Employee Attrition 2023

Olalekan Fagbuyi

2023-01-02

1. Introduction

This project aims to predict attrition of employee based on data sourced from HR department. This data includes columns for Age, Attrition, Department, Gender, Salary, Education Level, Role, Hours Worked among others.

The target variable here is the Attrition column with values Yes or No. The other 34 variables are used in predicting the target using a classification machine learning model. Performance of the model will also be evaluated using the McFadden R^2 and confusion matrix.

Both numerical

Lastly, factor importance will also be determined to ascertain the most important features driving staff to leave or stay with the company. This will enable HR can formulate appropriate staff retention strategies.

2. Importing Libraries and Loading Dataset

Libraries used for this project includes caret for Machine Learning classification models Tidyverse for manipulating dataframes Corrplot for viewing correlation among the numerical features pscl is used for calculating McFadden's R^2 to evaluate model's fit Random Forest for determining feature importance

```
library(caret)
```

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

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

```
library (tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
## ✓ purrr   0.3.4
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ purrr::lift()   masks caret::lift()
```

```
library (corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

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

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

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

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

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

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

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

```
#Loading dataset
HR_data <- read.csv("Employee Attrition.csv", header = TRUE, stringsAsFactors = FALSE)
head(HR_data)
```

```
##   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
## 6  32        No Travel_Frequently      1005 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
## 6                2         2  Life Sciences             1              8
##   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
## 6                       4   Male         79              3        1
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome MonthlyRate
## 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
## 6 Laboratory Technician               4        Single          3068       11864
##   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
## 6                  0      Y       No                13                 3
##   RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears
## 1                        1            80                0                 8
## 2                        4            80                1                10
## 3                        2            80                0                 7
## 4                        3            80                0                 8
## 5                        4            80                1                 6
## 6                        3            80                0                 8
##   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
## 6                     2               2              7                  7
##   YearsSinceLastPromotion YearsWithCurrManager
## 1                       0                    5
## 2                       1                    7
```

```
## 3                    0                    0
## 4                    3                    0
## 5                    2                    2
## 6                    3                    6
```

```
str(HR_data)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : chr  "Yes" "No" "Yes" "No" ...
##  $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Frequently" "Travel_Rarely" "Travel
_Frequently" ...
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : chr  "Sales" "Research & Development" "Research & Development"
"Research & Development" ...
##  $ 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          : chr  "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...
##  $ 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                  : chr  "Female" "Male" "Male" "Female" ...
##  $ 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                 : chr  "Sales Executive" "Research Scientist" "Laboratory Technici
an" "Research Scientist" ...
##  $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : chr  "Single" "Married" "Single" "Married" ...
##  $ 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                  : chr  "Y" "Y" "Y" "Y" ...
##  $ OverTime                : chr  "Yes" "No" "Yes" "Yes" ...
##  $ 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 ...
```

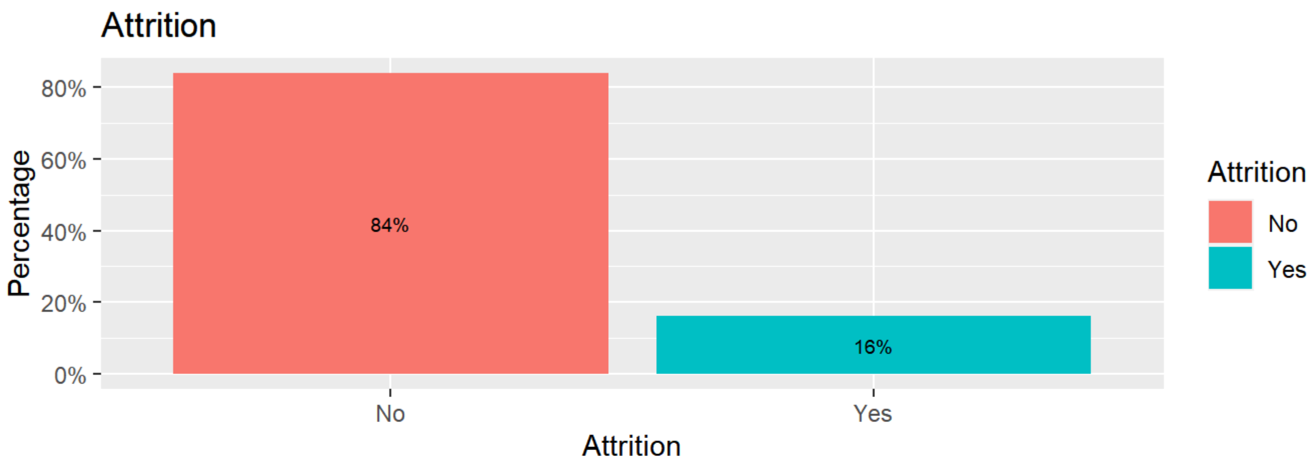
3.Exploratory Data Analyis using Visualizations

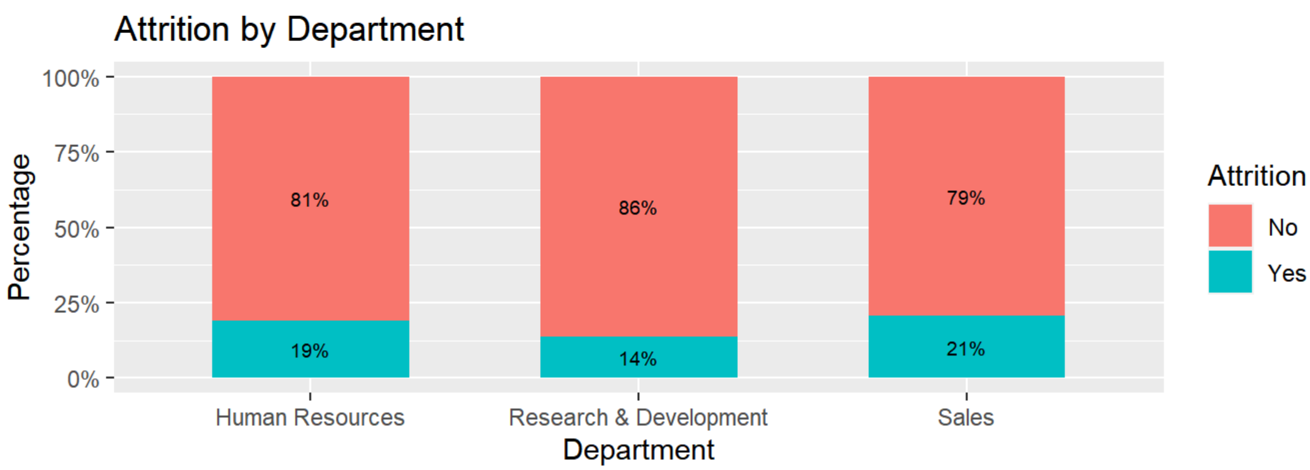
```
#Attrition rate at the company
p1 <- HR_data %>%  dplyr::group_by(Attrition) %>% dplyr::summarise(cnt = n()) %>%
  dplyr::mutate(freq = (cnt / sum(cnt))*100) %>%
  ggplot(aes(x = Attrition, y = freq, fill = Attrition)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(freq,0), "%")), position = position_stack(vjust = 0.5), siz
e = 2.5) +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +labs(title = "Attrition", x = "Attrit
ion", y ="Percentage")


p2 <- HR_data %>% dplyr::group_by(Department, Attrition) %>%  dplyr::summarise(cnt = n()) %>%
   dplyr::mutate(freq = (cnt / sum(cnt))*100) %>%  ggplot(aes(x = Department, y = freq, fill = A
ttrition)) +
   geom_bar(position = position_stack(), stat = "identity", width = .6) +
   geom_text(aes(label = paste0(round(freq,0), "%")), position = position_stack(vjust = 0.5), si
ze = 2.5) +
   scale_y_continuous(labels = function(x) paste0(x, "%")) +
  labs(title = "Attrition by Department", x = "Department", y = "Percentage")
```

