Employee Attrition

```
#----Set Working Directory
setwd("D:/Capstone Projects/R")
getwd()
```

```
## [1] "D:/Capstone Projects/R"
```

```
#----Library----
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(psych)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------
---- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.0     v forcats 0.5.0
## v readr   1.3.1
```

```
## -- Conflicts ------------------------------------------------------------ t
idyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(ggplot2)
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(superheat)
library(corrplot)

## corrplot 0.84 loaded

library(readr)
library(plotrix)

##
## Attaching package: 'plotrix'

## The following object is masked from 'package:gplots':
##
##     plotCI

## The following object is masked from 'package:psych':
##
##     rescale

library(ggcorrplot)
library(purrr)
library(moments)
library(psych)
library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

library(DMwR)

## Loading required package: lattice

## Loading required package: grid

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:purrr':
##
##     some
```

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

## The following object is masked from 'package:dplyr':
##
##     recode

library(caret)

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

library(ResourceSelection)

## ResourceSelection 0.3-5   2019-07-22

library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

library(PRROC)
library(ROCR)
library(plotROC)

##
## Attaching package: 'plotROC'

## The following object is masked from 'package:pROC':
##
##     ggroc

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:psych':
##
##     outlier

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.0.3
```

```
library(caTools)
library(tree)
```

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```
library(rpart)
```

```
#----Importing data----
data1 = read.csv("HR-Employee-Attrition.csv", stringsAsFactors = T)
str(data1)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ ï..Age                  : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1
1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequ
ently",..: 3 2 3 2 3 2 3 3 2 3 ...
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 21
6 1299 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2
2 2 2 2 2 2 2 2 ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : int  2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2
5 2 4 2 4 2 2 4 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2
2 1 2 2 2 ...
```

```
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : int  3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative
",..: 8 7 3 7 3 3 3 3 5 1 ...
##  $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3
2 3 2 2 3 2 1 3 2 ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693
9526 5237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964
13335 8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ..
.
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2
1 1 1 ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : int  3 4 3 3 3 3 4 4 4 3 ...
##  $ RelationshipSatisfaction: int  1 4 2 3 4 3 1 2 2 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : int  1 3 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

```
head(data1,5)
```

```
##   ï..Age Attrition    BusinessTravel DailyRate              Department
## 1     41       Yes     Travel_Rarely      1102                   Sales
## 2     49        No Travel_Frequently       279 Research & Development
## 3     37       Yes     Travel_Rarely      1373 Research & Development
## 4     33        No Travel_Frequently      1392 Research & Development
## 5     27        No     Travel_Rarely       591 Research & Development
##   DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 1                1         2  Life Sciences             1              1
## 2                8         1  Life Sciences             1              2
## 3                2         2          Other             1              4
## 4                3         4  Life Sciences             1              5
## 5                2         1        Medical             1              7
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2 Female         94              3        2
## 2                       3   Male         61              2        2
## 3                       4   Male         92              2        1
## 4                       4 Female         56              3        1
## 5                       1   Male         40              3        1
```

```
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome Monthl
yRate
## 1       Sales Executive               4        Single          5993
19479
## 2    Research Scientist               2       Married          5130
24907
## 3 Laboratory Technician               3        Single          2090
2396
## 4    Research Scientist               3       Married          2909
23159
## 5 Laboratory Technician               2       Married          3468
16632
##   NumCompaniesWorked Over18 OverTime PercentSalaryHike PerformanceRating
## 1                  8      Y      Yes                11                 3
## 2                  1      Y       No                23                 4
## 3                  6      Y      Yes                15                 3
## 4                  1      Y      Yes                11                 3
## 5                  9      Y       No                12                 3
##   RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYear
s
## 1                        1           80               0
8
## 2                        4           80               1                1
0
## 3                        2           80               0
7
## 4                        3           80               0
8
## 5                        4           80               1
6
##   TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole
## 1                     0               1              6                  4
## 2                     3               3             10                  7
## 3                     3               3              0                  0
## 4                     3               3              8                  7
## 5                     3               3              2                  2
##   YearsSinceLastPromotion YearsWithCurrManager
## 1                       0                    5
## 2                       1                    7
## 3                       0                    0
## 4                       3                    0
## 5                       2                    2

#----NA value check----
apply(is.na(data1), 2, sum)

##                   ï..Age                 Attrition             BusinessTravel
##                        0                         0                          0
##                DailyRate                Department          DistanceFromHome
##                        0                         0                          0
```

```
##                Education          EducationField             EmployeeCount
##                        0                       0                         0
##           EmployeeNumber  EnvironmentSatisfaction                    Gender
##                        0                       0                         0
##               HourlyRate          JobInvolvement                  JobLevel
##                        0                       0                         0
##                  JobRole         JobSatisfaction             MaritalStatus
##                        0                       0                         0
##            MonthlyIncome             MonthlyRate        NumCompaniesWorked
##                        0                       0                         0
##                   Over18              OverTime         PercentSalaryHike
##                        0                       0                         0
##        PerformanceRating RelationshipSatisfaction             StandardHours
##                        0                       0                         0
##         StockOptionLevel       TotalWorkingYears     TrainingTimesLastYear
##                        0                       0                         0
##          WorkLifeBalance           YearsAtCompany        YearsInCurrentRole
##                        0                       0                         0
##  YearsSinceLastPromotion    YearsWithCurrManager
##                        0                       0

#----Descriptive analysis----
summary(data1)

##      ï..Age      Attrition            BusinessTravel   DailyRate
##  Min.   :18.00   No :1233   Non-Travel       : 150   Min.   : 102.0
##  1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   1st Qu.: 465.0
##  Median :36.00              Travel_Rarely    :1043   Median : 802.0
##  Mean   :36.92                                       Mean   : 802.5
##  3rd Qu.:43.00                                       3rd Qu.:1157.0
##  Max.   :60.00                                       Max.   :1499.0
##
##                   Department  DistanceFromHome   Education
##  Human Resources       : 63   Min.   : 1.000   Min.   :1.000
##  Research & Development:961   1st Qu.: 2.000   1st Qu.:2.000
##  Sales                 :446   Median : 7.000   Median :3.000
##                               Mean   : 9.193   Mean   :2.913
##                               3rd Qu.:14.000   3rd Qu.:4.000
##                               Max.   :29.000   Max.   :5.000
##
##           EducationField EmployeeCount EmployeeNumber   EnvironmentSatisfa
ction
##  Human Resources : 27    Min.   :1     Min.   :   1.0   Min.   :1.000
##  Life Sciences   :606    1st Qu.:1     1st Qu.: 491.2   1st Qu.:2.000
##  Marketing       :159    Median :1     Median :1020.5   Median :3.000
##  Medical         :464    Mean   :1     Mean   :1024.9   Mean   :2.722
##  Other           : 82    3rd Qu.:1     3rd Qu.:1555.8   3rd Qu.:4.000
##  Technical Degree:132    Max.   :1     Max.   :2068.0   Max.   :4.000
##
##     Gender      HourlyRate     JobInvolvement    JobLevel
```

```
##  Female:588   Min.   : 30.00   Min.   :1.00   Min.   :1.000
##  Male  :882   1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000
##               Median : 66.00   Median :3.00   Median :2.000
##               Mean   : 65.89   Mean   :2.73   Mean   :2.064
##               3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000
##               Max.   :100.00   Max.   :4.00   Max.   :5.000
##
##                      JobRole    JobSatisfaction  MaritalStatus MonthlyInc
ome
##  Sales Executive          :326   Min.   :1.000   Divorced:327   Min.   : 1
009
##  Research Scientist       :292   1st Qu.:2.000   Married :673   1st Qu.: 2
911
##  Laboratory Technician    :259   Median :3.000   Single  :470   Median : 4
919
##  Manufacturing Director   :145   Mean   :2.729                  Mean   : 6
503
##  Healthcare Representative:131   3rd Qu.:4.000                  3rd Qu.: 8
379
##  Manager                  :102   Max.   :4.000                  Max.   :19
999
##  (Other)                  :215
##   MonthlyRate    NumCompaniesWorked Over18   OverTime   PercentSalaryHike
##  Min.   : 2094   Min.   :0.000      Y:1470   No :1054   Min.   :11.00
##  1st Qu.: 8047   1st Qu.:1.000               Yes: 416   1st Qu.:12.00
##  Median :14236   Median :2.000                          Median :14.00
##  Mean   :14313   Mean   :2.693                          Mean   :15.21
##  3rd Qu.:20462   3rd Qu.:4.000                          3rd Qu.:18.00
##  Max.   :26999   Max.   :9.000                          Max.   :25.00
##
##  PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Min.   :3.000     Min.   :1.000            Min.   :80    Min.   :0.0000
##  1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000
##  Median :3.000     Median :3.000            Median :80    Median :1.0000
##  Mean   :3.154     Mean   :2.712            Mean   :80    Mean   :0.7939
##  3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000
##  Max.   :4.000     Max.   :4.000            Max.   :80    Max.   :3.0000
##
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
##  Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
##  1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
##  Median :10.00     Median :3.000         Median :3.000   Median : 5.000
##  Mean   :11.28     Mean   :2.799         Mean   :2.761   Mean   : 7.008
##  3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000
##  Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000
##
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
##  1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
##  Median : 3.000     Median : 1.000          Median : 3.000
```

```
##  Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
##  3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
##  Max.   :18.000     Max.   :15.000          Max.   :17.000
##
```

```
describe(data1)
```

```
##                          vars    n     mean      sd  median  trimmed     m
ad
## ï..Age                      1 1470    36.92    9.14    36.0    36.47    8.
90
## Attrition*                  2 1470     1.16    0.37     1.0     1.08    0.
00
## BusinessTravel*             3 1470     2.61    0.67     3.0     2.76    0.
00
## DailyRate                   4 1470   802.49  403.51   802.0   803.83  510.
01
## Department*                 5 1470     2.26    0.53     2.0     2.25    0.
00
## DistanceFromHome            6 1470     9.19    8.11     7.0     8.08    7.
41
## Education                   7 1470     2.91    1.02     3.0     2.98    1.
48
## EducationField*             8 1470     3.25    1.33     3.0     3.10    1.
48
## EmployeeCount               9 1470     1.00    0.00     1.0     1.00    0.
00
## EmployeeNumber             10 1470  1024.87  602.02  1020.5  1023.40  790.
97
## EnvironmentSatisfaction    11 1470     2.72    1.09     3.0     2.78    1.
48
## Gender*                    12 1470     1.60    0.49     2.0     1.62    0.
00
## HourlyRate                 13 1470    65.89   20.33    66.0    66.02   26.
69
## JobInvolvement             14 1470     2.73    0.71     3.0     2.74    0.
00
## JobLevel                   15 1470     2.06    1.11     2.0     1.90    1.
48
## JobRole*                   16 1470     5.46    2.46     6.0     5.61    2.
97
## JobSatisfaction            17 1470     2.73    1.10     3.0     2.79    1.
48
## MaritalStatus*             18 1470     2.10    0.73     2.0     2.12    1.
48
## MonthlyIncome              19 1470  6502.93 4707.96  4919.0  5667.24 3260.
24
## MonthlyRate                20 1470 14313.10 7117.79 14235.5 14286.48 9201.
76
## NumCompaniesWorked         21 1470     2.69    2.50     2.0     2.36    1.
```

```
48
## Over18*                     22 1470     1.00    0.00     1.0     1.00    0.
00
## OverTime*                   23 1470     1.28    0.45     1.0     1.23    0.
00
## PercentSalaryHike           24 1470    15.21    3.66    14.0    14.80    2.
97
## PerformanceRating           25 1470     3.15    0.36     3.0     3.07    0.
00
## RelationshipSatisfaction    26 1470     2.71    1.08     3.0     2.77    1.
48
## StandardHours               27 1470    80.00    0.00    80.0    80.00    0.
00
## StockOptionLevel            28 1470     0.79    0.85     1.0     0.67    1.
48
## TotalWorkingYears           29 1470    11.28    7.78    10.0    10.37    5.
93
## TrainingTimesLastYear       30 1470     2.80    1.29     3.0     2.72    1.
48
## WorkLifeBalance             31 1470     2.76    0.71     3.0     2.77    0.
00
## YearsAtCompany              32 1470     7.01    6.13     5.0     5.99    4.
45
## YearsInCurrentRole          33 1470     4.23    3.62     3.0     3.85    4.
45
## YearsSinceLastPromotion     34 1470     2.19    3.22     1.0     1.48    1.
48
## YearsWithCurrManager        35 1470     4.12    3.57     3.0     3.77    4.
45
##                           min   max range  skew kurtosis     se
## ï..Age                     18    60    42  0.41    -0.41   0.24
## Attrition*                  1     2     1  1.84     1.39   0.01
## BusinessTravel*             1     3     2 -1.44     0.69   0.02
## DailyRate                 102  1499  1397  0.00    -1.21  10.52
## Department*                 1     3     2  0.17    -0.40   0.01
## DistanceFromHome            1    29    28  0.96    -0.23   0.21
## Education                   1     5     4 -0.29    -0.56   0.03
## EducationField*             1     6     5  0.55    -0.69   0.03
## EmployeeCount               1     1     0   NaN      NaN   0.00
## EmployeeNumber              1  2068  2067  0.02    -1.23  15.70
## EnvironmentSatisfaction     1     4     3 -0.32    -1.20   0.03
## Gender*                     1     2     1 -0.41    -1.83   0.01
## HourlyRate                 30   100    70 -0.03    -1.20   0.53
## JobInvolvement              1     4     3 -0.50     0.26   0.02
## JobLevel                    1     5     4  1.02     0.39   0.03
## JobRole*                    1     9     8 -0.36    -1.20   0.06
## JobSatisfaction             1     4     3 -0.33    -1.22   0.03
## MaritalStatus*              1     3     2 -0.15    -1.12   0.02
## MonthlyIncome            1009 19999 18990  1.37     0.99 122.79
## MonthlyRate              2094 26999 24905  0.02    -1.22 185.65
```

```
## NumCompaniesWorked          0     9     9  1.02     0.00   0.07
## Over18*                     1     1     0   NaN      NaN   0.00
## OverTime*                   1     2     1  0.96    -1.07   0.01
## PercentSalaryHike          11    25    14  0.82    -0.31   0.10
## PerformanceRating           3     4     1  1.92     1.68   0.01
## RelationshipSatisfaction    1     4     3 -0.30    -1.19   0.03
## StandardHours              80    80     0   NaN      NaN   0.00
## StockOptionLevel            0     3     3  0.97     0.35   0.02
## TotalWorkingYears           0    40    40  1.11     0.91   0.20
## TrainingTimesLastYear       0     6     6  0.55     0.48   0.03
## WorkLifeBalance             1     4     3 -0.55     0.41   0.02
## YearsAtCompany              0    40    40  1.76     3.91   0.16
## YearsInCurrentRole          0    18    18  0.92     0.47   0.09
## YearsSinceLastPromotion     0    15    15  1.98     3.59   0.08
## YearsWithCurrManager        0    17    17  0.83     0.16   0.09
```

comments: Out of the 35 variables we have 34 independent variables and one dependent/target variable which is Attrition !!!

```
#----Data Wrangling and cleaning
data1 <- data1 %>%
  mutate(Education = as.factor(if_else(Education == 1,"Below College", if_els
e(Education == 2, "College", if_else(Education == 3, "Bachelor", if_else(Educ
ation == 4, "Master","Doctor")))))
         ,EnvironmentSatisfaction = as.factor(if_else(EnvironmentSatisfaction
== 1,"Low",if_else(EnvironmentSatisfaction == 2, "Medium", if_else(Environmen
tSatisfaction == 3, "High", "Very High"))))
         ,JobInvolvement = as.factor(if_else(JobInvolvement == 1,"Low",if_els
e(JobInvolvement == 2, "Medium",if_else(JobInvolvement == 3, "High", "Very Hi
gh"))))
         ,JobSatisfaction = as.factor(if_else(JobSatisfaction == 1, "Low",if_
else(JobSatisfaction == 2, "Medium",if_else(JobSatisfaction == 3, "High","Ver
y High"))))
         ,PerformanceRating = as.factor(if_else(PerformanceRating == 1, "Low"
,if_else(PerformanceRating == 2, "Good", if_else(PerformanceRating == 3, "Exc
ellent", "Outstanding"))))
         ,RelationshipSatisfaction = as.factor(if_else(RelationshipSatisfacti
on == 1, "Low",if_else(RelationshipSatisfaction == 2, "Medium", if_else(Relat
ionshipSatisfaction == 3, "High", "Very High"))))
         ,WorkLifeBalance = as.factor(if_else(WorkLifeBalance == 1, "Bad",if_
else(WorkLifeBalance == 2, "Good", if_else(WorkLifeBalance == 3, "Better", "B
est")))),
         JobLevel = as.factor(JobLevel)
  )

#Removing unique values, no contribution in analysis
data2 <-  select(data1, -c("EmployeeCount", "EmployeeNumber",
                           "Over18", "StandardHours"))

#converting numeric to categorical
```

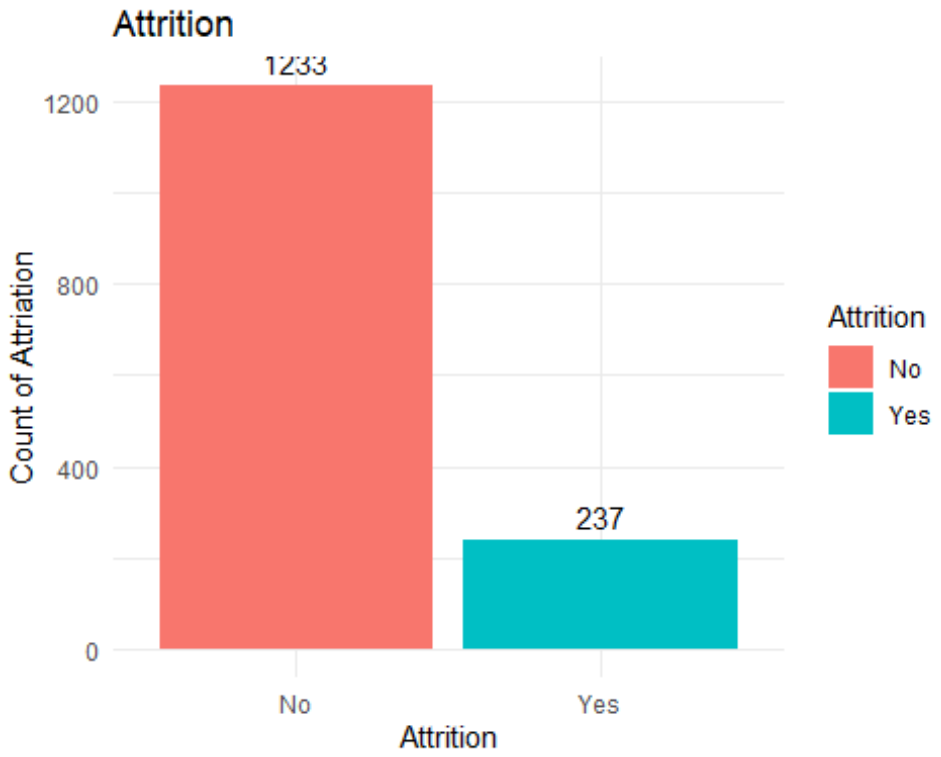
```r
data2$Education <- factor(data2$Education)
data2$EnvironmentSatisfaction <- factor(data2$EnvironmentSatisfaction)
data2$JobInvolvement <- factor(data2$JobInvolvement)
data2$JobLevel <- factor(data2$JobLevel)
data2$JobSatisfaction <- factor(data2$JobSatisfaction)
data2$PerformanceRating <- factor(data2$PerformanceRating)
data2$RelationshipSatisfaction <- factor(data2$RelationshipSatisfaction)
data2$StockOptionLevel <- factor(data2$StockOptionLevel)
data2$WorkLifeBalance <- factor(data2$WorkLifeBalance)

