##  1st Qu.:29.00   1st Qu.:3.000        1st Qu.: 3.00     1st Qu.:0.0000
##  Median :33.00   Median :3.000        Median : 5.00     Median :0.0000
##  Mean   :34.78   Mean   :3.339        Mean   : 5.81     Mean   :0.3588
##  3rd Qu.:39.00   3rd Qu.:4.000        3rd Qu.: 7.00     3rd Qu.:1.0000
##  Max.   :60.00   Max.   :5.000        Max.   :34.00     Max.   :1.0000
##                  NA's   :1812
##   awards_won.      avg_training_score
##  Min.   :0.00000   Min.   :39.00
##  1st Qu.:0.00000   1st Qu.:51.00
##  Median :0.00000   Median :60.00
##  Mean   :0.02278   Mean   :63.26
##  3rd Qu.:0.00000   3rd Qu.:76.00
##  Max.   :1.00000   Max.   :99.00
##
```

```
#Cleaning the dataset by removing and imputing rows


#mode imputation on column previous year rating
table(TrainHR$previous_year_rating)
```

```
##
##    1    2    3    4    5
## 2680 1731 7921 4249 5097
```

```
TrainHR$previous_year_rating[is.na(TrainHR$previous_year_rating)] <- 3
table(TrainHR$previous_year_rating)
```

```
##
##    1    2    3    4    5
## 2680 1731 9733 4249 5097
```

```
summary(TrainHR$previous_year_rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.313   4.000   5.000
```

```
#NA imputing on empty cells on column education
TrainHR[which(TrainHR$education==""),]$education <-NA
summary(TrainHR$education)
```

```
##                        Bachelor's  Below Secondary Master's & above
##                 0             15578              374             6504
##             NA's
##             1034
```

```
#removing all NA rows from the dataset
TrainHR <- na.omit(TrainHR)
summary(TrainHR)
```

```
##   employee_id                 department         region
##  Min.   :    3   Sales & Marketing:6652   region_2 :4923
##  1st Qu.:19380   Operations       :4672   region_22:2594
##  Median :38936   Procurement      :2980   region_7 :1893
##  Mean   :39026   Technology       :2969   region_13:1144
##  3rd Qu.:58704   Analytics        :2164   region_15:1057
##  Max.   :78295   Finance          :1077   region_26: 969
##                  (Other)          :1942   (Other)  :9876
##             education     gender    recruitment_channel no_of_trainings
##                  :    0   f: 6715   other   :12499      Min.   :1.000
##  Bachelor's      :15578   m:15741   referred:  448      1st Qu.:1.000
##  Below Secondary :  374             sourcing: 9509      Median :1.000
##  Master's & above: 6504                                 Mean   :1.257
##                                                         3rd Qu.:1.000
##                                                         Max.   :9.000
##
##       age        previous_year_rating length_of_service KPIs_met..80.
##  Min.   :20.00   Min.   :1.000        Min.   : 1.000    Min.   :0.0000
##  1st Qu.:29.00   1st Qu.:3.000        1st Qu.: 3.000    1st Qu.:0.0000
##  Median :33.00   Median :3.000        Median : 5.000    Median :0.0000
##  Mean   :34.97   Mean   :3.323        Mean   : 5.878    Mean   :0.3629
##  3rd Qu.:39.00   3rd Qu.:4.000        3rd Qu.: 7.000    3rd Qu.:1.0000
##  Max.   :60.00   Max.   :5.000        Max.   :34.000    Max.   :1.0000
##
##   awards_won.     avg_training_score
##  Min.   :0.0000   Min.   :39.00
##  1st Qu.:0.0000   1st Qu.:51.00
##  Median :0.0000   Median :60.00
##  Mean   :0.0232   Mean   :63.48
##  3rd Qu.:0.0000   3rd Qu.:76.00
##  Max.   :1.0000   Max.   :99.00
##
```

```
#To create dummy variables
TrainHR$gender_male<- ifelse(TrainHR$gender == "m", 1,0)
TrainHR$gender_female<- ifelse(TrainHR$gender == "f", 1,0)
TrainHR$education_bachelors<- ifelse(TrainHR$education == "Bachelor's", 1,0)
TrainHR$education_masters<- ifelse(TrainHR$education == "Master's & above", 1,0)
TrainHR$education_BelowSecondary<- ifelse(TrainHR$education == "Below Secondary", 1,0)


#Logistic regression prediction
res1 <- predict(model11,TrainHR, type='response')
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
table(PredictedValue= res1>0.45)
```

```
## PredictedValue
## FALSE  TRUE
## 22194   262
```

```
#decision tree Gini index model prediction
pred1 <- predict(model19, newdata = TrainHR, type = "class")
table(pred1)
```

```
## pred1
##     0     1
## 22225   231
```

```
TrainHR$Promoted_by_LR <- (PredictedValue= res1>0.45)
TrainHR$Promoted_by_LR<- ifelse(TrainHR$Promoted_by_LR == "TRUE", 1,0)
TrainHR$Promoted_by_GINI <- pred1
table(TrainHR$Promoted_by_LR)
```

```
##
##     0     1
## 22194   262
```

```
table(TrainHR$Promoted_by_GINI)
```

```
##
##     0     1
## 22225   231
```

```
FINALDATA_PREDICTION <- write.csv(TrainHR , "FINALDATA_PREDICTION.csv")
```