```
## Warning in gzfile(file, mode): cannot open compressed file 'C:/Users/ofagb/
## AppData/Local/Temp/RtmpO2KDUA\file41784b5f2ca3', probable reason 'No such file
## or directory'
```

```
## `summarise()` has grouped output by 'Department'. You can override using the
## `.groups` argument.
```

```
grid.arrange(p1, p2, nrow = 2, ncol = 1)
```

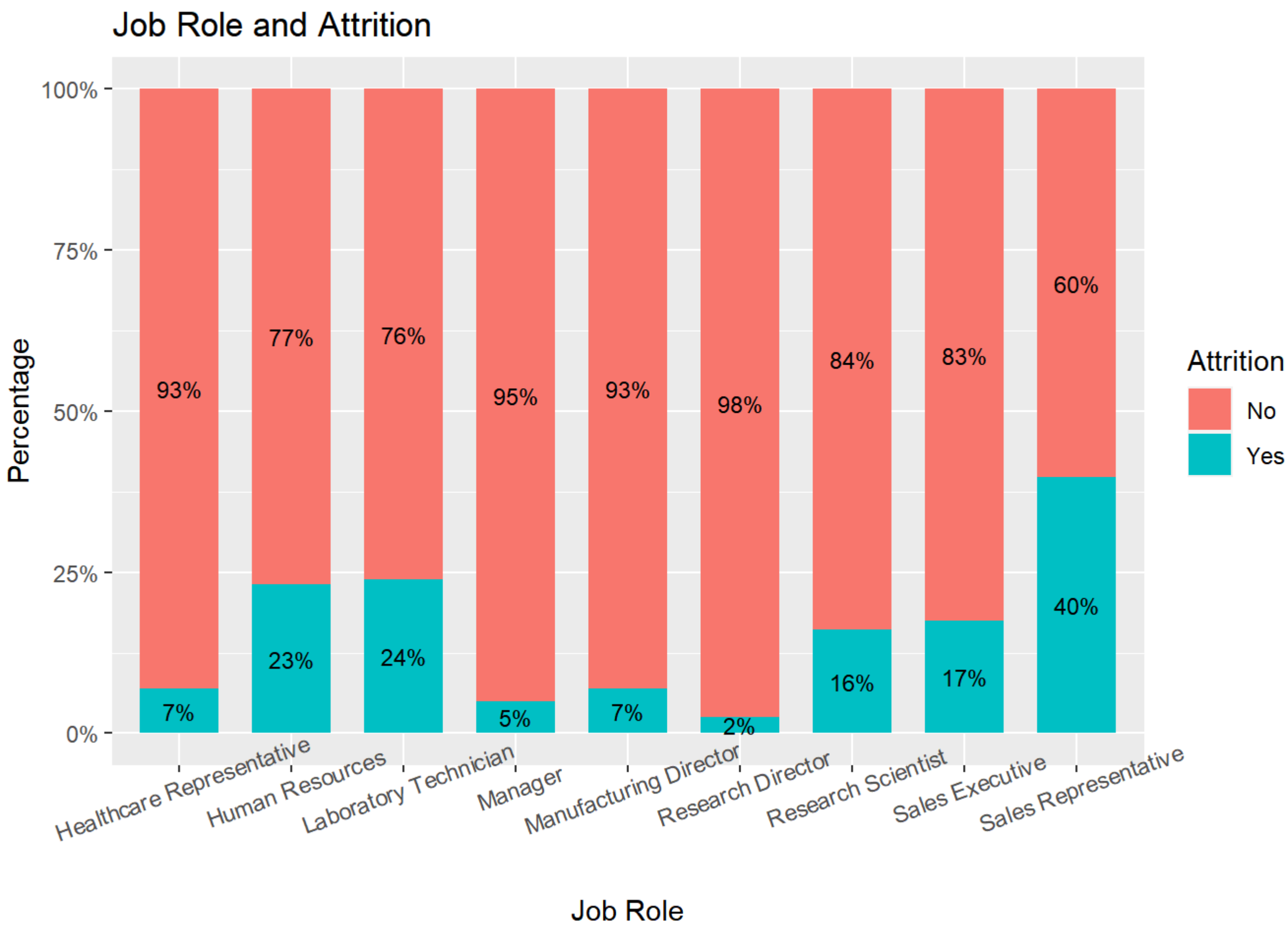




Comments - General attrition rate across the company is 16%. Sales (21%) and HR(19%) have a higher attrition than the company average while R&D department is slightly lower at 14%.