#Percentage of attrition----
d <- as.data.frame(table(data2$Attrition))
d
```

```
##   Var1 Freq
## 1   No 1233
## 2  Yes  237
```

```r
attrition_rate <- round((d[2,2] / sum(d$Freq))*100, 2)
print(attrition_rate)
```

```
## [1] 16.12
```

```r
data2 %>%
  group_by(Attrition) %>%
  tally() %>%
  ggplot(aes(x = Attrition, y = n,fill=Attrition)) +
  geom_bar(stat = "identity") +
  theme_minimal()+
  labs(x="Attrition", y="Count of Attriation")+
  ggtitle("Attrition")+
  geom_text(aes(label = n), vjust = -0.5, position = position_dodge(0.9))
```




```r
names(data2)[1] <- "Age"

library(janitor)

## Warning: package 'janitor' was built under R version 4.0.3

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test

library(CGPfunctions)

## Warning: package 'CGPfunctions' was built under R version 4.0.3

## Registered S3 method overwritten by 'DescTools':
##   method         from
##   reorder.factor gdata

## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```r
tabyl(data2, Gender, Attrition) %>%
  adorn_percentages("col") %>%
  adorn_pct_formatting(digits = 2)
```

```
##  Gender     No    Yes
##  Female 40.63% 36.71%
##    Male 59.37% 63.29%
```

```r
#Removing variables with no explanation of categories
data3 <-  select(data2, -c("StockOptionLevel", "JobLevel"))

data3$Department <- gsub("Human Resources", "HR", x = data3$Department)
data3$Department <- gsub("Research & Development", "R&D", x = data3$Departmen
t)

table(data3$Department)
```

```
##
##    HR   R&D Sales
##    63   961   446
```

```r
data_cat <- select_if(data3, is.factor)
str(data_cat)
```

```
## 'data.frame':    1470 obs. of  14 variables:
##  $ Attrition              : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1
1 1 1 ...
##  $ BusinessTravel         : Factor w/ 3 levels "Non-Travel","Travel_Frequ
ently",..: 3 2 3 2 3 2 3 3 2 3 ...
##  $ Education              : Factor w/ 5 levels "Bachelor","Below College"
,..: 3 2 3 5 2 3 1 2 1 1 ...
##  $ EducationField         : Factor w/ 6 levels "Human Resources",..: 2 2
5 2 4 2 4 2 2 4 ...
##  $ EnvironmentSatisfaction : Factor w/ 4 levels "High","Low","Medium",..:
3 1 4 4 2 4 1 4 4 1 ...
##  $ Gender                 : Factor w/ 2 levels "Female","Male": 1 2 2 1 2
2 1 2 2 2 ...
##  $ JobInvolvement         : Factor w/ 4 levels "High","Low","Medium",..:
1 3 3 1 1 1 4 1 3 1 ...
##  $ JobRole                : Factor w/ 9 levels "Healthcare Representative
",..: 8 7 3 7 3 3 3 3 5 1 ...
##  $ JobSatisfaction        : Factor w/ 4 levels "High","Low","Medium",..:
4 3 1 1 3 4 2 1 1 1 ...
##  $ MaritalStatus          : Factor w/ 3 levels "Divorced","Married",..: 3
2 3 2 2 3 2 1 3 2 ...
##  $ OverTime               : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2
1 1 1 ...
##  $ PerformanceRating      : Factor w/ 2 levels "Excellent","Outstanding":
1 2 1 1 1 1 2 2 2 1 ...
##  $ RelationshipSatisfaction: Factor w/ 4 levels "High","Low","Medium",..:
2 4 3 1 4 1 2 3 3 3 ...
```

```
##  $ WorkLifeBalance         : Factor w/ 4 levels "Bad","Best","Better",..:
1 3 3 3 3 4 4 3 3 4 ...

Attrition <- data_cat[,1]

data_num <- select_if(data3, is.numeric)
str(data_num)

## 'data.frame':    1470 obs. of  14 variables:
##  $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216
1299 ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9
526 5237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 1
3335 8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion: int  0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...

#Skewness Removal

length(data_num)

## [1] 14

data_num1 <- data_num
print(paste(colnames(data_num1)," ",skew(data_num1)))

##  [1] "Age   0.412443242937095"
##  [2] "DailyRate   -0.00351139085716317"
##  [3] "DistanceFromHome   0.95616353958535"
##  [4] "HourlyRate   -0.032245042085333"
##  [5] "MonthlyIncome   1.36702240441298"
##  [6] "MonthlyRate   0.0185399111919867"
##  [7] "NumCompaniesWorked   1.02437722300455"
##  [8] "PercentSalaryHike   0.81945296418638"
##  [9] "TotalWorkingYears   1.11489294424365"
## [10] "TrainingTimesLastYear   0.55199585815352"
## [11] "YearsAtCompany   1.76093000696624"
## [12] "YearsInCurrentRole   0.915491835640667"
## [13] "YearsSinceLastPromotion   1.9802422484982"
## [14] "YearsWithCurrManager   0.831750842992496"
```

```r
for (i in c(1:length(data_num1))){
  if(skew(data_num1[i]) > 0.75){
    data_num1[[i]] <- log1p(data_num1[[i]])
  }
}
print(paste(colnames(data_num1)," ",skew(data_num1)))
```

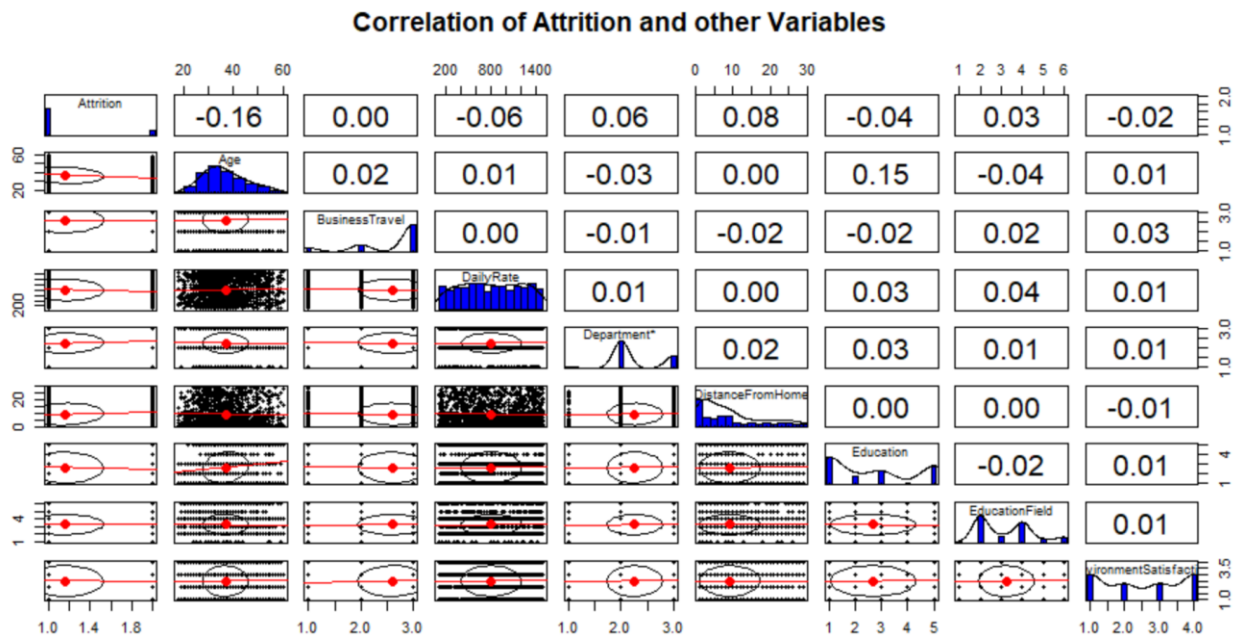
```
##  [1] "Age   0.412443242937095"
##  [2] "DailyRate   -0.00351139085716317"
##  [3] "DistanceFromHome   -0.0290613821426445"
##  [4] "HourlyRate   -0.032245042085333"
##  [5] "MonthlyIncome   0.285864052903371"
##  [6] "MonthlyRate   0.0185399111919867"
##  [7] "NumCompaniesWorked   0.0927067264493256"
##  [8] "PercentSalaryHike   0.512494950819636"
##  [9] "TotalWorkingYears   -0.620905696542672"
## [10] "TrainingTimesLastYear   0.55199585815352"
## [11] "YearsAtCompany   -0.207284251461488"
## [12] "YearsInCurrentRole   -0.382715446808629"
## [13] "YearsSinceLastPromotion   0.717338269762134"
## [14] "YearsWithCurrManager   -0.356956018688433"
```

```r
str(data3)
```

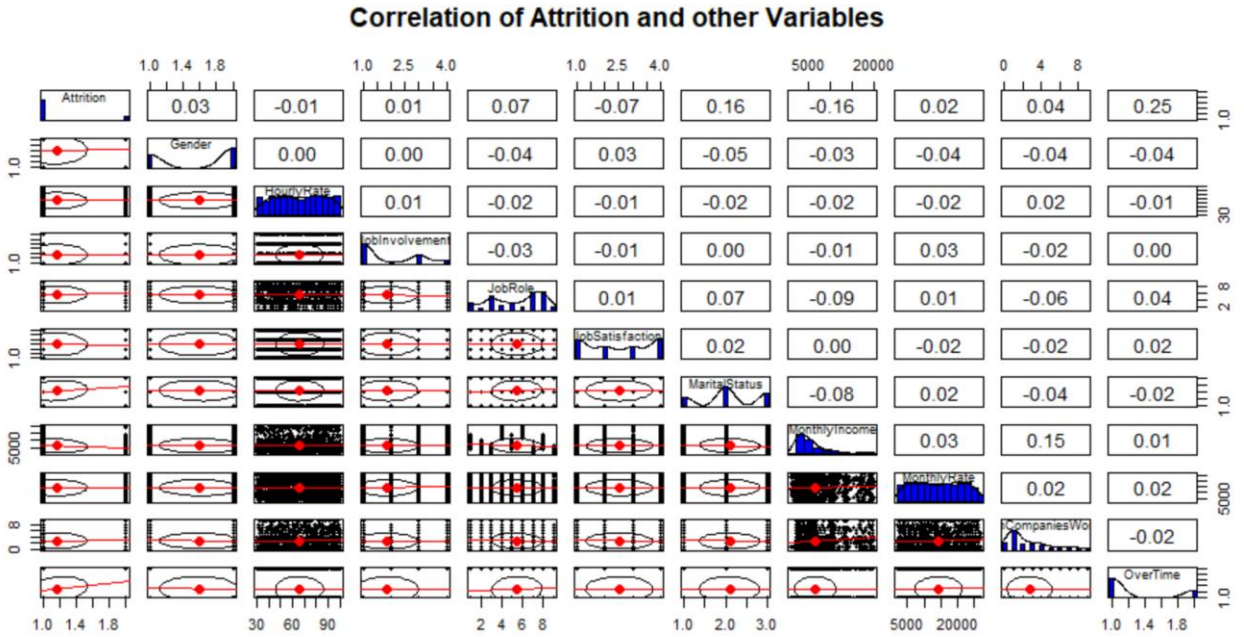
```
## 'data.frame':    1470 obs. of  29 variables:
##  $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1
1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequ
ently",..: 3 2 3 2 3 2 3 3 2 3 ...
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 21
6 1299 ...
##  $ Department              : chr  "Sales" "R&D" "R&D" "R&D" ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : Factor w/ 5 levels "Bachelor","Below College"
,..: 3 2 3 5 2 3 1 2 1 1 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2
5 2 4 2 4 2 2 4 ...
##  $ EnvironmentSatisfaction : Factor w/ 4 levels "High","Low","Medium",..:
3 1 4 4 2 4 1 4 4 1 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2
2 1 2 2 2 ...
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : Factor w/ 4 levels "High","Low","Medium",..:
1 3 3 1 1 1 4 1 3 1 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative
",..: 8 7 3 7 3 3 3 3 5 1 ...
##  $ JobSatisfaction         : Factor w/ 4 levels "High","Low","Medium",..:
4 3 1 1 3 4 2 1 1 1 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3
```

```
2 3 2 2 3 2 1 3 2 ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693
9526 5237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964
13335 8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2
1 1 1 ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : Factor w/ 2 levels "Excellent","Outstanding":
1 2 1 1 1 1 2 2 2 1 ...
##  $ RelationshipSatisfaction: Factor w/ 4 levels "High","Low","Medium",..:
2 4 3 1 4 1 2 3 3 3 ...
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : Factor w/ 4 levels "Bad","Best","Better",..:
1 3 3 3 3 4 4 3 3 4 ...
##  $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

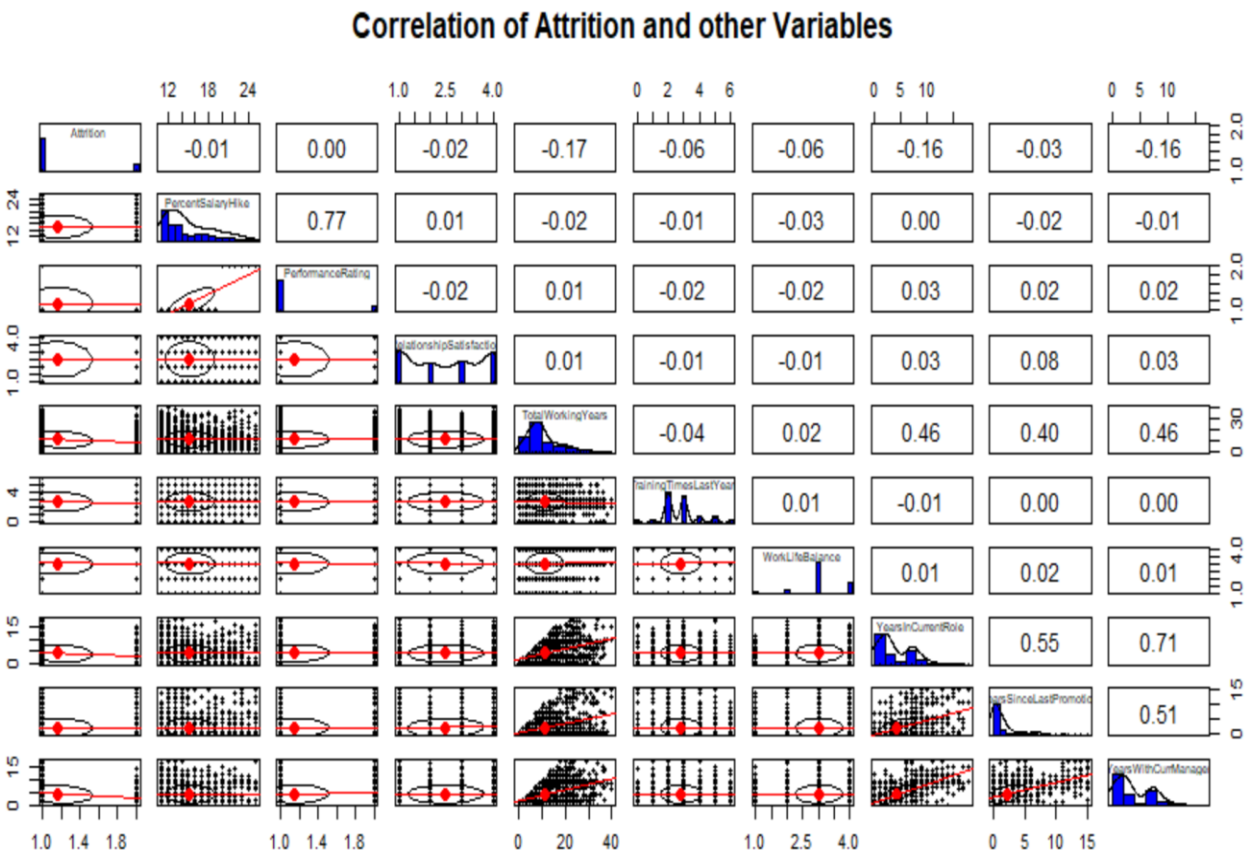
```r
pairs.panels(data3[, c(2,1,3:9)],
             method = "pearson", #coorelation method
             hist.col = "blue",
             main="Correlation of Attrition and other Variables",
             density = TRUE, # show density plots
             ellipses = TRUE, # show correlation ellipses
             lm=TRUE, #linear regression fits
             cex.cor = 5,
             cex.labels = 0.85
)
```




```
pairs.panels(data3[, c(2,10:19)],
             method = "pearson", #coorelation method
             hist.col = "blue",
             main="Correlation of Attrition and other Variables",
             density = TRUE, # show density plots
             ellipses = TRUE, # show correlation ellipses
             lm=TRUE, #linear regression fits
             cex.cor = 5,
             cex.labels = 0.85
)
```



```
pairs.panels(data3[, c(2,20:29)],
             method = "pearson", #coorelation method
             hist.col = "blue",
             main="Correlation of Attrition and other Variables",
             density = TRUE, # show density plots
             ellipses = TRUE, # show correlation ellipses
             lm=TRUE, #linear regression fits
             cex.cor = 4,
             cex.labels = 0.7
)
```




```
library(caret)
prep_num = preProcess(data_num1, method=c("center", "scale"))
data_num2 = predict(prep_num, data_num1)

cor_mat<- cor(data_num2)
high_corr <- findCorrelation(cor_mat, cutoff = 0.8)
names(data_num1)[high_corr]

## [1] "YearsAtCompany"
```

```r
#Removing highly correlated variable
data3 <-  select(data3, -c("YearsAtCompany"))
dim(data3)
```

```
## [1] 1470   28
```

```r
#----Outlier Detection----
lower_bound <- quantile(data_num$Attrition, 0.025)
upper_bound <- quantile(data_num$Attrition, 0.975)
outlier_ind <- which(data_num$Attrition < lower_bound | data_num$Attrition >
upper_bound)
data_num[outlier_ind,]
```

```
##  [1] Age                    DailyRate              DistanceFromHome
##  [4] HourlyRate             MonthlyIncome          MonthlyRate
##  [7] NumCompaniesWorked     PercentSalaryHike      TotalWorkingYears
## [10] TrainingTimesLastYear  YearsAtCompany         YearsInCurrentRole
## [13] YearsSinceLastPromotion YearsWithCurrManager
## <0 rows> (or 0-length row.names)
```

```r
dev.off()
```