```
#Plotting data to view distr of Job Roles
HR_data %>%
  dplyr::group_by(JobRole, Attrition) %>%
  dplyr::summarise(cnt = n()) %>%
  dplyr::mutate(freq = (cnt / sum(cnt))*100) %>%
  ggplot(aes(x = JobRole, y = freq, fill = Attrition)) +
  geom_bar(position = position_stack(), stat = "identity", width = .7) +
  geom_text(aes(label = paste0(round(freq,0), "%")),
            position = position_stack(vjust = 0.5), size = 3) +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  labs(title = "Job Role and Attrition", x = "Job Role", y = "Percentage") +
  theme(axis.text.x = element_text(angle = 20, hjust = 0.5))
```

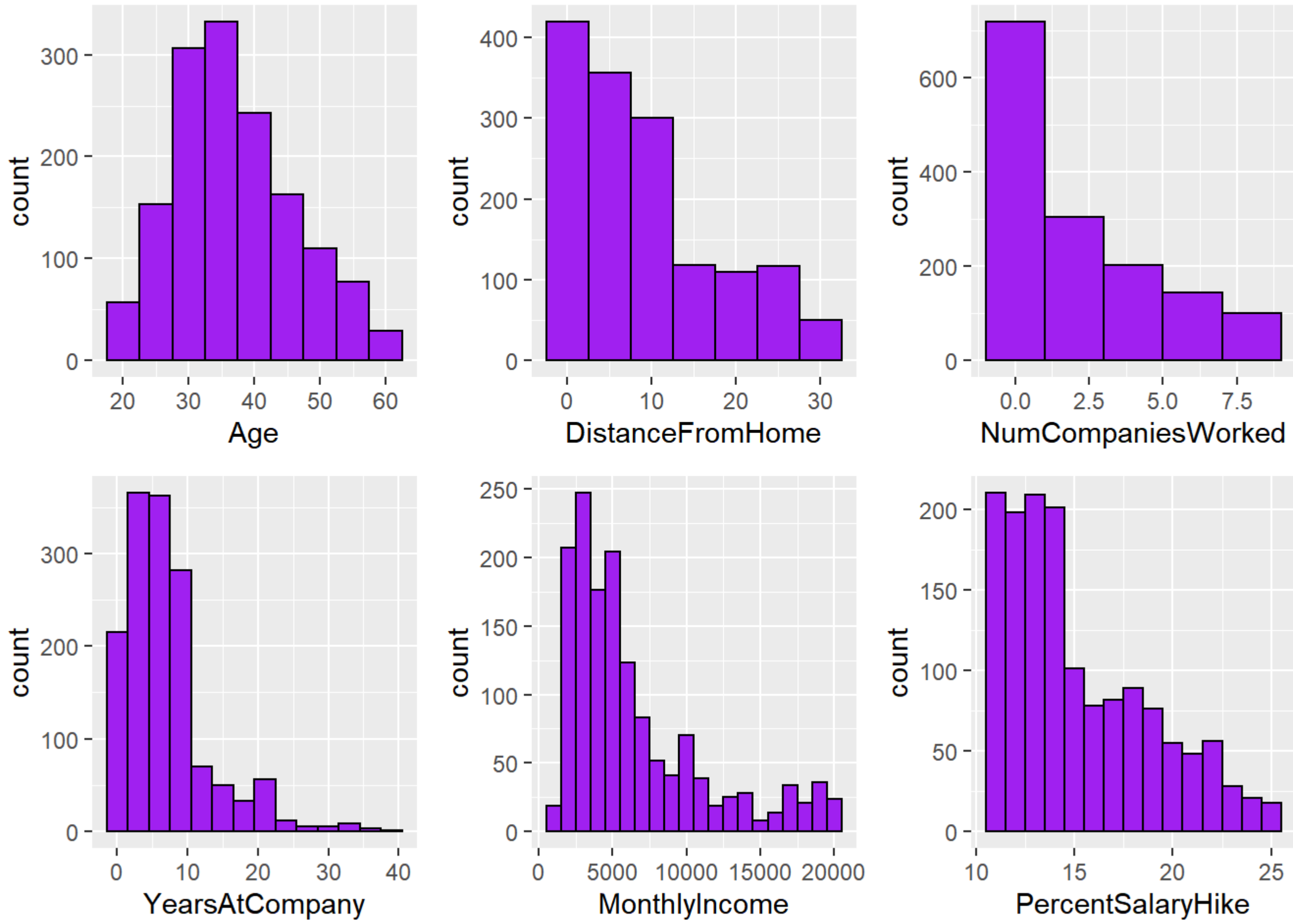
```
## `summarise()` has grouped output by 'JobRole'. You can override using the
## `.groups` argument.
```



Comments - A further dive into the Attrition numbers shows by job roles shows HR Representatives (23%), Laboratory Technicians (24%) and Sales Representatives (40%) are the key drivers of attrition at the company.

Analyzing Employee Demographics using Numerical Variables

```
p1 <- ggplot(HR_data) + geom_histogram(aes(Age), binwidth = 5, fill = "purple",col = "black")
p2 <- ggplot(HR_data) + geom_histogram(aes(DistanceFromHome), binwidth = 5, fill = "purple",col
 = "black")
p3 <- ggplot(HR_data) + geom_histogram(aes(NumCompaniesWorked), binwidth = 2, fill = "purple",co
l = "black")
p4 <- ggplot(HR_data) + geom_histogram(aes(YearsAtCompany ), binwidth = 3, fill = "purple",col =
"black")
p5 <- ggplot(HR_data) + geom_histogram(aes(MonthlyIncome), binwidth = 1000, fill = "purple",col
 = "black")
p6 <- ggplot(HR_data) + geom_histogram (aes(PercentSalaryHike), binwidth = 1, fill = "purple",co
l = "black")

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 2, ncol = 3)
```



Comments - The Age feature is close to being normally distrusted with most employees within the ages of 30 and 40. The other 5 numerical features are skewed to the right. Features will be normalized in subsequent sections

Bivariate Analysis using Numerical Variables

```
p1 <- HR_data %>%
  ggplot(aes(x = Age, fill = Attrition)) + geom_density(alpha = 0.5) + ggtitle("Age") + theme(pl
ot.title = element_text(size =10),axis.text.x = element_text(size =7,angle = 45, hjust = 1),axi
s.title.x=element_blank())


p2 <- HR_data %>%
  ggplot(aes(x = DistanceFromHome, fill = Attrition)) + geom_density(alpha = 0.5) + ggtitle("Dis
tance From Home") + theme(plot.title = element_text(size =10),axis.text.x = element_text(size =7
,angle = 45, hjust = 1),axis.title.x=element_blank())


p3 <- HR_data %>%
  ggplot(aes(x = NumCompaniesWorked, fill = Attrition)) + geom_density(alpha = 0.5) + ggtitle("N
umber of Companies Worked") + theme(plot.title = element_text(size =10),axis.text.x = element_te
xt(size =7,angle = 45, hjust = 1),axis.title.x=element_blank())


p4 <- HR_data %>%
  ggplot(aes(x = YearsAtCompany, fill = Attrition)) + geom_density(alpha = 0.5) + ggtitle("Years
at Company") + theme(plot.title = element_text(size =10),axis.text.x = element_text(size =7,angl
e = 45, hjust = 1),axis.title.x=element_blank())


p5 <- HR_data %>%
  ggplot(aes(x = MonthlyIncome, fill = Attrition)) + geom_density(alpha = 0.5) + ggtitle("Monthl
y Income") + theme(plot.title = element_text(size =10),axis.text.x = element_text(size =7,angle
 = 45, hjust = 1),axis.title.x=element_blank())


p6 <- HR_data %>%
  ggplot(aes(x = PercentSalaryHike, fill = Attrition)) + geom_density(alpha = 0.5) + ggtitle("Pe
rcent Salary Hike") + theme(plot.title = element_text(size =10),axis.text.x = element_text(size
 =7,angle = 45, hjust = 1),axis.title.x=element_blank())


grid.arrange(p1, p2, p3, p4, p5, p6 , nrow = 3, ncol = 2)
```

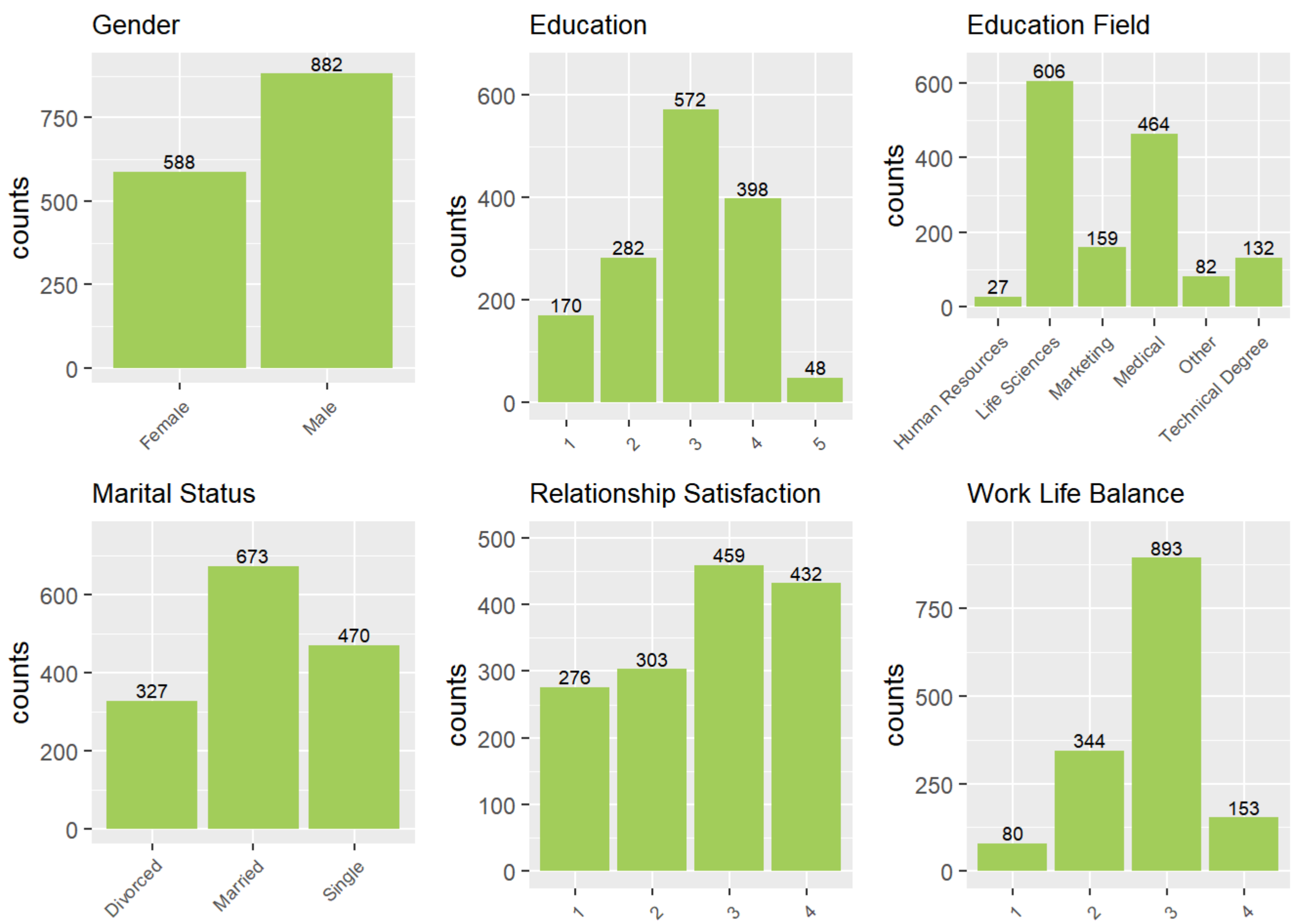
Comments - The Bivariate analysis applies EDA using 2 variables. In this case 6 of the numerical variables used earlier and the target variable (Attrition).

From the plots above, it can be seen that attritition is highest between the ages of 20-30 and also among staff that leave more than 10kms from work. In terms of salaries, staff that earn less than 5000 per month while staff that have worked for 5 or move companies have an higher attrition rate.

Analyzing Employee Demographics using Categorical Variables

```
p1<- HR_data %>%
  group_by(Gender) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(Gender), y = counts)) + geom_bar(stat = 'identity', fill = "darkolive
green3") + ggtitle("Gender") +geom_text(aes(label=counts), size = 2.5, position=position_dodge(w
idth=0.2), vjust=-0.25) + theme(plot.title = element_text(size =10),axis.text.x = element_text(s
ize =7,angle = 45, hjust = 1),axis.title.x=element_blank()) + scale_y_continuous(limits = c(0, 9
00))

p2<- HR_data %>%
  group_by(Education) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(Education), y = counts)) + geom_bar(stat = 'identity', fill = "darkol
ivegreen3") + ggtitle("Education") +geom_text(aes(label=counts), size = 2.5, position=position_d
odge(width=0.2), vjust=-0.25) + theme(plot.title = element_text(size =10),axis.text.x = element_
text(size =7,angle = 45, hjust = 1),axis.title.x=element_blank()) + scale_y_continuous(limits =
 c(0, 650))

p3 <- HR_data %>%
  group_by(EducationField) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(EducationField), y = counts)) + geom_bar(stat = 'identity', fill = "d
arkolivegreen3") + ggtitle("Education Field") +geom_text(aes(label=counts), size = 2.5, position
=position_dodge(width=0.2), vjust=-0.25) + theme(plot.title = element_text(size =10),axis.text.x
= element_text(size =7,angle = 45, hjust = 1),axis.title.x=element_blank()) + scale_y_continuous
(limits = c(0, 650))

p4 <- HR_data %>%
  group_by(MaritalStatus) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(MaritalStatus), y = counts)) + geom_bar(stat = 'identity', fill = "da
rkolivegreen3")+ ggtitle("Marital Status") +geom_text(aes(label=counts), size = 2.5, position=po
sition_dodge(width=0.2), vjust=-0.25) + theme(plot.title = element_text(size =10),axis.text.x =
 element_text(size =7,angle = 45, hjust = 1),axis.title.x=element_blank()) + scale_y_continuous
(limits = c(0, 750))

p5 <- HR_data %>%
  group_by(RelationshipSatisfaction) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(RelationshipSatisfaction), y = counts)) + geom_bar(stat = 'identity',
fill = "darkolivegreen3") + ggtitle("Relationship Satisfaction") +geom_text(aes(label=counts), s
ize = 2.5, position=position_dodge(width=0.2), vjust=-0.25) + theme(plot.title = element_text(si
ze =10),axis.text.x = element_text(size =7,angle = 45, hjust = 1),axis.title.x=element_blank())+
scale_y_continuous(limits = c(0, 500))

p6 <- HR_data %>%
  group_by(WorkLifeBalance) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(WorkLifeBalance), y = counts)) + geom_bar(stat = 'identity', fill =
"darkolivegreen3")+ ggtitle("Work Life Balance") +geom_text(aes(label=counts), size = 2.5, posit
ion=position_dodge(width=0.2), vjust=-0.25) + theme(plot.title = element_text(size =10),axis.tex
t.x = element_text(size =7,angle = 45, hjust = 1),axis.title.x=element_blank()) + scale_y_contin
```

```
uous(limits = c(0, 950))

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 2, ncol = 3)
```



Comments - The company employs more men (60%) than women (40%). Also, most employees (73%) are either from the life sciences and medical field. In terms of education qualifications, 50% of staff have at least a college education.

On a personal level, 46% of staff are married with 60% having either high or very high relationship satisfaction. Lastly, the data shows a high level of work life balance with 95% of staff choosing good to best option on the survey

4. Data Pre-processing

This stage involves making the data suitable for a machine learning model. Operations performed includes;

- Checking for null values
- Modifying/dropping highly correlated and redundant features
- Standardizing numerical features by removing outliers.
- Encoding categorical variables

Checking for null values

```
#checking for null values
sapply(HR_data, function(x) sum(is.na(x)))
```

```
##                      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
```

```
# Removing Zero and Near Zero-Variance Predictors - feature with very few unique values
nzv <- nearZeroVar(HR_data)
nzcol <- colnames(HR_data)[nzv]
nzcol
```

```
## [1] "EmployeeCount" "Over18"        "StandardHours"
```

```
#new df with redundant columns
HR_data1<- HR_data[, -nzv]
dim(HR_data1)
```

```
## [1] 1470   32
```

```
#Dropping other columns with little bearing with attrition or are better represented by other fe
atures
drop <- c("DailyRate", "EmployeeNumber","HourlyRate", "MonthlyRate" )
HR_data2 = HR_data1[,!(names(HR_data1) %in% drop)]
dim(HR_data2)
```

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

Subestting columns in the df to numeric and non-numeric

```
# numeric columns
num_cols <- unlist(lapply(HR_data2, is.numeric))
num_cols
```