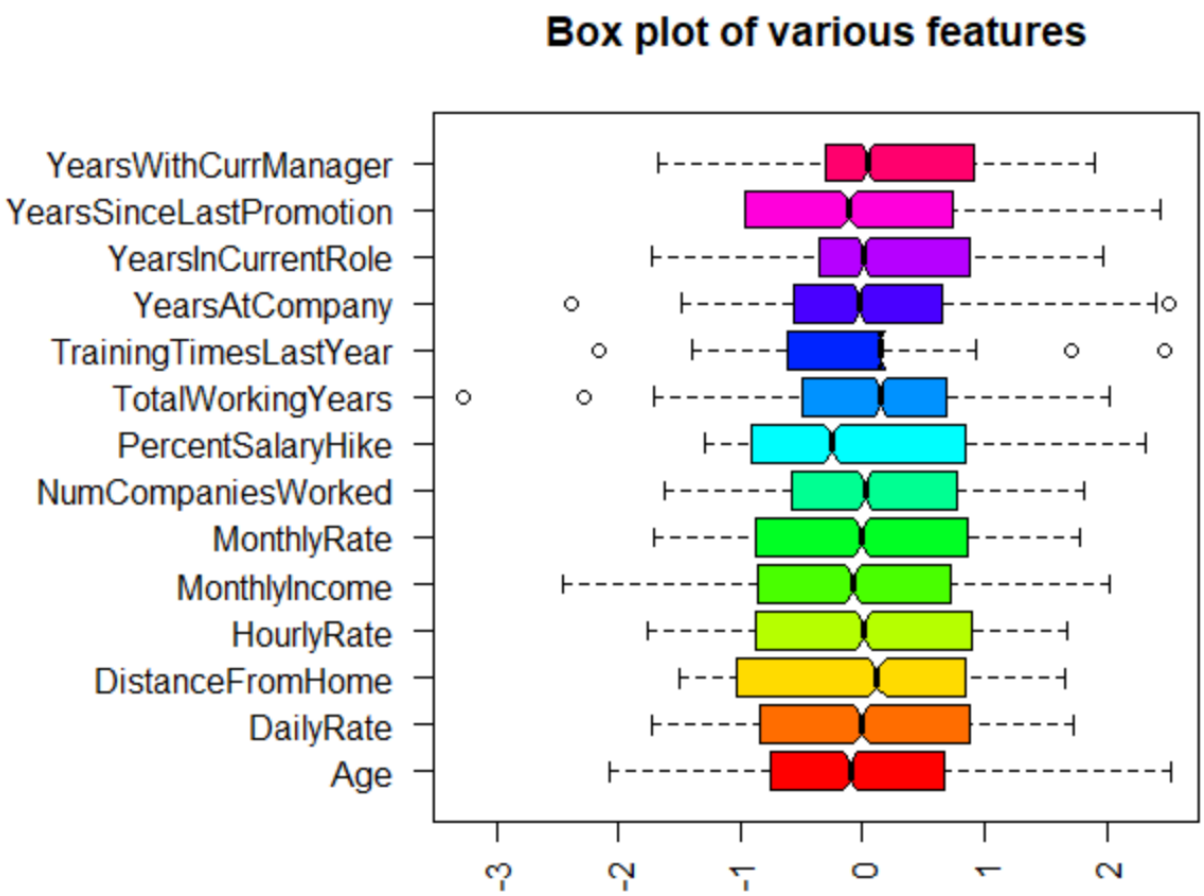
```
## null device
##           1
```

```r
par(mfrow=c(1,2))
dim(data_num2)
```

```
## [1] 1470   14
```

```r
plot.new()
boxplot(data_num2, main = "Box plot of various features", notch = T, col = ra
inbow(14),
        horizontal = T, las = 2, cex.axis=1, cex.names = 0., boxwe1x = 0.5)
```

```
## Warning in bxp(list(stats = structure(c(-2.07148722997616, -0.757912036332
496, :
## some notches went outside hinges ('box'): maybe set notch=FALSE
```




```
library(outliers)
```

```
##
## Attaching package: 'outliers'
```

```
## The following object is masked from 'package:randomForest':
##
##     outlier
```

```
## The following object is masked from 'package:psych':
##
##     outlier
```

```
outlier_scores <- scores(data_num2)
is_outlier <- outlier_scores > 3 | outlier_scores < -3
sum(is_outlier)
```

```
## [1] 11
```

```
# Outlier treatment using IQR
data_num3 <- data_num2
```

```r
livingar_lower <- (quantile(data_num3$YearsAtCompany, 0.25)) - 1.5*IQR(data_num3$YearsAtCompany)
livingar_upper <- (quantile(data_num3$YearsAtCompany, 0.75)) + 1.5*IQR(data_num3$YearsAtCompany)
data_num3$YearsAtCompany[data_num3$YearsAtCompany < livingar_lower | data_num3$YearsAtCompany > livingar_upper] <- NA
sum(is.na(data_num3$YearsAtCompany))
```

```
## [1] 45
```

```r
livingar_lower <- (quantile(data_num3$TrainingTimesLastYear, 0.25)) - 1.5*IQR(data_num3$TrainingTimesLastYear)
livingar_upper <- (quantile(data_num3$TrainingTimesLastYear, 0.75)) + 1.5*IQR(data_num3$TrainingTimesLastYear)
data_num3$TrainingTimesLastYear[data_num3$TrainingTimesLastYear < livingar_lower | data_num3$TrainingTimesLastYear > livingar_upper] <- NA
sum(is.na(data_num3$TrainingTimesLastYear))
```

```
## [1] 238
```

```r
livingar_lower <- (quantile(data_num3$TotalWorkingYears, 0.25)) - 1.5*IQR(data_num3$TotalWorkingYears)
livingar_upper <- (quantile(data_num3$TotalWorkingYears, 0.75)) + 1.5*IQR(data_num3$TotalWorkingYears)
data_num3$TotalWorkingYears[data_num3$TotalWorkingYears < livingar_lower | data_num3$TotalWorkingYears > livingar_upper] <- NA
sum(is.na(data_num3$TotalWorkingYears))
```

```
## [1] 92
```

```r
sum(is.na(data_num3))
```

```
## [1] 375
```

```r
data_num4 <- knnImputation(data = data_num3, k = 0.05*nrow(data_num3))
sum(is.na(data_num4))
```

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

```r
data_num4$Attrition <- ifelse(Attrition == "Yes",1,0 )

model_num <- glm(Attrition~., family = binomial, data = data_num4)
summary(model_num)
```

```
##
## Call:
## glm(formula = Attrition ~ ., family = binomial, data = data_num4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3843  -0.6284  -0.4451  -0.2943   2.9811
```

```
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.89905    0.09211 -20.617  < 2e-16 ***
## Age                      -0.31728    0.10337  -3.069 0.002145 **
## DailyRate                -0.14203    0.07508  -1.892 0.058517 .
## DistanceFromHome          0.25232    0.07671   3.289 0.001005 **
## HourlyRate               -0.01044    0.07541  -0.138 0.889865
## MonthlyIncome            -0.40136    0.11034  -3.637 0.000275 ***
## MonthlyRate               0.04223    0.07550   0.559 0.575993
## NumCompaniesWorked        0.29527    0.08690   3.398 0.000679 ***
## PercentSalaryHike        -0.07751    0.07674  -1.010 0.312534
## TotalWorkingYears        -0.06345    0.16840  -0.377 0.706335
## TrainingTimesLastYear     0.02769    0.14229   0.195 0.845718
## YearsAtCompany            0.07631    0.19755   0.386 0.699301
## YearsInCurrentRole       -0.30661    0.13542  -2.264 0.023560 *
## YearsSinceLastPromotion   0.39292    0.10416   3.772 0.000162 ***
## YearsWithCurrManager     -0.33703    0.12861  -2.621 0.008777 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1298.6  on 1469  degrees of freedom
## Residual deviance: 1162.1  on 1455  degrees of freedom
## AIC: 1192.1
##
## Number of Fisher Scoring iterations: 5
```

```
data_num1 <- select(data_num, c("Age", "DistanceFromHome", "MonthlyIncome",
                                "YearsSinceLastPromotion","NumCompaniesWorked
",
                                "YearsInCurrentRole", "YearsWithCurrManager")
)
```

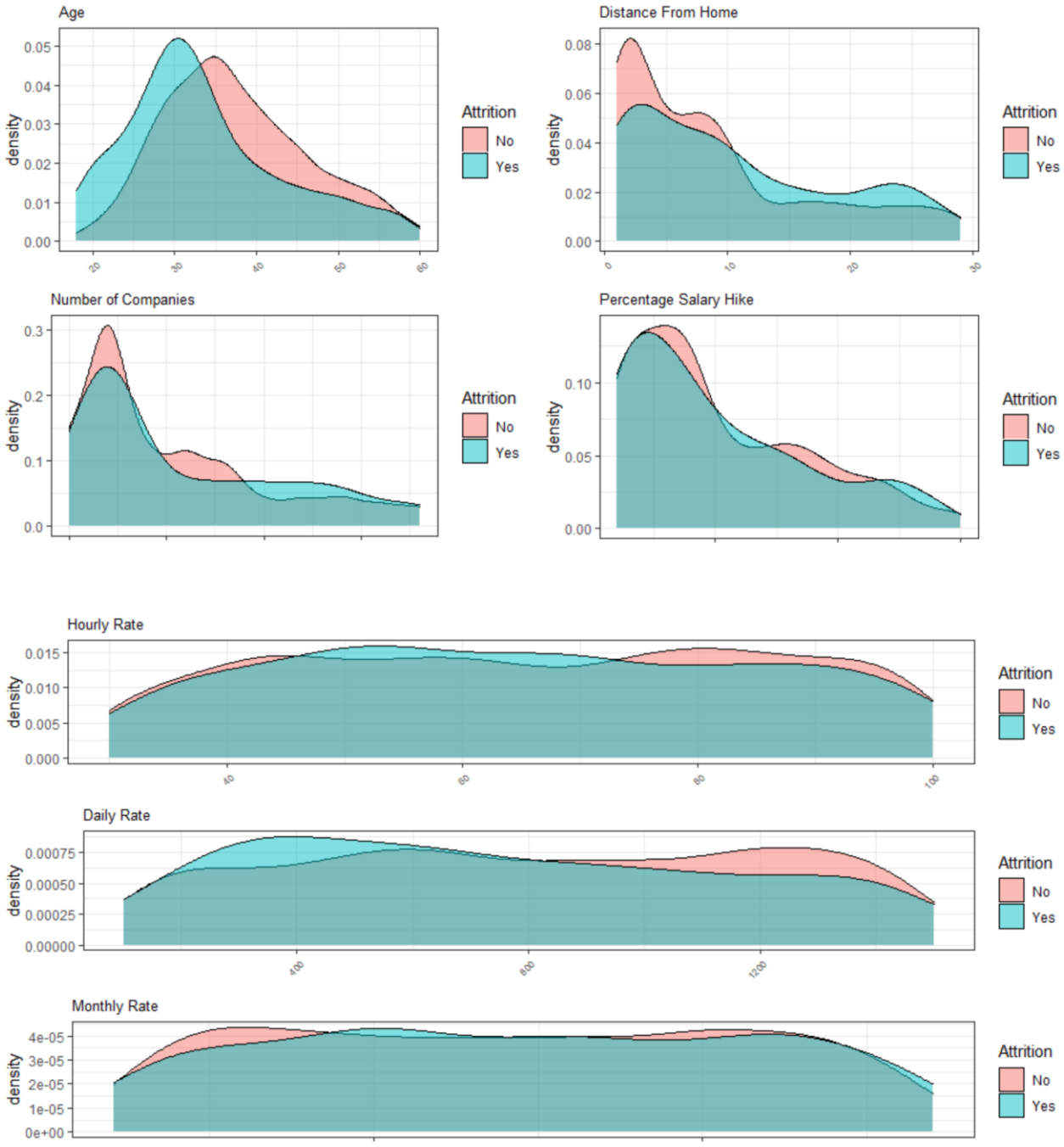
```
model_cat <- glm(Attrition~., family = binomial,data = data_cat)
summary(model_cat)
```

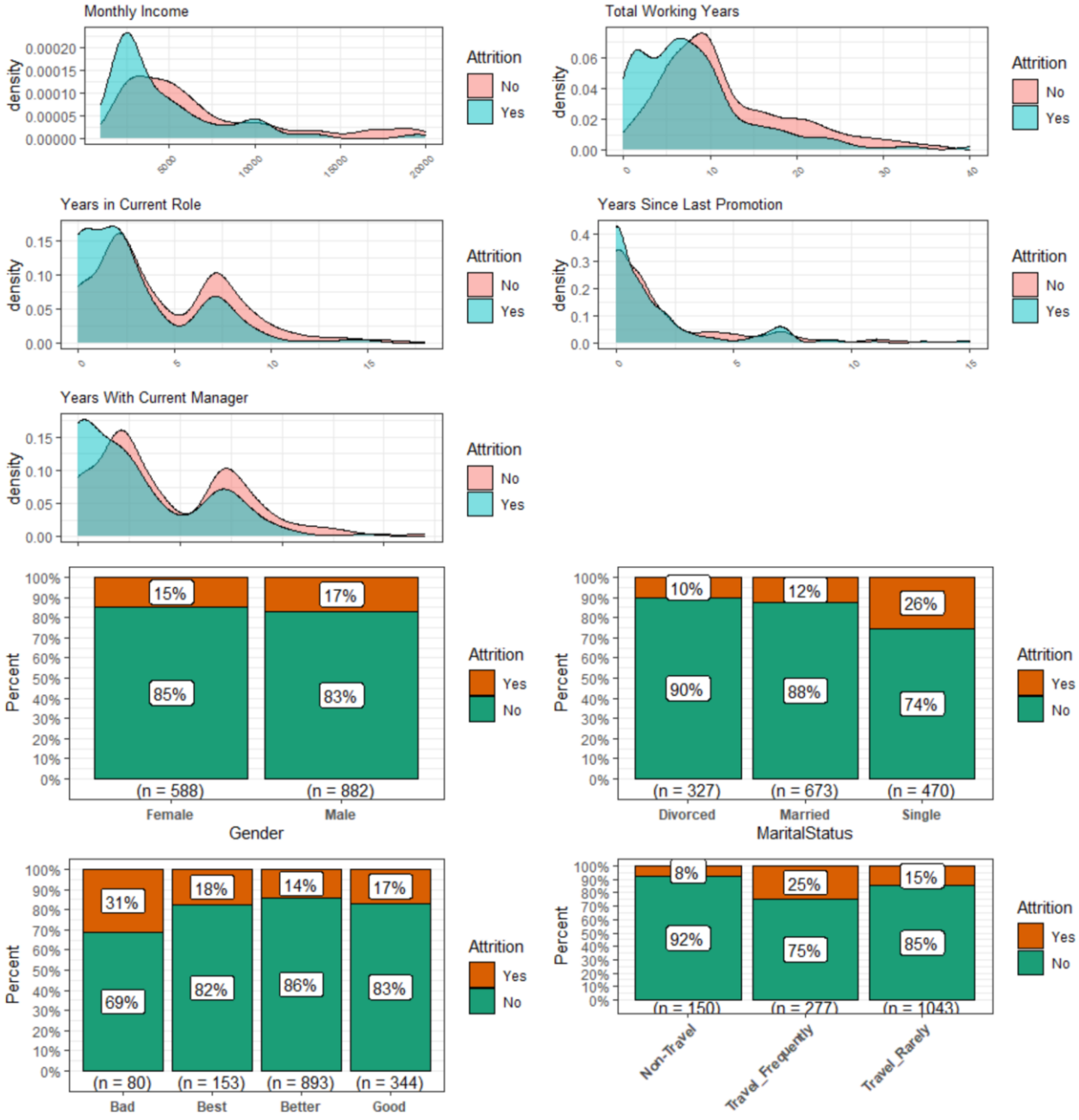
```
##
## Call:
## glm(formula = Attrition ~ ., family = binomial, data = data_cat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2246  -0.5117  -0.3008  -0.1391   3.3126
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -3.52669    0.96764  -3.645 0.000268 ***
## BusinessTravelTravel_Frequently   1.54839    0.38268   4.046 5.21e-05 ***
```