```
##                      Age                Attrition           BusinessTravel
##                     TRUE                    FALSE                    FALSE
##               Department         DistanceFromHome                Education
##                    FALSE                     TRUE                     TRUE
##           EducationField  EnvironmentSatisfaction                   Gender
##                    FALSE                     TRUE                    FALSE
##           JobInvolvement                 JobLevel                  JobRole
##                     TRUE                     TRUE                    FALSE
##          JobSatisfaction            MaritalStatus            MonthlyIncome
##                     TRUE                    FALSE                     TRUE
##       NumCompaniesWorked                 OverTime        PercentSalaryHike
##                     TRUE                    FALSE                     TRUE
##        PerformanceRating RelationshipSatisfaction         StockOptionLevel
##                     TRUE                     TRUE                     TRUE
##        TotalWorkingYears    TrainingTimesLastYear          WorkLifeBalance
##                     TRUE                     TRUE                     TRUE
##           YearsAtCompany       YearsInCurrentRole  YearsSinceLastPromotion
##                     TRUE                     TRUE                     TRUE
##     YearsWithCurrManager
##                     TRUE
```

```
HR_data_num <- HR_data2[ , num_cols]
dim(HR_data_num)
```

```
## [1] 1470   20
```

```
# non-numeric columns of data
char_cols <- unlist(lapply(HR_data2, is.character))
char_cols
```

```
##                      Age                Attrition           BusinessTravel
##                    FALSE                     TRUE                     TRUE
##               Department         DistanceFromHome                Education
##                     TRUE                    FALSE                    FALSE
##           EducationField  EnvironmentSatisfaction                   Gender
##                     TRUE                    FALSE                     TRUE
##           JobInvolvement                 JobLevel                  JobRole
##                    FALSE                    FALSE                     TRUE
##          JobSatisfaction            MaritalStatus            MonthlyIncome
##                    FALSE                     TRUE                    FALSE
##       NumCompaniesWorked                 OverTime        PercentSalaryHike
##                    FALSE                     TRUE                    FALSE
##        PerformanceRating RelationshipSatisfaction         StockOptionLevel
##                    FALSE                    FALSE                    FALSE
##        TotalWorkingYears    TrainingTimesLastYear          WorkLifeBalance
##                    FALSE                    FALSE                    FALSE
##           YearsAtCompany       YearsInCurrentRole  YearsSinceLastPromotion
##                    FALSE                    FALSE                    FALSE
##     YearsWithCurrManager
##                    FALSE
```

```
# non-numeric columns of data
char_cols <- unlist(lapply(HR_data2, is.character))
char_cols
```


```
##                      Age                Attrition           BusinessTravel
##                    FALSE                     TRUE                     TRUE
##               Department         DistanceFromHome                Education
##                     TRUE                    FALSE                    FALSE
##           EducationField  EnvironmentSatisfaction                   Gender
##                     TRUE                    FALSE                     TRUE
##           JobInvolvement                 JobLevel                  JobRole
##                    FALSE                    FALSE                     TRUE
##          JobSatisfaction            MaritalStatus            MonthlyIncome
##                    FALSE                     TRUE                    FALSE
##       NumCompaniesWorked                 OverTime        PercentSalaryHike
##                    FALSE                     TRUE                    FALSE
##        PerformanceRating RelationshipSatisfaction         StockOptionLevel
##                    FALSE                    FALSE                    FALSE
##        TotalWorkingYears    TrainingTimesLastYear          WorkLifeBalance
##                    FALSE                    FALSE                    FALSE
##           YearsAtCompany       YearsInCurrentRole  YearsSinceLastPromotion
##                    FALSE                    FALSE                    FALSE
##     YearsWithCurrManager
##                    FALSE
```


```
HR_data_char <- HR_data2[ , char_cols]
dim(HR_data_char)
```

```
## [1] 1470    8
```

Checking for and removing correlated features in the numeric df

```
Cor <- round(cor(HR_data_num),2)
Cor
```

```
##                           Age DistanceFromHome Education
## Age                      1.00             0.00      0.21
## DistanceFromHome         0.00             1.00      0.02
## Education                0.21             0.02      1.00
## EnvironmentSatisfaction  0.01            -0.02     -0.03
## JobInvolvement           0.03             0.01      0.04
## JobLevel                 0.51             0.01      0.10
## JobSatisfaction          0.00             0.00     -0.01
## MonthlyIncome            0.50            -0.02      0.09
## NumCompaniesWorked       0.30            -0.03      0.13
## PercentSalaryHike        0.00             0.04     -0.01
## PerformanceRating        0.00             0.03     -0.02
## RelationshipSatisfaction 0.05             0.01     -0.01
## StockOptionLevel         0.04             0.04      0.02
## TotalWorkingYears        0.68             0.00      0.15
## TrainingTimesLastYear   -0.02            -0.04     -0.03
## WorkLifeBalance         -0.02            -0.03      0.01
## YearsAtCompany           0.31             0.01      0.07
## YearsInCurrentRole       0.21             0.02      0.06
## YearsSinceLastPromotion  0.22             0.01      0.05
## YearsWithCurrManager     0.20             0.01      0.07
##                          EnvironmentSatisfaction JobInvolvement JobLevel
## Age                                         0.01           0.03     0.51
## DistanceFromHome                           -0.02           0.01     0.01
## Education                                  -0.03           0.04     0.10
## EnvironmentSatisfaction                     1.00          -0.01     0.00
## JobInvolvement                             -0.01           1.00    -0.01
## JobLevel                                    0.00          -0.01     1.00
## JobSatisfaction                            -0.01          -0.02     0.00
## MonthlyIncome                              -0.01          -0.02     0.95
## NumCompaniesWorked                          0.01           0.02     0.14
## PercentSalaryHike                          -0.03          -0.02    -0.03
## PerformanceRating                          -0.03          -0.03    -0.02
## RelationshipSatisfaction                    0.01           0.03     0.02
## StockOptionLevel                            0.00           0.02     0.01
## TotalWorkingYears                           0.00          -0.01     0.78
## TrainingTimesLastYear                      -0.02          -0.02    -0.02
## WorkLifeBalance                             0.03          -0.01     0.04
## YearsAtCompany                              0.00          -0.02     0.53
## YearsInCurrentRole                          0.02           0.01     0.39
## YearsSinceLastPromotion                     0.02          -0.02     0.35
## YearsWithCurrManager                        0.00           0.03     0.38
##                          JobSatisfaction MonthlyIncome NumCompaniesWorked
## Age                                 0.00          0.50               0.30
## DistanceFromHome                    0.00         -0.02              -0.03
## Education                          -0.01          0.09               0.13
## EnvironmentSatisfaction            -0.01         -0.01               0.01
## JobInvolvement                     -0.02         -0.02               0.02
## JobLevel                            0.00          0.95               0.14
## JobSatisfaction                     1.00         -0.01              -0.06
## MonthlyIncome                      -0.01          1.00               0.15
## NumCompaniesWorked                 -0.06          0.15               1.00
```

```
## PercentSalaryHike               0.02        -0.03              -0.01
## PerformanceRating               0.00        -0.02              -0.01
## RelationshipSatisfaction       -0.01         0.03               0.05
## StockOptionLevel                0.01         0.01               0.03
## TotalWorkingYears              -0.02         0.77               0.24
## TrainingTimesLastYear          -0.01        -0.02              -0.07
## WorkLifeBalance                -0.02         0.03              -0.01
## YearsAtCompany                  0.00         0.51              -0.12
## YearsInCurrentRole              0.00         0.36              -0.09
## YearsSinceLastPromotion        -0.02         0.34              -0.04
## YearsWithCurrManager           -0.03         0.34              -0.11
##                          PercentSalaryHike PerformanceRating
## Age                                   0.00              0.00
## DistanceFromHome                      0.04              0.03
## Education                            -0.01             -0.02
## EnvironmentSatisfaction              -0.03             -0.03
## JobInvolvement                       -0.02             -0.03
## JobLevel                             -0.03             -0.02
## JobSatisfaction                       0.02              0.00
## MonthlyIncome                        -0.03             -0.02
## NumCompaniesWorked                   -0.01             -0.01
## PercentSalaryHike                     1.00              0.77
## PerformanceRating                     0.77              1.00
## RelationshipSatisfaction             -0.04             -0.03
## StockOptionLevel                      0.01              0.00
## TotalWorkingYears                    -0.02              0.01
## TrainingTimesLastYear                -0.01             -0.02
## WorkLifeBalance                       0.00              0.00
## YearsAtCompany                       -0.04              0.00
## YearsInCurrentRole                    0.00              0.03
## YearsSinceLastPromotion              -0.02              0.02
## YearsWithCurrManager                 -0.01              0.02
##                          RelationshipSatisfaction StockOptionLevel
## Age                                          0.05             0.04
## DistanceFromHome                             0.01             0.04
## Education                                   -0.01             0.02
## EnvironmentSatisfaction                      0.01             0.00
## JobInvolvement                               0.03             0.02
## JobLevel                                     0.02             0.01
## JobSatisfaction                             -0.01             0.01
## MonthlyIncome                                0.03             0.01
## NumCompaniesWorked                           0.05             0.03
## PercentSalaryHike                           -0.04             0.01
## PerformanceRating                           -0.03             0.00
## RelationshipSatisfaction                     1.00            -0.05
## StockOptionLevel                            -0.05             1.00
## TotalWorkingYears                            0.02             0.01
## TrainingTimesLastYear                        0.00             0.01
## WorkLifeBalance                              0.02             0.00
## YearsAtCompany                               0.02             0.02
## YearsInCurrentRole                          -0.02             0.05
## YearsSinceLastPromotion                      0.03             0.01
```

```
## YearsWithCurrManager                       0.00             0.02
##                          TotalWorkingYears TrainingTimesLastYear
## Age                                   0.68                 -0.02
## DistanceFromHome                      0.00                 -0.04
## Education                             0.15                 -0.03
## EnvironmentSatisfaction               0.00                 -0.02
## JobInvolvement                       -0.01                 -0.02
## JobLevel                              0.78                 -0.02
## JobSatisfaction                      -0.02                 -0.01
## MonthlyIncome                         0.77                 -0.02
## NumCompaniesWorked                    0.24                 -0.07
## PercentSalaryHike                    -0.02                 -0.01
## PerformanceRating                     0.01                 -0.02
## RelationshipSatisfaction              0.02                  0.00
## StockOptionLevel                      0.01                  0.01
## TotalWorkingYears                     1.00                 -0.04
## TrainingTimesLastYear                -0.04                  1.00
## WorkLifeBalance                       0.00                  0.03
## YearsAtCompany                        0.63                  0.00
## YearsInCurrentRole                    0.46                 -0.01
## YearsSinceLastPromotion               0.40                  0.00
## YearsWithCurrManager                  0.46                  0.00
##                          WorkLifeBalance YearsAtCompany YearsInCurrentRole
## Age                                -0.02           0.31               0.21
## DistanceFromHome                   -0.03           0.01               0.02
## Education                           0.01           0.07               0.06
## EnvironmentSatisfaction             0.03           0.00               0.02
## JobInvolvement                     -0.01          -0.02               0.01
## JobLevel                            0.04           0.53               0.39
## JobSatisfaction                    -0.02           0.00               0.00
## MonthlyIncome                       0.03           0.51               0.36
## NumCompaniesWorked                 -0.01          -0.12              -0.09
## PercentSalaryHike                   0.00          -0.04               0.00
## PerformanceRating                   0.00           0.00               0.03
## RelationshipSatisfaction            0.02           0.02              -0.02
## StockOptionLevel                    0.00           0.02               0.05
## TotalWorkingYears                   0.00           0.63               0.46
## TrainingTimesLastYear               0.03           0.00              -0.01
## WorkLifeBalance                     1.00           0.01               0.05
## YearsAtCompany                      0.01           1.00               0.76
## YearsInCurrentRole                  0.05           0.76               1.00
## YearsSinceLastPromotion             0.01           0.62               0.55
## YearsWithCurrManager                0.00           0.77               0.71
##                          YearsSinceLastPromotion YearsWithCurrManager
## Age                                         0.22                 0.20
## DistanceFromHome                            0.01                 0.01
## Education                                   0.05                 0.07
## EnvironmentSatisfaction                     0.02                 0.00
## JobInvolvement                             -0.02                 0.03
## JobLevel                                    0.35                 0.38
## JobSatisfaction                            -0.02                -0.03
## MonthlyIncome                               0.34                 0.34
```

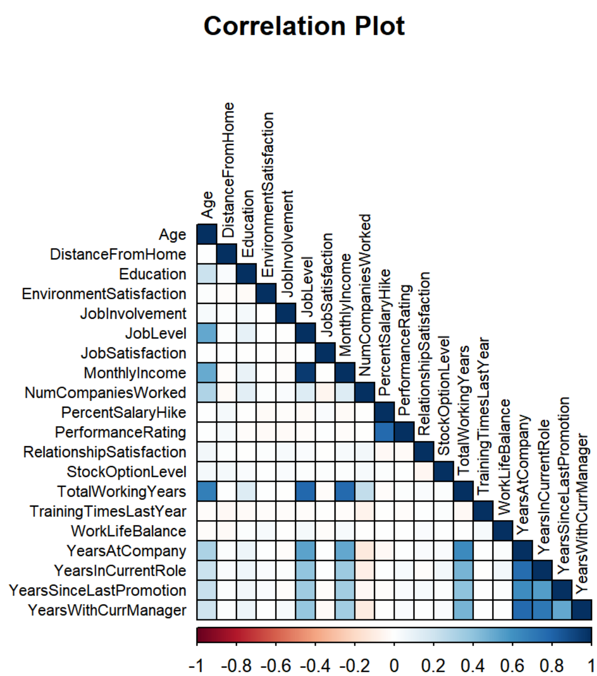
```
## NumCompaniesWorked                     -0.04               -0.11
## PercentSalaryHike                      -0.02               -0.01
## PerformanceRating                       0.02                0.02
## RelationshipSatisfaction                0.03                0.00
## StockOptionLevel                        0.01                0.02
## TotalWorkingYears                       0.40                0.46
## TrainingTimesLastYear                   0.00                0.00
## WorkLifeBalance                         0.01                0.00
## YearsAtCompany                          0.62                0.77
## YearsInCurrentRole                      0.55                0.71
## YearsSinceLastPromotion                 1.00                0.51
## YearsWithCurrManager                    0.51                1.00
```

```
summary(Cor[upper.tri(Cor)])
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.12000 -0.01000  0.01000  0.08589  0.04000  0.95000
```

Vizualizing Correlation Plot