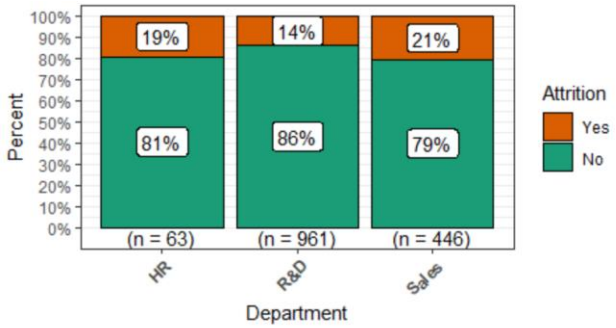
```
## BusinessTravelTravel_Rarely        0.79128    0.35643   2.220 0.026419 *
## EducationBelow College            -0.13928    0.27525  -0.506 0.612845
## EducationCollege                  -0.17996    0.23768  -0.757 0.448966
## EducationDoctor                   -0.47443    0.58312  -0.814 0.415863
## EducationMaster                   -0.19260    0.21829  -0.882 0.377628
## EducationFieldLife Sciences       -0.94973    0.72370  -1.312 0.189412
## EducationFieldMarketing           -0.56465    0.76878  -0.734 0.462662
## EducationFieldMedical             -0.95360    0.72960  -1.307 0.191208
## EducationFieldOther               -1.07949    0.79757  -1.353 0.175908
## EducationFieldTechnical Degree    -0.16560    0.74541  -0.222 0.824192
## EnvironmentSatisfactionLow         1.03816    0.23566   4.405 1.06e-05 ***
## EnvironmentSatisfactionMedium      0.04302    0.25540   0.168 0.866225
## EnvironmentSatisfactionVery High  -0.18622    0.22672  -0.821 0.411442
## GenderMale                         0.36589    0.17761   2.060 0.039393 *
## JobInvolvementLow                  1.42101    0.31503   4.511 6.46e-06 ***
## JobInvolvementMedium               0.33364    0.19467   1.714 0.086556 .
## JobInvolvementVery High           -0.63919    0.34961  -1.828 0.067503 .
## JobRoleHuman Resources             1.40291    0.63382   2.213 0.026869 *
## JobRoleLaboratory Technician       1.70299    0.41080   4.145 3.39e-05 ***
## JobRoleManager                    -0.26972    0.63685  -0.424 0.671916
## JobRoleManufacturing Director      0.17629    0.51115   0.345 0.730179
## JobRoleResearch Director          -1.13305    0.83527  -1.356 0.174941
## JobRoleResearch Scientist          0.89254    0.41402   2.156 0.031099 *
## JobRoleSales Executive             1.11091    0.42416   2.619 0.008816 **
## JobRoleSales Representative        2.32470    0.47819   4.861 1.17e-06 ***
## JobSatisfactionLow                 0.54788    0.22903   2.392 0.016749 *
## JobSatisfactionMedium              0.07973    0.24824   0.321 0.748076
## JobSatisfactionVery High          -0.72634    0.23234  -3.126 0.001771 **
## MaritalStatusMarried               0.36753    0.24581   1.495 0.134870
## MaritalStatusSingle                1.32131    0.24771   5.334 9.60e-08 ***
## OverTimeYes                        1.90107    0.18466  10.295  < 2e-16 ***
## PerformanceRatingOutstanding      -0.16852    0.24560  -0.686 0.492615
## RelationshipSatisfactionLow        0.58879    0.23892   2.464 0.013723 *
## RelationshipSatisfactionMedium    -0.01317    0.24964  -0.053 0.957923
## RelationshipSatisfactionVery High -0.09650    0.22283  -0.433 0.664973
## WorkLifeBalanceBest               -0.83574    0.39135  -2.136 0.032717 *
## WorkLifeBalanceBetter             -1.38114    0.32235  -4.285 1.83e-05 ***
## WorkLifeBalanceGood               -0.95942    0.34548  -2.777 0.005485 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1298.58  on 1469  degrees of freedom
## Residual deviance:  934.19  on 1430  degrees of freedom
## AIC: 1014.2
##
## Number of Fisher Scoring iterations: 6
```

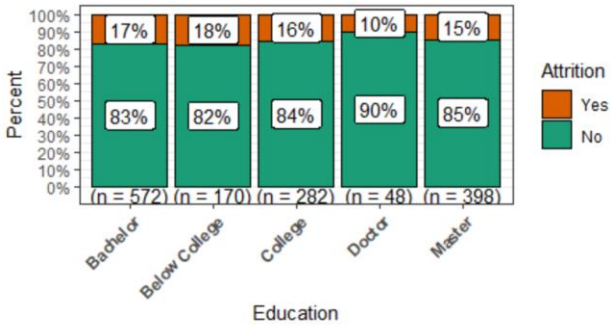
```
data_cat1 <- select(data_cat, c("BusinessTravel", "EnvironmentSatisfaction",
                                "JobInvolvement", "JobRole", "OverTime", "Mar
italStatus",
                                "WorkLifeBalance", "JobSatisfaction"))

## null device
##           1

## null device
##           1

## 'data.frame':    1470 obs. of  14 variables:
##  $ Attrition              : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1
1 1 1 ...
##  $ BusinessTravel         : Factor w/ 3 levels "Non-Travel","Travel_Frequ
ently",..: 3 2 3 2 3 2 3 3 2 3 ...
##  $ Education              : Factor w/ 5 levels "Bachelor","Below College"
,..: 3 2 3 5 2 3 1 2 1 1 ...
##  $ EducationField         : Factor w/ 6 levels "Human Resources",..: 2 2
5 2 4 2 4 2 2 4 ...
##  $ EnvironmentSatisfaction : Factor w/ 4 levels "High","Low","Medium",..:
3 1 4 4 2 4 1 4 4 1 ...
##  $ Gender                 : Factor w/ 2 levels "Female","Male": 1 2 2 1 2
2 1 2 2 2 ...
##  $ JobInvolvement         : Factor w/ 4 levels "High","Low","Medium",..:
1 3 3 1 1 1 4 1 3 1 ...
##  $ JobRole                : Factor w/ 9 levels "Healthcare Representative
",..: 8 7 3 7 3 3 3 3 5 1 ...
##  $ JobSatisfaction        : Factor w/ 4 levels "High","Low","Medium",..:
4 3 1 1 3 4 2 1 1 1 ...
##  $ MaritalStatus          : Factor w/ 3 levels "Divorced","Married",..: 3
2 3 2 2 3 2 1 3 2 ...
##  $ OverTime               : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2
1 1 1 ...
##  $ PerformanceRating      : Factor w/ 2 levels "Excellent","Outstanding":
1 2 1 1 1 1 2 2 2 1 ...
##  $ RelationshipSatisfaction: Factor w/ 4 levels "High","Low","Medium",..:
2 4 3 1 4 1 2 3 3 3 ...
##  $ WorkLifeBalance        : Factor w/ 4 levels "Bad","Best","Better",..:
1 3 3 3 3 4 4 3 3 4 ...
```

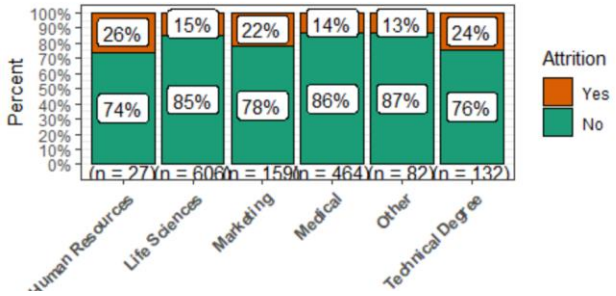
#----Data Visualization----

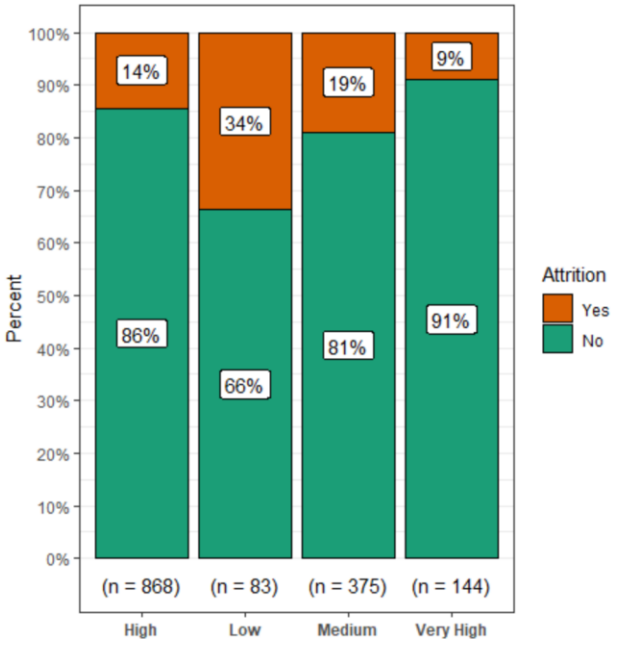

Age
density
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
Attrition
No
Yes
Distance From Home
density
0.08
0.06
0.04
0.02
0.00
0
10
20
30
Attrition
No
Yes
Number of Companies
density
0.3
0.2
0.1
0.0
Attrition
No
Yes
Percentage Salary Hike
density
0.10
0.05
0.00
Attrition
No
Yes
Hourly Rate
density
0.015
0.010
0.005
0.000
40
60
80
100
Attrition
No
Yes
Daily Rate
density
0.00075
0.00050
0.00025
0.00000
400
800
1200
Attrition
No
Yes
Monthly Rate
density
4e-05
3e-05
2e-05
1e-05
0e+00
Attrition
No
Yes

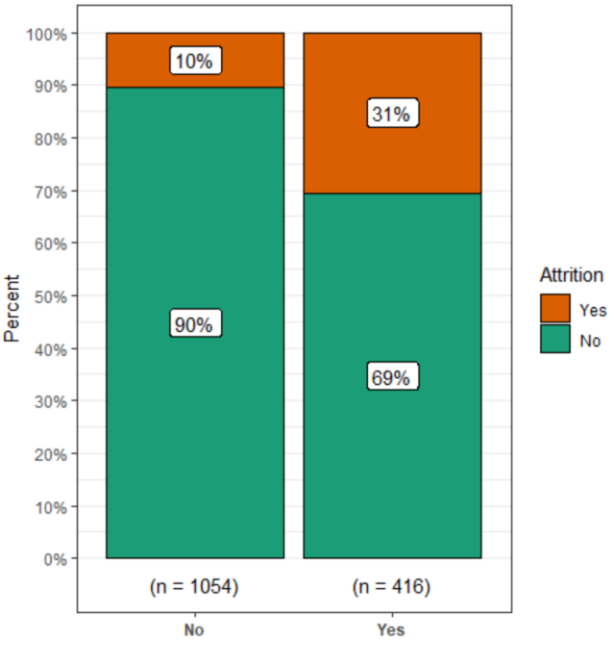

Monthly Income
density
0.00020
0.00015
0.00010
0.00005
0.00000
5000
10000
15000
20000
Attrition
No
Yes
Total Working Years
density
0.06
0.04
0.02
0.00
0
10
20
30
40
Attrition
No
Yes
Years in Current Role
density
0.15
0.10
0.05
0.00
0
5
10
15
Attrition
No
Yes
Years Since Last Promotion
density
0.4
0.3
0.2
0.1
0.0
0
5
10
15
Attrition
No
Yes
Years With Current Manager
density
0.15
0.10
0.05
0.00
Attrition
No
Yes
Percent
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
15%
17%
85%
83%
(n = 588)
(n = 882)
Female
Male
Gender
Attrition
Yes
No
10%
12%
26%
90%
88%
74%
(n = 327)
(n = 673)
(n = 470)
Divorced
Married
Single
MaritalStatus
Attrition
Yes
No
31%
18%
14%
17%
69%
82%
86%
83%
(n = 80)
(n = 153)
(n = 893)
(n = 344)
Bad
Best
Better
Good
Attrition
Yes
No
8%
25%
15%
92%
75%
85%
(n = 150)
(n = 277)
(n = 1043)
Non-Travel
Travel_Frequently
Travel_Rarely
Attrition
Yes
No

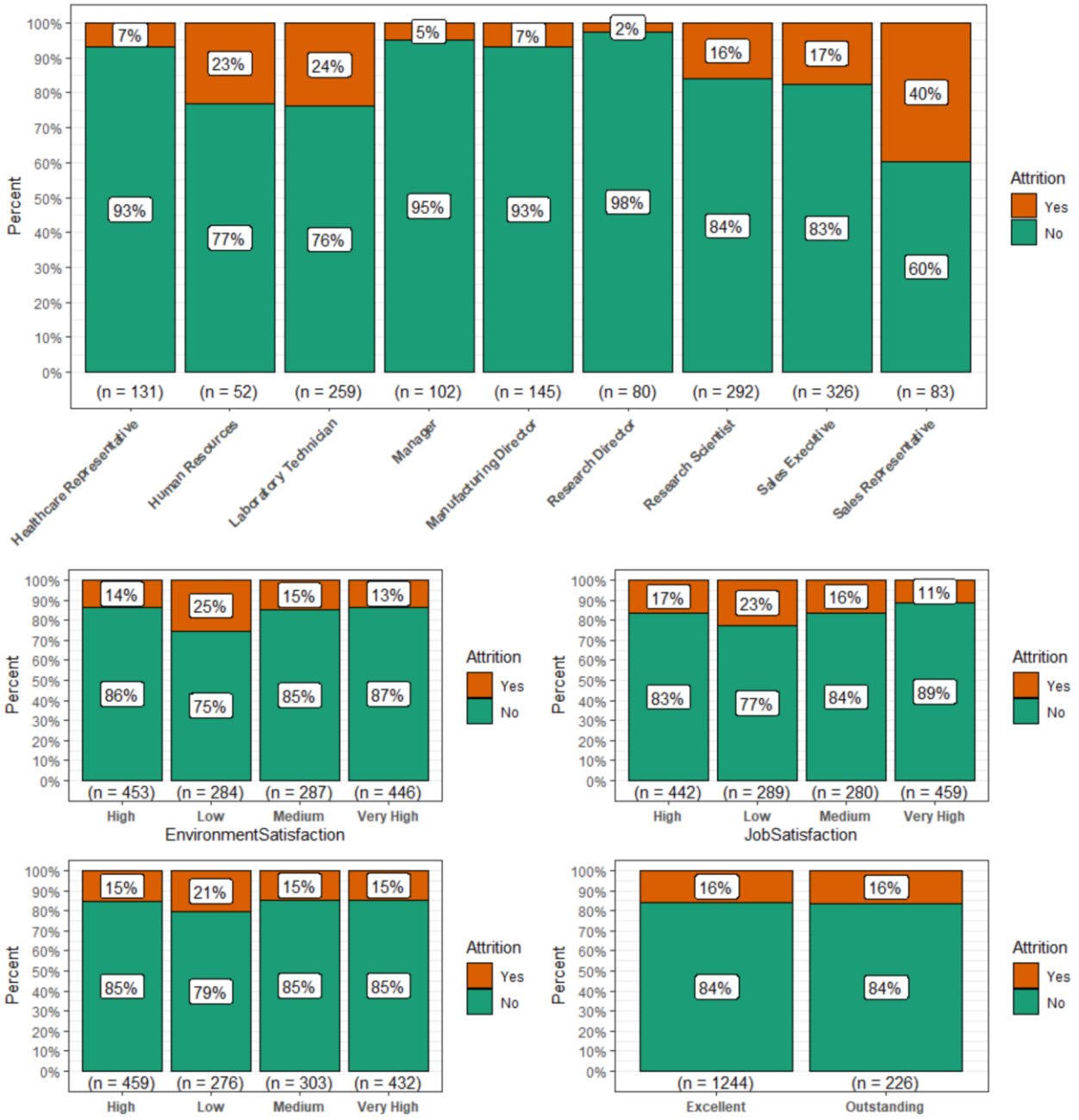

Percent
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
19%
14%
21%
81%
86%
79%
(n = 63)
(n = 961)
(n = 446)
HR
R&D
Sales
Department
Attrition
Yes
No


Percent
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
17%
18%
16%
10%
15%
83%
82%
84%
90%
85%
(n = 572)
(n = 170)
(n = 282)
(n = 48)
(n = 398)
Bachelor
Below College
College
Doctor
Master
Education
Attrition
Yes
No


Percent
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
26%
15%
22%
14%
13%
24%
74%
85%
78%
86%
87%
76%
(n = 27)
(n = 606)
(n = 159)
(n = 464)
(n = 82)
(n = 132)
Life Sciences
Marketing
Medical
Other
Technical Degree
Attrition
Yes
No


Percent
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
14%
34%
19%
9%
86%
66%
81%
91%
(n = 868)
(n = 83)
(n = 375)
(n = 144)
High
Low
Medium
Very High
Attrition
Yes
No


Percent
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
10%
31%
90%
69%
(n = 1054)
(n = 416)
No
Yes
Attrition
Yes
No



Observations : 1. Age is more or less normally distributed. 2. Distance from Home, Num of Companies worked and Total working Years is right skewed and should be transformed to curb the skewness. 3. More than 50% of Male population. 4. More than 37% is having a bachelors degree and 27% is having a masters degree. 5. Almost 75% of the employees come from Life Sciences and Medical background. 6. 45% of the employees are marrried whereas 22% of them are divorced. 7. More than 30% of the employees have a High or Very High Relationship Satisfaction. 8. More than 60% of the employees feel they have a better work life balance. 9. More than 70% of the employees Travel rarely for work. 10. 30% of the employees have a high and very high environment satisfaction each. 11. Almost 59% of the employees think they have a High job involvement at work. 12. Again in Job