```
corrplot(Cor, type="lower",method ="color", title = "Correlation Plot",
         mar=c(0,1,1,1), tl.cex= 0.7, outline= T, tl.col= rgb(0, 0, 0))
```




```
#Setting correlation cutoff
highlyCorrelated <- findCorrelation(Cor, cutoff = 0.6)
highlyCorCol <- colnames(HR_data_num)[highlyCorrelated]
highlyCorCol
```

```
## [1] "TotalWorkingYears"  "YearsAtCompany"     "JobLevel"
## [4] "YearsInCurrentRole" "PercentSalaryHike"
```

```
HR_data_num1 <- HR_data_num[, -which(colnames(HR_data_num) %in% highlyCorCol)]
dim(HR_data_num1)
```

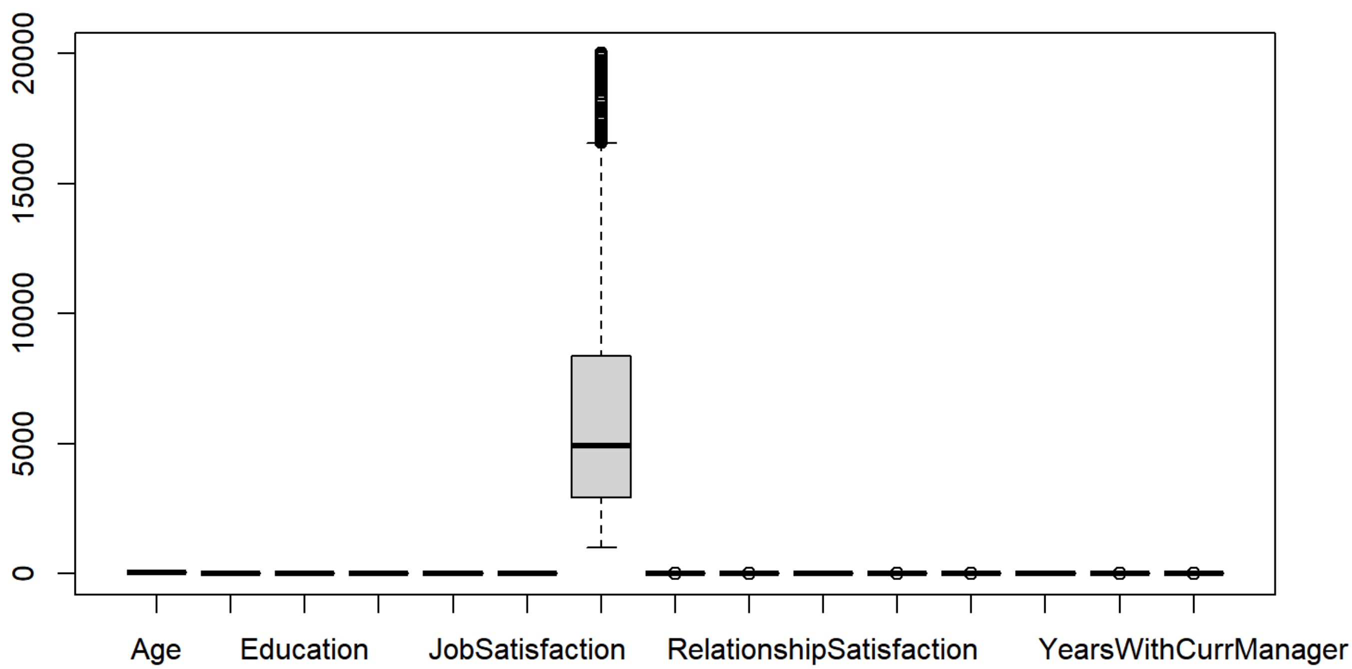
```
## [1] 1470   15
```

Standardizing data to reduce effect of outliers.

Outliers should be handled before building a statistical model as they reduce the fit and stability of the model. In order to avoid this, features are scaled using a technique called Standardization, which is a process of rescaling data so that the data have a mean of '0' and standard deviation of '1'.

Viewing distribution of numerical features using a boxplot

```
boxplot(HR_data_num1)
```



Numerical features are standardized using the scale function

```
HR_data_num2 <- HR_data_num1 %>% mutate_all(~(scale(.) %>% as.vector))
head(HR_data_num2)
```

```
##           Age DistanceFromHome  Education EnvironmentSatisfaction
## 1  0.44619856       -1.0105654 -0.8913849              -0.6603060
## 2  1.32191535       -0.1470997 -1.8677901               0.2545383
## 3  0.00834016       -0.8872132 -0.8913849               1.1693826
## 4 -0.42951824       -0.7638609  1.0614255               1.1693826
## 5 -1.08630583       -0.8872132 -1.8677901              -1.5751502
## 6 -0.53898284       -0.8872132 -0.8913849               1.1693826
##   JobInvolvement JobSatisfaction MonthlyIncome NumCompaniesWorked
## 1       0.379543       1.1528613    -0.1083127          2.1244130
## 2      -1.025818      -0.6606284    -0.2916193         -0.6778187
## 3      -1.025818       0.2461164    -0.9373347          1.3237753
## 4       0.379543       0.2461164    -0.7633739         -0.6778187
## 5       0.379543      -0.6606284    -0.6446387          2.5247318
## 6       0.379543       1.1528613    -0.7296013         -1.0781375
##   PerformanceRating RelationshipSatisfaction StockOptionLevel
## 1         -0.426085               -1.5836393       -0.9316973
## 2          2.345353                1.1910327        0.2419060
## 3         -0.426085               -0.6587487       -0.9316973
## 4         -0.426085                0.2661420       -0.9316973
## 5         -0.426085                1.1910327        0.2419060
## 6         -0.426085                0.2661420       -0.9316973
##   TrainingTimesLastYear WorkLifeBalance YearsSinceLastPromotion
## 1            -2.1712429      -2.4929720             -0.67891464
## 2             0.1556541       0.3379811             -0.36858985
## 3             0.1556541       0.3379811             -0.67891464
## 4             0.1556541       0.3379811              0.25205973
## 5             0.1556541       0.3379811             -0.05826506
## 6            -0.6199782      -1.0774954              0.25205973
##   YearsWithCurrManager
## 1            0.2457504
## 2            0.8062671
## 3           -1.1555415
## 4           -1.1555415
## 5           -0.5950247
## 6            0.5260087
```

Feature Selection for Categorical variables using Chi-Square Test

The chi-square test is a statistical test of independence to determine the dependency of two variables. It shares similarities with coefficient of determination, R^2. However, chi-square test is only applicable to categorical or nominal data while R^2 is only applicable to numeric data.

The chi-square statistics is calculated between every feature variable and the target variable. The null hypothesis for this test is the two variables are independent, and the alternative hypothesis is the variables are not independent. In order to reject the null hypothesis and keep variables in the model, the p-value of this test must have a p-value below .05

```
glimpse(HR_data_char)
```

```
## Rows: 1,470
## Columns: 8
## $ Attrition      <chr> "Yes", "No", "Yes", "No", "No", "No", "No", "No", "No",…
## $ BusinessTravel <chr> "Travel_Rarely", "Travel_Frequently", "Travel_Rarely", …
## $ Department     <chr> "Sales", "Research & Development", "Research & Developm…
## $ EducationField <chr> "Life Sciences", "Life Sciences", "Other", "Life Scienc…
## $ Gender         <chr> "Female", "Male", "Male", "Female", "Male", "Male", "Fe…
## $ JobRole        <chr> "Sales Executive", "Research Scientist", "Laboratory Te…
## $ MaritalStatus  <chr> "Single", "Married", "Single", "Married", "Married", "S…
## $ OverTime       <chr> "Yes", "No", "Yes", "Yes", "No", "No", "Yes", "No", "No…
```