Satisifaction we see that 30% employees have a high and a very high job satisfaction each. 13. More than 70% of the people seem to be working over time. 14. 85% of the employees have an excellent performance rating. 15. More than half of the employees work for the R&D department. 16. Majority of the employees work as Sales Executives, Research Scientists and Laboratory Technicians. 17. Younger employees within 25-35 years have a higher attrition rate. 18. Lower attrition rate is seen when the Distance from home is within 10 kms. The attrition rate increase post 10kms. 19. The attrition rate tends to be higher with employees who have worked with 5 to 7 companies. 20. Attrition rate seems to be extremely high amongst employees who have a total working experience between 0 to 7 years approximately. 21. Attrition Rate is slightly more in Males as compared to Females. 22. 18% attrition rate is observed amongst employees have below college education. 23. Attrition rate is very high amongst employees from HR, Marketing and Technical backgrounds. 24. As expected, the attrition rate is very high amongst employees who have a bad work life balance. 25. Attrition rate is higher amongst people who travel frequently. 26. Its also higher amongst employees who have a low environment satisfaction, low job involvement and low job satisfaction. 27. The attrition rate is almost 30% amongst employees who work over time. 28. Sales department have the highest attrition at 20% whereas Sales Representatives have the highest attrition at 40%.

```
#----Final Data Building----
final_data = cbind(data_num1, data_cat1)
final_data$Attrition = Attrition
final_data$Attrition <- ifelse(final_data$Attrition == "Yes",1,0 )
str(final_data)

## 'data.frame':    1470 obs. of  16 variables:
##  $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9
526 5237 ...
##  $ YearsSinceLastPromotion: int  0 1 0 3 2 3 0 0 1 7 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Freque
ntly",..: 3 2 3 2 3 2 3 3 2 3 ...
##  $ EnvironmentSatisfaction: Factor w/ 4 levels "High","Low","Medium",..: 3
1 4 4 2 4 1 4 4 1 ...
##  $ JobInvolvement          : Factor w/ 4 levels "High","Low","Medium",..: 1
3 3 1 1 1 4 1 3 1 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative"
,..: 8 7 3 7 3 3 3 3 5 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1
1 1 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3
2 3 2 2 3 2 1 3 2 ...
##  $ WorkLifeBalance         : Factor w/ 4 levels "Bad","Best","Better",..: 1
3 3 3 3 4 4 3 3 4 ...
##  $ JobSatisfaction         : Factor w/ 4 levels "High","Low","Medium",..: 4
```

```
3 1 1 3 4 2 1 1 1 ...
##  $ Attrition              : num  1 0 1 0 0 0 0 0 0 0 ...

#Train & Test data
Train <- createDataPartition(final_data$Attrition, p=0.7, list=FALSE)
names(final_data)

##  [1] "Age"                    "DistanceFromHome"
##  [3] "MonthlyIncome"          "YearsSinceLastPromotion"
##  [5] "NumCompaniesWorked"     "YearsInCurrentRole"
##  [7] "YearsWithCurrManager"   "BusinessTravel"
##  [9] "EnvironmentSatisfaction" "JobInvolvement"
## [11] "JobRole"                "OverTime"
## [13] "MaritalStatus"          "WorkLifeBalance"
## [15] "JobSatisfaction"        "Attrition"

training <- (final_data[ Train, ])
testing <- final_data[ -Train, -16]
str(training)

## 'data.frame':    1029 obs. of  16 variables:
##  $ Age                    : int  41 33 27 32 59 30 38 36 35 31 ...
##  $ DistanceFromHome       : int  1 3 2 2 3 24 23 27 16 26 ...
##  $ MonthlyIncome          : int  5993 2909 3468 3068 2670 2693 9526 5237 2
426 2911 ...
##  $ YearsSinceLastPromotion: int  0 3 2 3 0 0 1 7 0 4 ...
##  $ NumCompaniesWorked     : int  8 1 9 0 4 1 0 6 0 1 ...
##  $ YearsInCurrentRole     : int  4 7 2 7 0 0 7 7 4 2 ...
##  $ YearsWithCurrManager   : int  5 0 2 6 0 0 8 7 3 3 ...
##  $ BusinessTravel         : Factor w/ 3 levels "Non-Travel","Travel_Freque
ntly",..: 3 2 3 2 3 3 2 3 3 3 ...
##  $ EnvironmentSatisfaction: Factor w/ 4 levels "High","Low","Medium",..: 3
4 2 4 1 4 4 1 2 2 ...
##  $ JobInvolvement         : Factor w/ 4 levels "High","Low","Medium",..: 1
1 1 1 4 1 3 1 4 1 ...
##  $ JobRole                : Factor w/ 9 levels "Healthcare Representative"
,..: 8 7 3 3 3 3 5 1 3 7 ...
##  $ OverTime               : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1
1 1 ...
##  $ MaritalStatus          : Factor w/ 3 levels "Divorced","Married",..: 3
2 2 3 2 1 3 2 2 1 ...
##  $ WorkLifeBalance        : Factor w/ 4 levels "Bad","Best","Better",..: 1
3 3 4 4 3 3 4 3 4 ...
##  $ JobSatisfaction        : Factor w/ 4 levels "High","Low","Medium",..: 4
1 3 4 2 1 1 1 3 1 ...
##  $ Attrition              : num  1 0 0 0 0 0 0 0 0 0 ...

str(testing)

## 'data.frame':    441 obs. of  15 variables:
##  $ Age                    : int  49 37 29 28 53 38 21 32 44 46 ...
```

```
##  $ DistanceFromHome       : int  8 2 15 24 2 2 15 16 7 2 ...
##  $ MonthlyIncome          : int  5130 2090 4193 2028 15427 3944 1232 3919
10248 18947 ...
##  $ YearsSinceLastPromotion: int  1 0 0 0 3 1 0 6 5 2 ...
##  $ NumCompaniesWorked     : int  1 6 0 5 2 5 1 1 3 3 ...
##  $ YearsInCurrentRole     : int  7 0 5 2 8 2 0 2 6 2 ...
##  $ YearsWithCurrManager   : int  7 0 8 3 7 2 0 7 17 1 ...
##  $ BusinessTravel         : Factor w/ 3 levels "Non-Travel","Travel_Freque
ntly",..: 2 3 3 3 3 3 3 2 3 3 ...
##  $ EnvironmentSatisfaction: Factor w/ 4 levels "High","Low","Medium",..: 1
4 4 1 2 4 1 3 2 3 ...
##  $ JobInvolvement         : Factor w/ 4 levels "High","Low","Medium",..: 3
3 3 3 3 1 1 2 3 1 ...
##  $ JobRole                : Factor w/ 9 levels "Healthcare Representative"
,..: 7 3 3 3 4 7 7 7 1 4 ...
##  $ OverTime               : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 2 1 2
1 1 ...
##  $ MaritalStatus          : Factor w/ 3 levels "Divorced","Married",..: 2
3 3 3 2 3 3 3 2 3 ...
##  $ WorkLifeBalance        : Factor w/ 4 levels "Bad","Best","Better",..: 3
3 3 3 3 3 3 3 3 4 ...
##  $ JobSatisfaction        : Factor w/ 4 levels "High","Low","Medium",..: 3
1 1 1 4 4 4 2 4 2 ...
```

Comments: We need to ensure that the Attrition data in both training and test set is of the same proporation as we have it in our main dataset to avoid any bias in prediction. For this I have use createDataPartition function. Training dataset with have 70% of the rows whereas the test dataset will have the remaining 30% !!!

```
#----Regression Model Development----
#Basic Model - MOdel0
model <- glm(Attrition~., family = binomial,data = training)
summary(model)

##
## Call:
## glm(formula = Attrition ~ ., family = binomial, data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9246  -0.4740  -0.2417  -0.0972   3.3054
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -3.165e+00  1.056e+00  -2.997 0.002729 **
## Age                          -5.673e-02  1.491e-02  -3.805 0.000142 **
*
## DistanceFromHome              5.969e-02  1.340e-02   4.453 8.47e-06 **
*
## MonthlyIncome                -1.711e-05  6.627e-05  -0.258 0.796251
```

```
## YearsSinceLastPromotion           1.884e-01  4.979e-02   3.784 0.000154 **
*
## NumCompaniesWorked                1.679e-01  4.537e-02   3.700 0.000215 **
*
## YearsInCurrentRole               -8.582e-02  5.256e-02  -1.633 0.102509
## YearsWithCurrManager             -1.262e-01  5.110e-02  -2.470 0.013493 *
## BusinessTravelTravel_Frequently   1.958e+00  5.181e-01   3.780 0.000157 **
*
## BusinessTravelTravel_Rarely       1.064e+00  4.846e-01   2.195 0.028157 *
## EnvironmentSatisfactionLow        1.217e+00  3.057e-01   3.980 6.88e-05 **
*
## EnvironmentSatisfactionMedium    -1.657e-01  3.361e-01  -0.493 0.622091
## EnvironmentSatisfactionVery High -1.191e-01  2.991e-01  -0.398 0.690483
## JobInvolvementLow                 1.603e+00  3.893e-01   4.117 3.83e-05 **
*
## JobInvolvementMedium              2.444e-01  2.565e-01   0.953 0.340576
## JobInvolvementVery High          -3.827e-01  4.294e-01  -0.891 0.372782
## JobRoleHuman Resources            1.655e+00  7.342e-01   2.254 0.024200 *
## JobRoleLaboratory Technician      1.537e+00  5.682e-01   2.704 0.006848 **
## JobRoleManager                    4.640e-01  9.216e-01   0.503 0.614672
## JobRoleManufacturing Director    -1.534e-01  6.749e-01  -0.227 0.820242
## JobRoleResearch Director         -2.490e-01  1.070e+00  -0.233 0.816051
## JobRoleResearch Scientist         7.445e-01  5.807e-01   1.282 0.199850
## JobRoleSales Executive            1.258e+00  5.084e-01   2.475 0.013337 *
## JobRoleSales Representative       2.351e+00  6.444e-01   3.648 0.000265 **
*
## OverTimeYes                       1.960e+00  2.364e-01   8.295  < 2e-16 **
*
## MaritalStatusMarried              5.691e-01  3.211e-01   1.772 0.076359 .
## MaritalStatusSingle               1.666e+00  3.263e-01   5.107 3.28e-07 **
*
## WorkLifeBalanceBest              -1.069e+00  5.016e-01  -2.132 0.033020 *
## WorkLifeBalanceBetter            -1.665e+00  4.151e-01  -4.012 6.03e-05 **
*
## WorkLifeBalanceGood              -1.419e+00  4.425e-01  -3.208 0.001338 **
## JobSatisfactionLow                4.719e-01  3.016e-01   1.565 0.117650
## JobSatisfactionMedium             3.071e-01  3.070e-01   1.000 0.317123
## JobSatisfactionVery High         -6.765e-01  2.912e-01  -2.323 0.020183 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 885.90  on 1028  degrees of freedom
## Residual deviance: 582.76  on  996  degrees of freedom
## AIC: 648.76
##
## Number of Fisher Scoring iterations: 6
```

```r
# Multicollinearity
library(car)
car::vif(model)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## Age                      1.531403  1        1.237499
## DistanceFromHome         1.138388  1        1.066952
## MonthlyIncome            5.190282  1        2.278219
## YearsSinceLastPromotion 2.078741  1        1.441784
## NumCompaniesWorked       1.294272  1        1.137661
## YearsInCurrentRole       2.279329  1        1.509745
## YearsWithCurrManager     2.230990  1        1.493650
## BusinessTravel           1.170929  2        1.040238
## EnvironmentSatisfaction 1.250551  3        1.037967
## JobInvolvement           1.209487  3        1.032207
## JobRole                  6.244537  8        1.121292
## OverTime                 1.224152  1        1.106414
## MaritalStatus            1.159695  2        1.037734
## WorkLifeBalance          1.204154  3        1.031447
## JobSatisfaction          1.195327  3        1.030183
```

```r
#Autocoreation
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
dwt(model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.05480111       1.88787   0.074
##  Alternative hypothesis: rho != 0
```

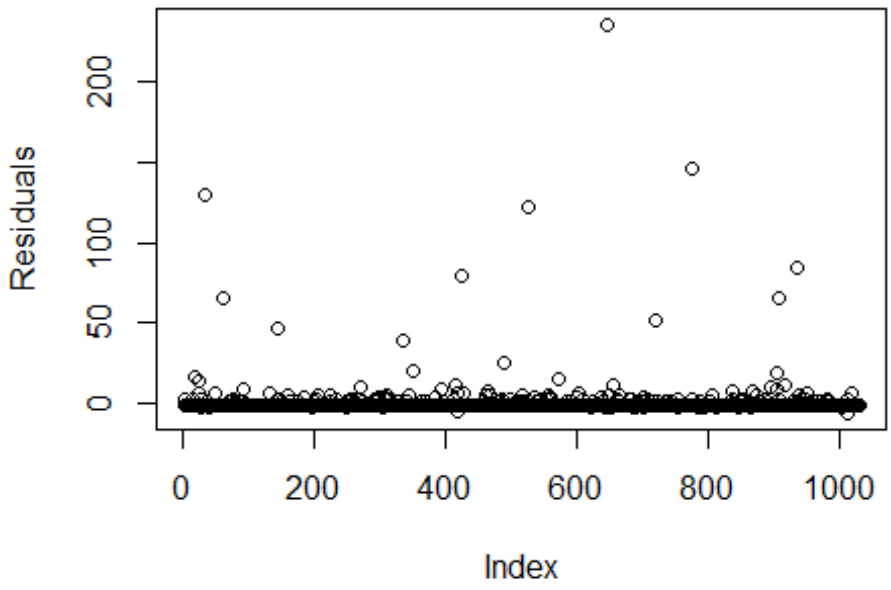
```r
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.96, p-value = 0.2592
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
#Hetroscadisticity
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 141.78, df = 32, p-value = 9.027e-16
```

```
plot(model$residuals, ylab = "Residuals")
```




```
#Model1
training1 <- select(training, c("Age", "DistanceFromHome", "YearsSinceLastPro
motion",
                                "BusinessTravel","EnvironmentSatisfaction","M
aritalStatus",
                                "JobInvolvement", "OverTime", "WorkLifeBalanc
e",
                                "JobSatisfaction", "JobRole", "Attrition"))

model1 <- glm(Attrition~., family = binomial,data = training1)
summary(model1)
```

```
##
## Call:
## glm(formula = Attrition ~ ., family = binomial, data = training1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9266  -0.5090  -0.2842  -0.1315   3.1605
```

```
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -3.05262    0.97014  -3.147 0.001652 **
## Age                              -0.04644    0.01335  -3.478 0.000505 ***
## DistanceFromHome                  0.05324    0.01283   4.149 3.34e-05 ***
## YearsSinceLastPromotion           0.04787    0.03799   1.260 0.207664
## BusinessTravelTravel_Frequently   1.76296    0.48868   3.608 0.000309 ***
## BusinessTravelTravel_Rarely       0.93996    0.45881   2.049 0.040491 *
## EnvironmentSatisfactionLow        1.11155    0.28974   3.836 0.000125 ***
## EnvironmentSatisfactionMedium    -0.25908    0.32499  -0.797 0.425335
## EnvironmentSatisfactionVery High -0.17232    0.28773  -0.599 0.549244
## MaritalStatusMarried              0.53362    0.31315   1.704 0.088370 .
## MaritalStatusSingle               1.56799    0.31604   4.961 7.00e-07 ***
## JobInvolvementLow                 1.51634    0.37121   4.085 4.41e-05 ***
## JobInvolvementMedium              0.23602    0.24695   0.956 0.339190
## JobInvolvementVery High          -0.43667    0.40701  -1.073 0.283335
## OverTimeYes                       1.82373    0.22351   8.159 3.37e-16 ***
## WorkLifeBalanceBest              -0.96258    0.48233  -1.996 0.045967 *
## WorkLifeBalanceBetter            -1.63044    0.39669  -4.110 3.95e-05 ***
## WorkLifeBalanceGood              -1.36588    0.42404  -3.221 0.001277 **
## JobSatisfactionLow                0.30052    0.28931   1.039 0.298921
## JobSatisfactionMedium             0.25956    0.29693   0.874 0.382050
## JobSatisfactionVery High         -0.78540    0.28334  -2.772 0.005573 **
## JobRoleHuman Resources            1.45833    0.68298   2.135 0.032742 *
## JobRoleLaboratory Technician      1.45543    0.49234   2.956 0.003115 **
## JobRoleManager                    0.09579    0.74118   0.129 0.897172
## JobRoleManufacturing Director    -0.42732    0.65854  -0.649 0.516405
## JobRoleResearch Director         -0.68436    0.91218  -0.750 0.453109
## JobRoleResearch Scientist         0.68884    0.50069   1.376 0.168887
## JobRoleSales Executive            1.08862    0.48518   2.244 0.024849 *
## JobRoleSales Representative       2.26373    0.56217   4.027 5.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 885.90  on 1028  degrees of freedom
## Residual deviance: 621.16  on 1000  degrees of freedom
## AIC: 679.16
##
## Number of Fisher Scoring iterations: 6
```

```
car::vif(model1)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## Age                      1.240302  1        1.113688
## DistanceFromHome         1.098225  1        1.047962
## YearsSinceLastPromotion 1.236994  1        1.112202
## BusinessTravel           1.123832  2        1.029616
```

```
## EnvironmentSatisfaction 1.199201  3        1.030739
## MaritalStatus           1.136726  2        1.032557
## JobInvolvement          1.153529  3        1.024090
## OverTime                1.159474  1        1.076789
## WorkLifeBalance         1.160902  3        1.025178
## JobSatisfaction         1.161186  3        1.025220
## JobRole                 1.663206  8        1.032308
```

```
dwtest(model1)
```

```
##
##  Durbin-Watson test
##
## data:  model1
## DW = 1.9445, p-value = 0.1857
## alternative hypothesis: true autocorrelation is greater than 0
```

```
#Model2
training2 <- select(training, c("Age", "BusinessTravel","EnvironmentSatisfact
ion",
                                "MaritalStatus","JobInvolvement", "OverTime",
                                "WorkLifeBalance", "JobRole", "Attrition"))

model2 <- glm(Attrition~., family = binomial,data = training2)
summary(model2)
```

```
##
## Call:
## glm(formula = Attrition ~ ., family = binomial, data = training2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9124  -0.5262  -0.3112  -0.1607   3.2460
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -2.38015    0.86501  -2.752 0.005931 **
## Age                              -0.04021    0.01264  -3.182 0.001461 **
## BusinessTravelTravel_Frequently   1.53588    0.46320   3.316 0.000914 ***
## BusinessTravelTravel_Rarely       0.81311    0.43614   1.864 0.062276 .
## EnvironmentSatisfactionLow        1.04695    0.27522   3.804 0.000142 ***
## EnvironmentSatisfactionMedium    -0.28298    0.31767  -0.891 0.373040
## EnvironmentSatisfactionVery High -0.23131    0.27521  -0.840 0.400632
## MaritalStatusMarried              0.52676    0.30640   1.719 0.085582 .
## MaritalStatusSingle               1.41038    0.30694   4.595 4.33e-06 ***
## JobInvolvementLow                 1.33550    0.35786   3.732 0.000190 ***
## JobInvolvementMedium              0.21597    0.23796   0.908 0.364105
## JobInvolvementVery High          -0.33181    0.39442  -0.841 0.400204
## OverTimeYes                       1.67226    0.21281   7.858 3.90e-15 ***
## WorkLifeBalanceBest              -0.96830    0.46706  -2.073 0.038155 *
## WorkLifeBalanceBetter            -1.50482    0.38127  -3.947 7.92e-05 ***
```

```
## WorkLifeBalanceGood               -1.31413    0.40976  -3.207 0.001341 **
## JobRoleHuman Resources             1.23895    0.67042   1.848 0.064600 .
## JobRoleLaboratory Technician       1.29045    0.47454   2.719 0.006541 **
## JobRoleManager                     0.06586    0.71342   0.092 0.926451
## JobRoleManufacturing Director     -0.44594    0.64703  -0.689 0.490690
## JobRoleResearch Director          -0.78650    0.89496  -0.879 0.379505
## JobRoleResearch Scientist          0.52786    0.48545   1.087 0.276879
## JobRoleSales Executive             0.94878    0.47185   2.011 0.044351 *
## JobRoleSales Representative        1.99153    0.54045   3.685 0.000229 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 885.90  on 1028  degrees of freedom
## Residual deviance: 657.01  on 1005  degrees of freedom
## AIC: 705.01
##
## Number of Fisher Scoring iterations: 6
```

```
car::vif(model2)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## Age                      1.187086  1        1.089535
## BusinessTravel           1.063308  2        1.015465
## EnvironmentSatisfaction 1.154510  3        1.024235
## MaritalStatus            1.092704  2        1.022411
## JobInvolvement           1.105188  3        1.016809
## OverTime                 1.110082  1        1.053605
## WorkLifeBalance          1.119591  3        1.019006
## JobRole                  1.381613  8        1.020409
```

```
dwtest(model2)
```

```
##
##  Durbin-Watson test
##
## data:  model2
## DW = 1.9421, p-value = 0.1747
## alternative hypothesis: true autocorrelation is greater than 0
```

```
#----Model Analysis----
#Goodness of fit - hoslem.test
library(ResourceSelection)
library(rcompanion)
```

```
##
## Attaching package: 'rcompanion'
```

```
## The following object is masked from 'package:psych':
##
##     phi
```

```
hoslem.test(training$Attrition, fitted(model))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  training$Attrition, fitted(model)
## X-squared = 26.028, df = 8, p-value = 0.001039
```

```
nagelkerke(model)
```

```
## $Models
##
## Model: "glm, Attrition ~ ., binomial, training"
## Null:  "glm, Attrition ~ 1, binomial, training"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                             0.342191
## Cox and Snell (ML)                   0.255174
## Nagelkerke (Cragg and Uhler)         0.442064
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff  Chisq    p.value
##      -32     -151.57 303.15 6.4561e-46
##
## $Number.of.observations
##
## Model: 1029
## Null:  1029
##
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitti
ng with ML"
##
## $Warnings
## [1] "None"
```

```
hoslem.test(training1$Attrition, fitted(model1))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  training1$Attrition, fitted(model1)
## X-squared = 4.8881, df = 8, p-value = 0.7695
```

```
nagelkerke(model1)
```

```
## $Models
##
## Model: "glm, Attrition ~ ., binomial, training1"
## Null:  "glm, Attrition ~ 1, binomial, training1"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                            0.298844
## Cox and Snell (ML)                  0.226853
## Nagelkerke (Cragg and Uhler)        0.393000
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff  Chisq    p.value
##      -28     -132.37 264.75 2.2107e-40
##
## $Number.of.observations
##
## Model: 1029
## Null:  1029
##
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitti
ng with ML"
##
## $Warnings
## [1] "None"
```

```
hoslem.test(training2$Attrition, fitted(model2))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  training2$Attrition, fitted(model2)
## X-squared = 7.1314, df = 8, p-value = 0.5225
```

```
nagelkerke(model2)
```

```
## $Models
##
## Model: "glm, Attrition ~ ., binomial, training2"
## Null:  "glm, Attrition ~ 1, binomial, training2"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                            0.258372
## Cox and Snell (ML)                  0.199438
## Nagelkerke (Cragg and Uhler)        0.345507
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff  Chisq   p.value
##      -23     -114.45 228.89 7.547e-36
```

```
##
## $Number.of.observations
##
## Model: 1029
## Null:  1029
##
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitti
ng with ML"
##
## $Warnings
## [1] "None"
```

```
#Prediction value mutation for three models

predict1 <- predict(model,type = "response")
predict1
```

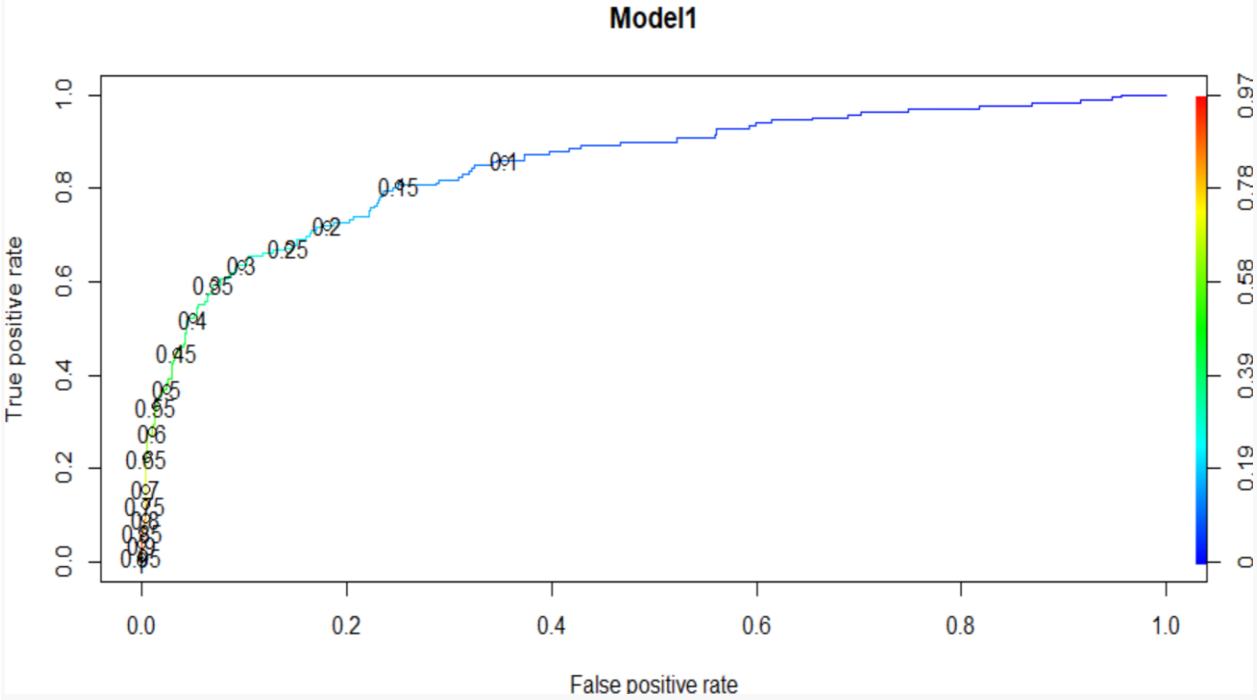
```
##            1            4            5            6            7
predict2 <- predict(model1,type = "response")
predict2
```

```
##            1            4            5            6            7
8
## 0.3710764971 0.2183334006 0.1923830165 0.1387369298 0.0845447264 0.0707411
749


predict3 <- predict(model2,type = "response")
predict3
```

```
##           1           4           5           6           7           8
## 0.630141523 0.234975583 0.215287329 0.273770996 0.109500972 0.038457416


predict_glm <- predict(model1,type = "response")

training$predict1 <- predict1
training$predict2 <- predict2
training$predict3 <- predict3

#Calculating performance for the three models
library(ROCR)

pred1 <- prediction(training$predict1,training$Attrition)
perf1 <- performance(pred1,"tpr","fpr")
perf1
```

```
## A performance instance
##   'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')
##   with 1030 data points
```

```r
pred2 <- prediction(training$predict2,training$Attrition)
perf2 <- performance(pred2,"tpr","fpr")
perf2

## A performance instance
##   'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')
##   with 1030 data points

pred3 <- prediction(training$predict3,training$Attrition)
perf3 <- performance(pred3,"tpr","fpr")
perf3

## A performance instance
##   'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')
##   with 1015 data points

prediction_glm <- prediction(training$predict2,training$Attrition)
perf_glm <- performance(pred2,"tpr","fpr")

dev.off()

## null device
##           1

#par(mfrow = c(1,2))
plot(perf1, colorize = T, print.cutoffs.at = seq(0.1, by=0.05), main = "Model
0")
plot(perf2, colorize = T, print.cutoffs.at = seq(0.1, by=0.05), main = "Model
1")
plot(perf3, colorize = T, print.cutoffs.at = seq(0.1, by=0.05), main = "Model
2")
```




```
#-----ROC Method------
library(ggplot2)
library(pROC)
library(PRROC)
library(ROCR)
library(plotROC)

#Model0 confusion matrix
predicted11 <- ifelse(predict1 > 0.48, 1, 0)
cnfmtrx11 <- table(prd=predicted11, training$Attrition)
confusionMatrix(cnfmtrx11)

## Confusion Matrix and Statistics
##
##
## prd   0   1
##   0 844  81
##   1  26  78
##
##                Accuracy : 0.896
##                  95% CI : (0.8757, 0.914)
##     No Information Rate : 0.8455
##     P-Value [Acc > NIR] : 1.565e-06
##
##                   Kappa : 0.5365
##
##  Mcnemar's Test P-Value : 1.786e-07
```

```
##
##             Sensitivity : 0.9701
##             Specificity : 0.4906
##          Pos Pred Value : 0.9124
##          Neg Pred Value : 0.7500
##              Prevalence : 0.8455
##          Detection Rate : 0.8202
##    Detection Prevalence : 0.8989
##       Balanced Accuracy : 0.7303
##
##        'Positive' Class : 0
##
```

```
OAA_m0 <- ((cnfmtrx11[1,1]+cnfmtrx11[2,2])/sum(cnfmtrx11))
OAA_m0
```

```
## [1] 0.8960155
```

```
#Model 1 confusion matrix
predicted22 <- ifelse(predict2 > 0.48, 1, 0)
cnfmtrx22 <- table(prd=predicted22, training$Attrition)
confusionMatrix(cnfmtrx22)
```

```
## Confusion Matrix and Statistics
##
##
## prd   0   1
##   0 837  93
##   1  33  66
##
##                Accuracy : 0.8776
##                  95% CI : (0.856, 0.897)
##     No Information Rate : 0.8455
##     P-Value [Acc > NIR] : 0.001996
##
##                   Kappa : 0.4459
##
##  Mcnemar's Test P-Value : 1.471e-07
##
##             Sensitivity : 0.9621
##             Specificity : 0.4151
##          Pos Pred Value : 0.9000
##          Neg Pred Value : 0.6667
##              Prevalence : 0.8455
##          Detection Rate : 0.8134
##    Detection Prevalence : 0.9038
##       Balanced Accuracy : 0.6886
##
##        'Positive' Class : 0
##
```

```r
OAA_m1 <- ((cnfmtrx22[1,1]+cnfmtrx22[2,2])/sum(cnfmtrx22))
OAA_m1
```

```
## [1] 0.877551
```

```r
#Model 2 confusion matrix
predicted33 <- ifelse(predict3 > 0.48, 1, 0)
cnfmtrx33 <- table(prd=predicted33, training$Attrition)
confusionMatrix(cnfmtrx33)
```

```
## Confusion Matrix and Statistics
##
##
## prd   0   1
##   0 846 111
##   1  24  48
##
##                Accuracy : 0.8688
##                  95% CI : (0.8466, 0.8888)
##     No Information Rate : 0.8455
##     P-Value [Acc > NIR] : 0.01969
##
##                   Kappa : 0.3533
##
##  Mcnemar's Test P-Value : 1.345e-13
##
##             Sensitivity : 0.9724
##             Specificity : 0.3019
##          Pos Pred Value : 0.8840
##          Neg Pred Value : 0.6667
##              Prevalence : 0.8455
##          Detection Rate : 0.8222
##    Detection Prevalence : 0.9300
##       Balanced Accuracy : 0.6372
##
##        'Positive' Class : 0
##
```

```r
OAA_m2 <- ((cnfmtrx33[1,1]+cnfmtrx33[2,2])/sum(cnfmtrx33))
OAA_m2
```

```
## [1] 0.8688047
```

```r
OAA_LG <- OAA_m1
#----Plotting Logistic Curve----
Plotmodl0 <- mutate(training, PrdVal=predict1, POutcome=predicted11)
head(Plotmodl0)
```

```
##   Age DistanceFromHome MonthlyIncome YearsSinceLastPromotion NumCompaniesW
orked
## 1  41               1          5993                        0
```

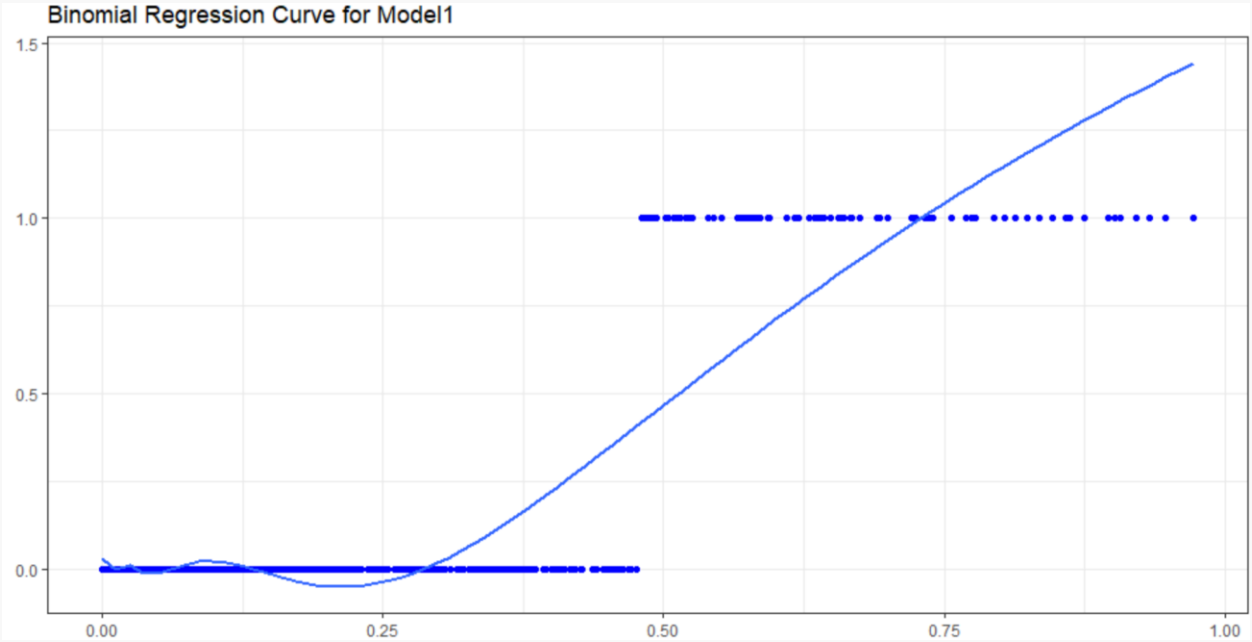
```
8
## 2  33                3          2909                        3
1
## 3  27                2          3468                        2
9
## 4  32                2          3068                        3
0
## 5  59                3          2670                        0
4
## 6  30               24          2693                        0
1
##   YearsInCurrentRole YearsWithCurrManager    BusinessTravel
## 1                  4                    5     Travel_Rarely
## 2                  7                    0 Travel_Frequently
## 3                  2                    2     Travel_Rarely
## 4                  7                    6 Travel_Frequently
## 5                  0                    0     Travel_Rarely
## 6                  0                    0     Travel_Rarely
##   EnvironmentSatisfaction JobInvolvement              JobRole OverTime
## 1                  Medium           High       Sales Executive      Yes
## 2               Very High           High    Research Scientist      Yes
## 3                     Low           High Laboratory Technician       No
## 4               Very High           High Laboratory Technician       No
## 5                    High      Very High Laboratory Technician      Yes
## 6               Very High           High Laboratory Technician       No
##   MaritalStatus WorkLifeBalance JobSatisfaction Attrition   predict1   pre
dict2
## 1        Single             Bad       Very High         1 0.48525492 0.371
07650
## 2       Married          Better            High         0 0.20953990 0.218
33340
## 3       Married          Better          Medium         0 0.46382153 0.192
38302
## 4        Single            Good       Very High         0 0.05949131 0.138
73693
## 5       Married            Good             Low         0 0.12924833 0.084
54473
## 6      Divorced          Better            High         0 0.07609667 0.070
74117
##     predict3     PrdVal POutcome
## 1 0.63014152 0.48525492        1
## 2 0.23497558 0.20953990        0
## 3 0.21528733 0.46382153        0
## 4 0.27377100 0.05949131        0
## 5 0.10950097 0.12924833        0
## 6 0.03845742 0.07609667        0

Plotmodl1 <- mutate(training, PrdVal=predict2, POutcome=predicted22)
head(Plotmodl1)
```

```
##   Age DistanceFromHome MonthlyIncome YearsSinceLastPromotion NumCompaniesW
orked
## 1  41               1          5993                       0
8
## 2  33               3          2909                       3
1
## 3  27               2          3468                       2
9
## 4  32               2          3068                       3
0
## 5  59               3          2670                       0
4
## 6  30              24          2693                       0
1
##   YearsInCurrentRole YearsWithCurrManager    BusinessTravel
## 1                  4                    5     Travel_Rarely
## 2                  7                    0 Travel_Frequently
## 3                  2                    2     Travel_Rarely
## 4                  7                    6 Travel_Frequently
## 5                  0                    0     Travel_Rarely
## 6                  0                    0     Travel_Rarely
##   EnvironmentSatisfaction JobInvolvement              JobRole OverTime
## 1                  Medium          High       Sales Executive      Yes
## 2               Very High          High    Research Scientist      Yes
## 3                     Low          High Laboratory Technician       No
## 4               Very High          High Laboratory Technician       No
## 5                    High     Very High Laboratory Technician      Yes
## 6               Very High          High Laboratory Technician       No
##   MaritalStatus WorkLifeBalance JobSatisfaction Attrition   predict1   pre
dict2
## 1        Single             Bad       Very High         1 0.48525492 0.371
07650
## 2       Married          Better            High         0 0.20953990 0.218
33340
## 3       Married          Better          Medium         0 0.46382153 0.192
38302
## 4        Single            Good       Very High         0 0.05949131 0.138
73693
## 5       Married            Good             Low         0 0.12924833 0.084
54473
## 6      Divorced          Better            High         0 0.07609667 0.070
74117
##     predict3     PrdVal POutcome
## 1 0.63014152 0.37107650        0
## 2 0.23497558 0.21833340        0
## 3 0.21528733 0.19238302        0
## 4 0.27377100 0.13873693        0
## 5 0.10950097 0.08454473        0
## 6 0.03845742 0.07074117        0
```

```
Plotmodl2 <- mutate(training, PrdVal=predict3, POutcome=predicted33)
head(Plotmodl2)

##   Age DistanceFromHome MonthlyIncome YearsSinceLastPromotion NumCompaniesW
orked
## 1  41                1          5993                        0
8
## 2  33                3          2909                        3
1
## 3  27                2          3468                        2
9
## 4  32                2          3068                        3
0
## 5  59                3          2670                        0
4
## 6  30               24          2693                        0
1
##   YearsInCurrentRole YearsWithCurrManager    BusinessTravel
## 1                  4                    5     Travel_Rarely
## 2                  7                    0 Travel_Frequently
## 3                  2                    2     Travel_Rarely
## 4                  7                    6 Travel_Frequently
## 5                  0                    0     Travel_Rarely
## 6                  0                    0     Travel_Rarely
##   EnvironmentSatisfaction JobInvolvement              JobRole OverTime
## 1                  Medium           High       Sales Executive      Yes
## 2               Very High           High    Research Scientist      Yes
## 3                     Low           High Laboratory Technician       No
## 4               Very High           High Laboratory Technician       No
## 5                    High      Very High Laboratory Technician      Yes
## 6               Very High           High Laboratory Technician       No
##   MaritalStatus WorkLifeBalance JobSatisfaction Attrition   predict1   pre
dict2
## 1        Single             Bad       Very High         1 0.48525492 0.371
07650
## 2       Married          Better            High         0 0.20953990 0.218
33340
## 3       Married          Better          Medium         0 0.46382153 0.192
38302
## 4        Single            Good       Very High         0 0.05949131 0.138
73693
## 5       Married            Good             Low         0 0.12924833 0.084
54473
## 6      Divorced          Better            High         0 0.07609667 0.070
74117
##     predict3     PrdVal POutcome
## 1 0.63014152 0.63014152        1
## 2 0.23497558 0.23497558        0
## 3 0.21528733 0.21528733        0
## 4 0.27377100 0.27377100        0
```

```
## 5 0.10950097 0.10950097        0
## 6 0.03845742 0.03845742        0
```

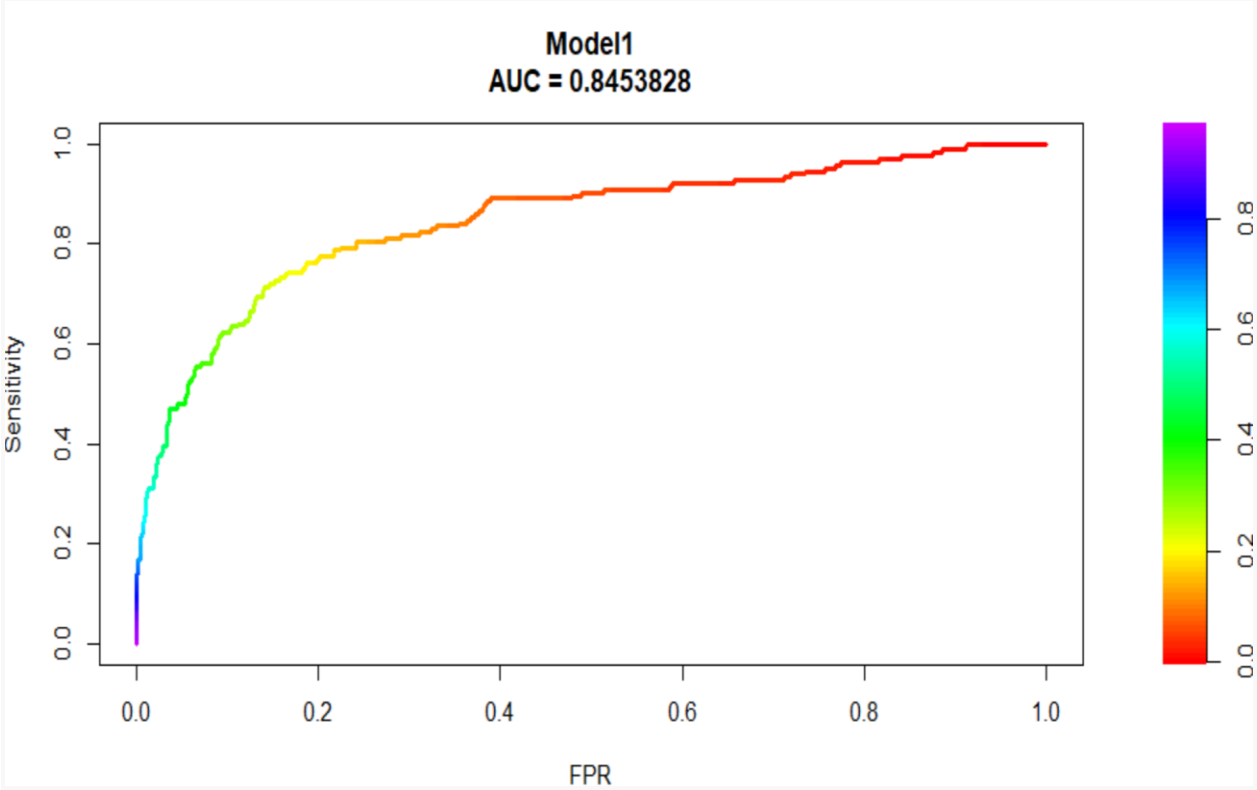
```
ggplot(Plotmodl0, aes(x=predict1, y=POutcome))  +
  geom_point(shape=19, colour="blue", fill="blue") +
  geom_smooth(method="gam", formula=y~s(log(x)), se=FALSE) +
  labs(title="Binomial Regression Curve for Model0") +
  labs(x="") +
  labs(y="")

ggplot(Plotmodl1, aes(x=predict2, y=POutcome))  +
  geom_point(shape=19, colour="blue", fill="blue") +
  geom_smooth(method="gam", formula=y~s(log(x)), se=FALSE) +
  labs(title="Binomial Regression Curve for Model1") +
  labs(x="") +
  labs(y="")

ggplot(Plotmodl2, aes(x=predict2, y=POutcome))  +
  geom_point(shape=19, colour="blue", fill="blue") +
  geom_smooth(method="gam", formula=y~s(log(x)), se=FALSE) +
  labs(title="Binomial Regression Curve for Model1") +
  labs(x="") +
  labs(y="")
```




```
#Method for Finding AUC- Area under curve

dev.off()
```

```
## null device
##           1
```

```
par(mfrow = c(2,2))
PRROC_obj1 <- roc.curve(scores.class0 = training$predict1, weights.class0=tra
ining$Attrition,
                        curve=TRUE)
plot(PRROC_obj1, main = "Model0")

PRROC_obj2 <- roc.curve(scores.class0 = training$predict2, weights.class0=tra
ining$Attrition,
                        curve=TRUE)
plot(PRROC_obj2, main = "Model1")

PRROC_obj3 <- roc.curve(scores.class0 = training$predict3, weights.class0=tra
ining$Attrition,
                        curve=TRUE)
plot(PRROC_obj3, main = "Model2")
```




```
#Finding precision and recall
library(precrec)

##
## Attaching package: 'precrec'
```

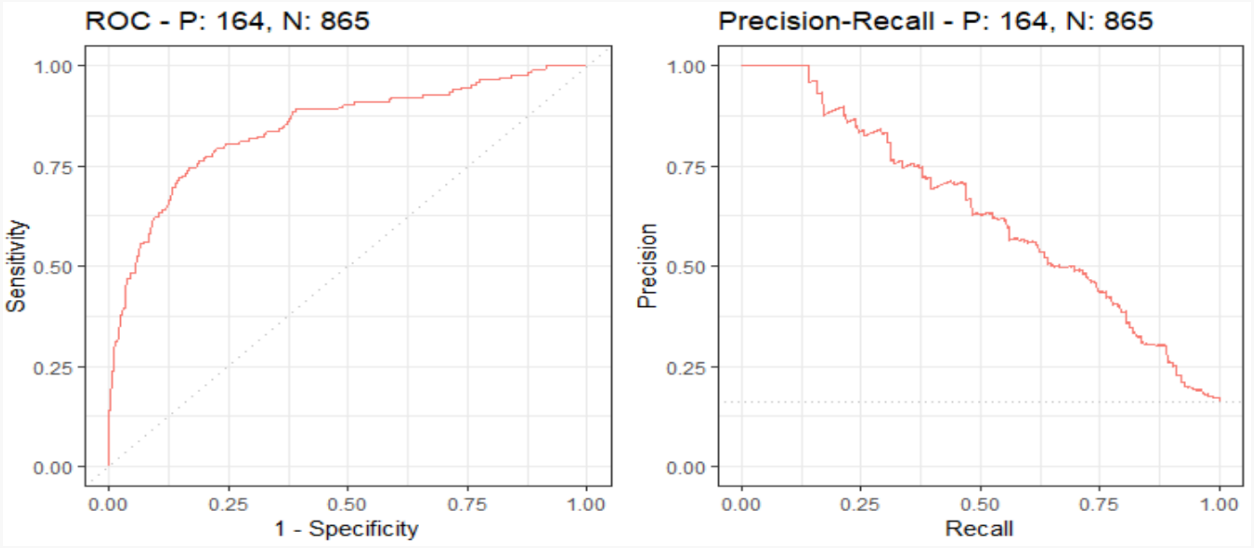
```
## The following object is masked from 'package:pROC':
##
##     auc
```

```
dev.off()
```

```
## null device
##           1
```

```
precrec_obj1 <- evalmod(scores = training$predict2, labels = training$Attriti
on)
autoplot(precrec_obj1, main = "Model0")
```

```
precrec_obj2 <- evalmod(scores = training$predict3, labels = training$Attriti
on)
autoplot(precrec_obj2, main = "Model1")
```




```
#----Predict test data----
```

```
Predict_test <- predict(model1, testing, type="response")
Predict_testbin <- ifelse(Predict_test > 0.48, 1, 0)
Predict_testbin <- as.factor(Predict_testbin)
levels(Predict_testbin) <- c("0", "1")
Prd_Test <- mutate(testing, Result=Predict_test, Prd_Outcome = Predict_testbi
n)
head(Prd_Test)
```

```
##   Age DistanceFromHome MonthlyIncome YearsSinceLastPromotion NumCompaniesW
orked
## 1  49               8          5130                        1
1
## 2  37               2          2090                        0
6
## 3  29              15          4193                        0
```

```
0
## 4  28               24          2028                        0
5
## 5  53                2         15427                        3
2
## 6  38                2          3944                        1
5
##   YearsInCurrentRole YearsWithCurrManager    BusinessTravel
## 1                  7                     7 Travel_Frequently
## 2                  0                     0     Travel_Rarely
## 3                  5                     8     Travel_Rarely
## 4                  2                     3     Travel_Rarely
## 5                  8                     7     Travel_Rarely
## 6                  2                     2     Travel_Rarely
##   EnvironmentSatisfaction JobInvolvement              JobRole OverTime
## 1                    High         Medium    Research Scientist       No
## 2               Very High         Medium Laboratory Technician      Yes
## 3               Very High         Medium Laboratory Technician      Yes
## 4                    High         Medium Laboratory Technician      Yes
## 5                     Low         Medium               Manager       No
## 6               Very High           High    Research Scientist      Yes
##   MaritalStatus WorkLifeBalance JobSatisfaction      Result Prd_Outcome
## 1       Married          Better          Medium 0.047254775           0
## 2        Single          Better            High 0.390777726           0
## 3        Single          Better            High 0.650135818           1
## 4        Single          Better            High 0.788779709           1
## 5       Married          Better       Very High 0.008469738           0
## 6        Single          Better       Very High 0.097033603           0
```

```
table(Prd_Test$Prd_Outcome)
```

```
##
##   0   1
## 402  39
```

```
write.csv(Prd_Test,"Test_Prediction.csv")
```

Observations: 1. Regression Model gives important variables that effect the attrition rate as - “Age”, “BusinessTravel”,“EnvironmentSatisfaction”, “MaritalStatus”,“JobInvolvement”, “OverTime”, “WorkLifeBalance”, “JobRole” 2. The performance of the model can be concluded by - AUC = 0.84, OAA(Overall Accuracy) = 86%, cutoff level = 0.48

```
#----Classification Model Development----
#---Decision Tree----
library(ISLR)
library(caTools)
library(tree)

training_DT <- training[, -c(17:19)]
testing_DT <- final_data[ -Train,]
str(training_DT)
```

```
## 'data.frame':    1029 obs. of  16 variables:
##  $ Age                     : int  41 33 27 32 59 30 38 36 35 31 ...
##  $ DistanceFromHome        : int  1 3 2 2 3 24 23 27 16 26 ...
##  $ MonthlyIncome           : int  5993 2909 3468 3068 2670 2693 9526 5237 2
426 2911 ...
##  $ YearsSinceLastPromotion: int  0 3 2 3 0 0 1 7 0 4 ...
##  $ NumCompaniesWorked      : int  8 1 9 0 4 1 0 6 0 1 ...
##  $ YearsInCurrentRole      : int  4 7 2 7 0 0 7 7 4 2 ...
##  $ YearsWithCurrManager    : int  5 0 2 6 0 0 8 7 3 3 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Freque
ntly",..: 3 2 3 2 3 3 2 3 3 3 ...
##  $ EnvironmentSatisfaction: Factor w/ 4 levels "High","Low","Medium",..: 3
4 2 4 1 4 4 1 2 2 ...
##  $ JobInvolvement          : Factor w/ 4 levels "High","Low","Medium",..: 1
1 1 1 4 1 3 1 4 1 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative"
,..: 8 7 3 3 3 3 5 1 3 7 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1
1 1 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3
2 2 3 2 1 3 2 2 1 ...
##  $ WorkLifeBalance         : Factor w/ 4 levels "Bad","Best","Better",..: 1
3 3 4 4 3 3 4 3 4 ...
##  $ JobSatisfaction         : Factor w/ 4 levels "High","Low","Medium",..: 4
1 3 4 2 1 1 1 3 1 ...
##  $ Attrition               : num  1 0 0 0 0 0 0 0 0 0 ...
```

```
str(testing_DT)
```

```
## 'data.frame':    441 obs. of  16 variables:
##  $ Age                     : int  49 37 29 28 53 38 21 32 44 46 ...
##  $ DistanceFromHome        : int  8 2 15 24 2 2 15 16 7 2 ...
##  $ MonthlyIncome           : int  5130 2090 4193 2028 15427 3944 1232 3919
10248 18947 ...
##  $ YearsSinceLastPromotion: int  1 0 0 0 3 1 0 6 5 2 ...
##  $ NumCompaniesWorked      : int  1 6 0 5 2 5 1 1 3 3 ...
##  $ YearsInCurrentRole      : int  7 0 5 2 8 2 0 2 6 2 ...
##  $ YearsWithCurrManager    : int  7 0 8 3 7 2 0 7 17 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Freque
ntly",..: 2 3 3 3 3 3 3 2 3 3 ...
##  $ EnvironmentSatisfaction: Factor w/ 4 levels "High","Low","Medium",..: 1
4 4 1 2 4 1 3 2 3 ...
##  $ JobInvolvement          : Factor w/ 4 levels "High","Low","Medium",..: 3
3 3 3 3 1 1 2 3 1 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative"
,..: 7 3 3 3 4 7 7 7 1 4 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 2 1 2
1 1 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 2
3 3 3 2 3 3 3 2 3 ...
```

```
##  $ WorkLifeBalance        : Factor w/ 4 levels "Bad","Best","Better",..: 3
3 3 3 3 3 3 3 3 4 ...
##  $ JobSatisfaction        : Factor w/ 4 levels "High","Low","Medium",..: 3
1 1 1 4 4 4 2 4 2 ...
##  $ Attrition              : num  0 1 0 1 0 0 0 1 0 0 ...
```

```
final_data$Attrition <- as.factor(final_data$Attrition)
training$Attrition <- as.factor(training$Attrition)
testing_DT$Attrition<- as.factor(testing_DT$Attrition)
```

```
table(final_data$Attrition)
```

```
##
##    0    1
## 1233  237
```

```
prop.table(table(final_data$Attrition))
```

```
##
##         0         1
## 0.8387755 0.1612245
```

```
table(training$Attrition)
```

```
##
##   0   1
## 870 159
```

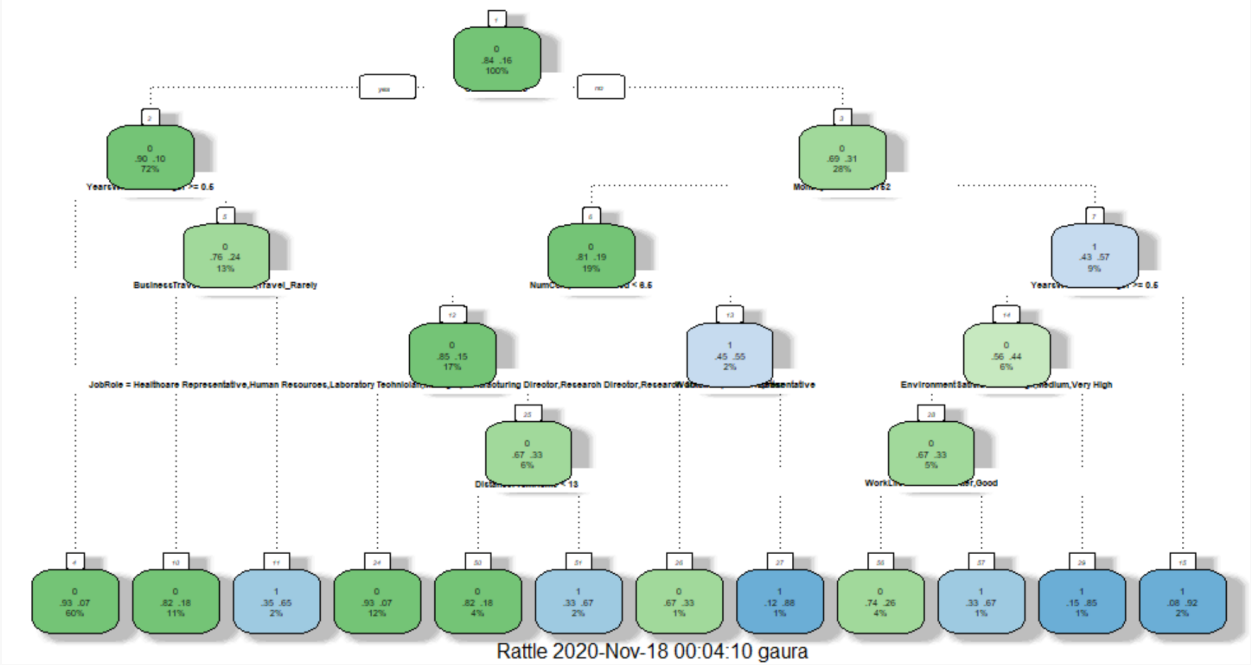
```
prop.table(table(training$Attrition))
```

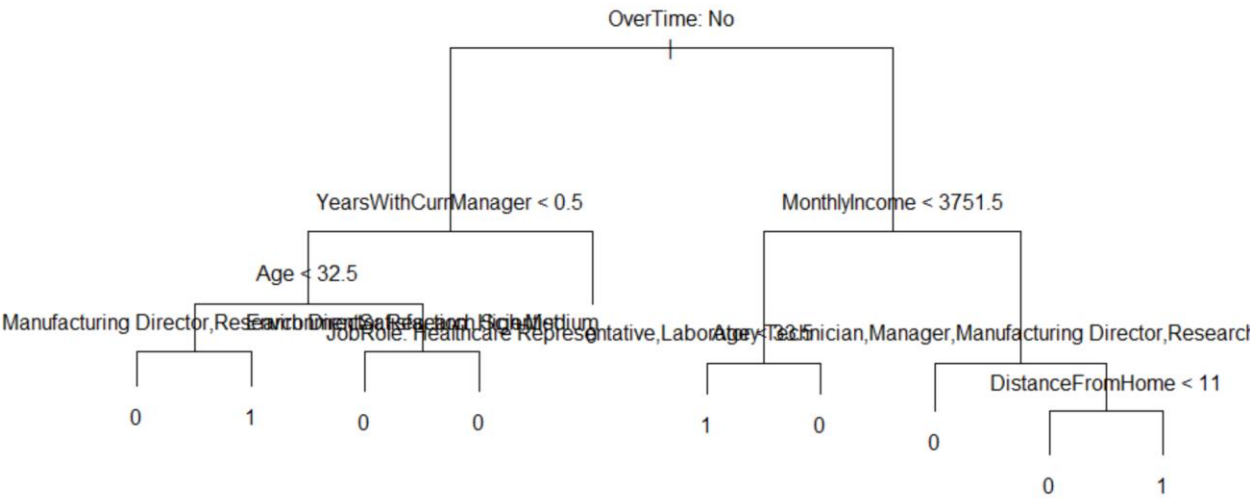
```
##
##        0        1
## 0.845481 0.154519
```

```
#------Building the model on entire Attrition data set----
tree.attrition <- tree(Attrition~.,final_data)
summary(tree.attrition)
```

```
##
## Classification tree:
## tree(formula = Attrition ~ ., data = final_data)
## Variables actually used in tree construction:
## [1] "OverTime"                "YearsWithCurrManager"
## [3] "Age"                     "JobRole"
## [5] "EnvironmentSatisfaction" "MonthlyIncome"
## [7] "DistanceFromHome"
## Number of terminal nodes:  10
## Residual mean deviance:  0.6989 = 1020 / 1460
## Misclassification error rate: 0.134 = 197 / 1470
```

```
plot(tree.attrition)
text(tree.attrition,pretty = 0)
```



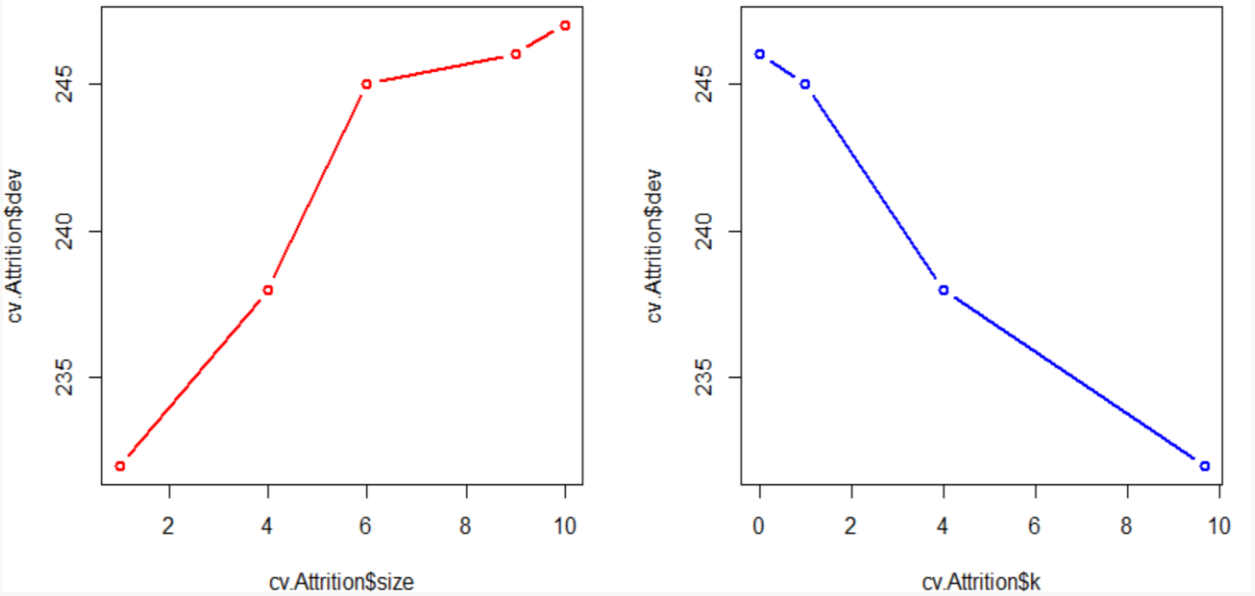



```
library(rattle)

## Warning: package 'rattle' was built under R version 4.0.3

## Loading required package: bitops

## Registered S3 method overwritten by 'rattle':
##   method          from
##   predict.kmeans parameters

## 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 object is masked from 'package:randomForest':
##
##     importance

library(rpart)
dev.off()

## null device
##           1

DTreeModel <- rpart(Attrition~.,data=training_DT,method="class")
fancyRpartPlot(DTreeModel, tweak = 1.25, cex= 0.35)

## Warning: cex and tweak both specified, applying both

predDT <- predict(DTreeModel,newdata = testing_DT,type = "class")
pred_table <- table(testing_DT$Attrition,predDT)
OAA_DT <- ((pred_table[1,1]+pred_table[2,2])/sum(pred_table))
OAA_DT

## [1] 0.8321995

#----Prediction on same Attrition ie entire data set---
tree.pred <- predict(tree.attrition,final_data,type = "class")
conf_tree.pred <- table(tree.pred,final_data$Attrition)
conf_tree.pred

##
## tree.pred    0    1
##         0 1167  131
##         1   66  106

#-----Calculating Classification Accuracy of Attrition data set---
OAA_DT2 <- ((conf_tree.pred[1,1]+conf_tree.pred[2,2])/sum(conf_tree.pred))
OAA_DT2

## [1] 0.8659864

#----Prune the tree----
set.seed(123)
cv.Attrition <- cv.tree(tree.attrition, FUN = prune.misclass)
names(cv.Attrition)

## [1] "size"   "dev"    "k"      "method"

cv.Attrition

## $size
## [1] 10  9  6  4  1
##
```

```
## $dev
## [1] 247 246 245 238 232
##
## $k
## [1]     -Inf 0.000000 1.000000 4.000000 9.666667
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

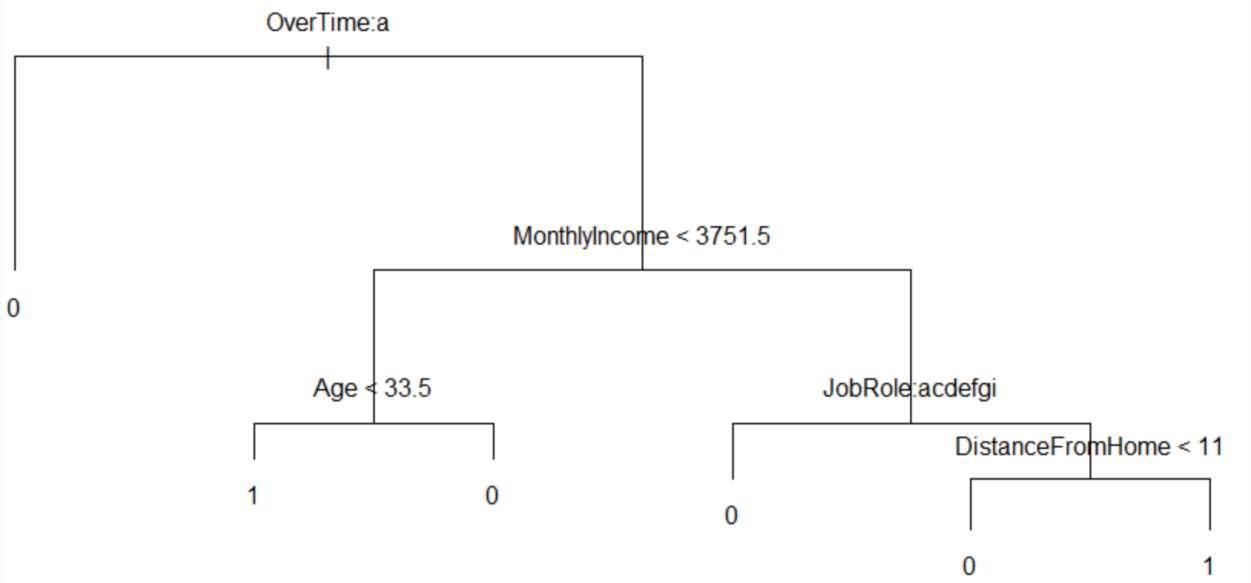
```
#----Plotting error----
par(mfrow=c(1,2))
plot(cv.Attrition$size,cv.Attrition$dev,type="b",col="red",lwd=2)
plot(cv.Attrition$k,cv.Attrition$dev,type="b",col="blue",lwd=2)
```




```
#---Again build a tree with 6 terminal nodes---
prune.Attrition <- prune.misclass(tree.attrition,best=6)
dev.off()
```

```
## null device
##           1
```

```
plot(prune.Attrition)
text(prune.Attrition,petty=0)
```




```
## Warning in text.default(xy$x[ind], xy$y[ind] + 0.5 * charht, rows[ind], :
## "petty" is not a graphical parameter

## Warning in text.default(xy$x[leaves], xy$y[leaves] - 0.5 * charht, labels
=
## stat, : "petty" is not a graphical parameter
```

```
#----Using the prune model on test data for prediction---
pred_prune <- predict(prune.Attrition, testing_DT,type = "class")
conf_prune_tree_pred1 <- table(pred_prune, testing_DT$Attrition)
conf_prune_tree_pred1
```

```
##
## pred_prune   0   1
##          0 353  53
##          1  10  25
```

```
OAA_PR <- ((conf_prune_tree_pred1[1,1]+conf_prune_tree_pred1[2,2])/sum(conf_p
rune_tree_pred1))
OAA_PR
```

```
## [1] 0.8571429
```

```
#-------Bagging------
library(randomForest)
#----Bagging will take all variables so mtry=15 ie all 15 variables except "A
ttrition"-----
set.seed(123)
bag.attrition <- randomForest(Attrition~.,final_data, subset=Train, mtry=22)
```

```
## Warning in randomForest.default(m, y, ...): invalid mtry: reset to within
valid
## range
```

```r
dim(final_data)
```

```
## [1] 1470   16
```

```r
#importance(bag.attrition)
varImpPlot(bag.attrition,col="red",pch=10,cex=1.25)
bag.attrition
```

```
##
## Call:
##  randomForest(formula = Attrition ~ ., data = final_data, mtry = 22,
subset = Train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 15
##
##         OOB estimate of  error rate: 13.51%
## Confusion matrix:
##     0  1 class.error
## 0 848 22  0.02528736
## 1 117 42  0.73584906
```

```r
str(train)
```

```
## function (x, ...)
```

```r
#----Using bagging model on test data for prediction----
pred_bag <- predict(bag.attrition,testing_DT,type = "class")
conf_bag_pred_test<-table(pred_bag,testing_DT$Attrition)
conf_bag_pred_test
```

```
##
## pred_bag   0   1
##        0 355  54
##        1   8  24
```

```r
OAA_BG <- (conf_bag_pred_test[1,1]+conf_bag_pred_test[2,2])/sum(conf_bag_pred
_test)
OAA_BG
```

```
## [1] 0.8594104
```

```r
#----Random forest will take selected variables ie ~SQRT22(as there are 22 pr
edictors)= mtry=4.6 ~ 5
rf.attrition <- randomForest(Attrition~.,final_data, subset = Train, mtry=5)
dim(final_data)
```

```
## [1] 1470   16
```

```r
#importance(rf.attrition)
varImpPlot(rf.attrition, col = "red", pch = 10, cex = 1.25)
```

```r
#-----Using random forest on test data for prediction----
pred_rf <- predict(rf.attrition, testing_DT, type = "class")
conf_test_pred_rf <- table(pred_rf, testing_DT$Attrition)
conf_test_pred_rf
```

```
##
## pred_rf   0   1
##       0 359  60
##       1   4  18
```

```r
OAA_RF <- (conf_test_pred_rf[1,1]+conf_test_pred_rf[2,2])/sum(conf_test_pred_
rf)
OAA_RF
```

```
## [1] 0.8548753
```

```r
#---Final Conclusion----

glm_ROC <- predict(model1,type = "response")
pred_glm <- prediction(training$predict2,training$Attrition)
perf_glm <- performance(pred2,"tpr","fpr")

dt_ROC = predict(tree.attrition,testing_DT)
pred_dt = prediction(dt_ROC[,2],testing_DT$Attrition)
perf_dt = performance(pred_dt,"tpr","fpr")

RF_ROC = predict(rf.attrition,testing_DT,type="prob")
pred_RF = prediction(RF_ROC[,2],testing_DT$Attrition)
perf_RF = performance(pred_RF,"tpr","fpr")

BG_ROC = predict(bag.attrition,testing_DT,type="prob")
pred_BG = prediction(BG_ROC[,2],testing_DT$Attrition)
perf_BG = performance(pred_BG,"tpr","fpr")

PR_ROC = predict(prune.Attrition,testing_DT)
pred_PR = prediction(PR_ROC[,2],testing_DT$Attrition)
perf_PR = performance(pred_PR,"tpr","fpr")

auc_glm <- performance(pred_glm,"auc")
auc_glm <- round(as.numeric(auc_glm@y.values),3)
auc_dt <- performance(pred_dt,"auc")
auc_dt <- round(as.numeric(auc_dt@y.values),3)
auc_RF <- performance(pred_RF,"auc")
auc_RF <- round(as.numeric(auc_RF@y.values),3)
auc_BG <- performance(pred_BG,"auc")
auc_BG <- round(as.numeric(auc_BG@y.values),3)
auc_PR <- performance(pred_PR,"auc")
auc_PR <- round(as.numeric(auc_PR@y.values),3)
print(paste('AUC of Logistic Regression:',auc_glm))
```

```
## [1] "AUC of Logistic Regression: 0.854"
```

```
print(paste('AUC of Decision Tree:',auc_dt))
```

```
## [1] "AUC of Decision Tree: 0.754"
```

```
print(paste('AUC of Random Forest:',auc_RF))
```

```
## [1] "AUC of Random Forest: 0.808"
```

```
print(paste('AUC of Bagging Tree:',auc_BG))
```

```
## [1] "AUC of Bagging Tree: 0.807"
```

```
print(paste('AUC of Pruning Tree:',auc_PR))
```

```
## [1] "AUC of Pruning Tree: 0.673"
```

```
print(paste('Accuracy of Logistic Regression Model:',round(OAA_LG*100,2), "%"
))
```

```
## [1] "Accuracy of Logistic Regression Model: 87.76 %"
```

```
print(paste('Accuracy of Decision Tree Model:',round(OAA_DT*100,2), "%"))
```

```
## [1] "Accuracy of Decision Tree Model: 83.22 %"
```

```
print(paste('Accuracy of Random Forest Model:',round(OAA_RF*100,2), "%"))
```

```
## [1] "Accuracy of Random Forest Model: 85.49 %"
```

```
print(paste('Accuracy of Bagging Tree MOdel:',round(OAA_BG*100,2), "%"))
```

```
## [1] "Accuracy of Bagging Tree MOdel: 85.94 %"
```

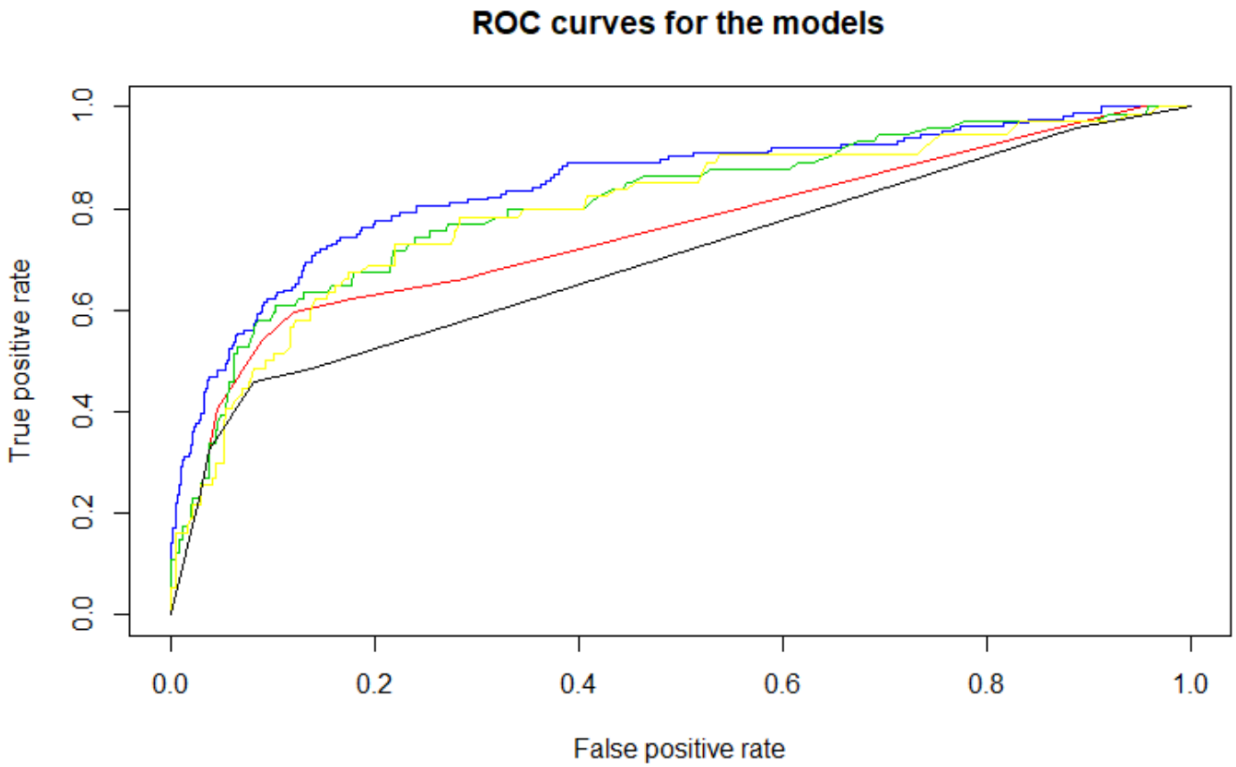
```
print(paste('Accuracy of Pruning Tree Model:',round(OAA_PR*100,2), "%"))
```

```
## [1] "Accuracy of Pruning Tree Model: 85.71 %"
```

```
dev.off()
```

```
## null device
##           1
```

```
plot(perf_glm, main = "ROC curves for the models", col='blue')
plot(perf_dt,add=TRUE, col='red')
plot(perf_RF, add=TRUE, col='green3')
plot(perf_BG,add=TRUE, col='yellow')
plot(perf_PR, add=TRUE, col='black')
legend('bottomright', c("Logistic Regression","Decision Tree",
                        "Random Forest", "Bagging","Pruning"),
       fill = c('blue','red','green3','yellow','black'),
       bty='n', cex = 0.8)
```



Final Conclusions:

1. Five modelling techniques are used to study the attrition rate of the company i.e. Logistic Regression, Decision Tree, Random Forest, Baging Tree and Pruning Tree.
2. Accuracy of Random forest is similar to the Regression Model i.e 87.5%
3. Decision tree has accuracy of 84.5% while pruning the tree increases the accuracy significantly by 3%.
4. Area under the Curve as a measure of permance of model is maximum for regression model.
5. Random Forest is simple in terms of development and also gives good accuracy.
6. As per the above analysis Regression Model gives best results with good accuracy and AUC.