```
chisq.test(HR_data_char$BusinessTravel, HR_data_char$Attrition)
```

```
##
##  Pearson's Chi-squared test
##
## data:  HR_data_char$BusinessTravel and HR_data_char$Attrition
## X-squared = 24.182, df = 2, p-value = 5.609e-06
```

```
chisq.test(HR_data_char$Department, HR_data_char$Attrition)
```

```
##
##  Pearson's Chi-squared test
##
## data:  HR_data_char$Department and HR_data_char$Attrition
## X-squared = 10.796, df = 2, p-value = 0.004526
```

```
chisq.test(HR_data_char$EducationField, HR_data_char$Attrition)
```

```
## Warning in chisq.test(HR_data_char$EducationField, HR_data_char$Attrition): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  HR_data_char$EducationField and HR_data_char$Attrition
## X-squared = 16.025, df = 5, p-value = 0.006774
```

```
chisq.test(HR_data_char$Gender, HR_data_char$Attrition)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  HR_data_char$Gender and HR_data_char$Attrition
## X-squared = 1.117, df = 1, p-value = 0.2906
```

```
chisq.test(HR_data_char$JobRole, HR_data_char$Attrition)
```

```
##
##  Pearson's Chi-squared test
##
## data:  HR_data_char$JobRole and HR_data_char$Attrition
## X-squared = 86.19, df = 8, p-value = 2.752e-15
```

```
chisq.test(HR_data_char$MaritalStatus, HR_data_char$Attrition)
```

```
##
##  Pearson's Chi-squared test
##
## data:  HR_data_char$MaritalStatus and HR_data_char$Attrition
## X-squared = 46.164, df = 2, p-value = 9.456e-11
```

```
chisq.test(HR_data_char$OverTime, HR_data_char$Attrition)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  HR_data_char$OverTime and HR_data_char$Attrition
## X-squared = 87.564, df = 1, p-value < 2.2e-16
```

From chi-square tests carried out, the gender features with be dropped because it has a p-value of > 0.05 i.e. 0.2906

```
#dropping gender column
HR_data_char = subset(HR_data_char, select = -c(Gender) )
head(HR_data_char)
```

```
##   Attrition    BusinessTravel             Department EducationField
## 1       Yes     Travel_Rarely                  Sales  Life Sciences
## 2        No Travel_Frequently Research & Development  Life Sciences
## 3       Yes     Travel_Rarely Research & Development          Other
## 4        No Travel_Frequently Research & Development  Life Sciences
## 5        No     Travel_Rarely Research & Development        Medical
## 6        No Travel_Frequently Research & Development  Life Sciences
##                 JobRole MaritalStatus OverTime
## 1       Sales Executive        Single      Yes
## 2    Research Scientist       Married       No
## 3 Laboratory Technician        Single      Yes
## 4    Research Scientist       Married      Yes
## 5 Laboratory Technician       Married       No
## 6 Laboratory Technician        Single       No
```

Encoding categorical varialbes

```
# Label encoding columns with 2 unique values
HR_data_char$Attrition[HR_data_char$Attrition == 'Yes'] <- 1
HR_data_char$Attrition[HR_data_char$Attrition == 'No'] <- 0

HR_data_char$OverTime[HR_data_char$OverTime == 'Yes'] <- 1
HR_data_char$OverTime[HR_data_char$OverTime == 'No'] <- 0

#converting columns to numeric
HR_data_char$Attrition <- as.numeric(HR_data_char$Attrition)
HR_data_char$OverTime <- as.numeric(HR_data_char$OverTime)

str(HR_data_char)
```

```
## 'data.frame':	1470 obs. of  7 variables:
##  $ Attrition     : num  1 0 1 0 0 0 0 0 0 0 ...
##  $ BusinessTravel: chr  "Travel_Rarely" "Travel_Frequently" "Travel_Rarely" "Travel_Frequentl
y" ...
##  $ Department    : chr  "Sales" "Research & Development" "Research & Development" "Research &
Development" ...
##  $ EducationField: chr  "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...
##  $ JobRole       : chr  "Sales Executive" "Research Scientist" "Laboratory Technician" "Resea
rch Scientist" ...
##  $ MaritalStatus : chr  "Single" "Married" "Single" "Married" ...
##  $ OverTime      : num  1 0 1 1 0 0 1 0 0 0 ...
```

```
#one-hot encoding columns with more than 2 unique values
dummy <- dummyVars(" ~ .", data = HR_data_char)
HR_data_char1 <- data.frame(predict(dummy, newdata = HR_data_char))

str(HR_data_char1)
```

```
## 'data.frame':    1470 obs. of  26 variables:
##  $ Attrition                        : num  1 0 1 0 0 0 0 0 0 0 ...
##  $ BusinessTravelNon.Travel         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BusinessTravelTravel_Frequently : num  0 1 0 1 0 1 0 0 1 0 ...
##  $ BusinessTravelTravel_Rarely      : num  1 0 1 0 1 0 1 1 0 1 ...
##  $ DepartmentHuman.Resources        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DepartmentResearch...Development: num  0 1 1 1 1 1 1 1 1 1 ...
##  $ DepartmentSales                  : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ EducationFieldHuman.Resources    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ EducationFieldLife.Sciences      : num  1 1 0 1 0 1 0 1 1 0 ...
##  $ EducationFieldMarketing          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ EducationFieldMedical            : num  0 0 0 0 1 0 1 0 0 1 ...
##  $ EducationFieldOther              : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ EducationFieldTechnical.Degree   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JobRoleHealthcare.Representative: num  0 0 0 0 0 0 0 0 0 1 ...
##  $ JobRoleHuman.Resources           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JobRoleLaboratory.Technician     : num  0 0 1 0 1 1 1 1 0 0 ...
##  $ JobRoleManager                   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JobRoleManufacturing.Director    : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ JobRoleResearch.Director         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JobRoleResearch.Scientist        : num  0 1 0 1 0 0 0 0 0 0 ...
##  $ JobRoleSales.Executive           : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ JobRoleSales.Representative      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MaritalStatusDivorced            : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ MaritalStatusMarried             : num  0 1 0 1 1 0 1 0 0 1 ...
##  $ MaritalStatusSingle              : num  1 0 1 0 0 1 0 0 1 0 ...
##  $ OverTime                         : num  1 0 1 1 0 0 1 0 0 0 ...
```

```
#Binding categorical and numerical dfs to form new (complete) df
HR_data2 <- cbind(HR_data_char1, HR_data_num2)
glimpse(HR_data2)
```

```
## Rows: 1,470
## Columns: 41
## $ Attrition                        <dbl> 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ BusinessTravelNon.Travel         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ BusinessTravelTravel_Frequently  <dbl> 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0…
## $ BusinessTravelTravel_Rarely      <dbl> 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1…
## $ DepartmentHuman.Resources        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ DepartmentResearch...Development <dbl> 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ DepartmentSales                  <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ EducationFieldHuman.Resources    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ EducationFieldLife.Sciences      <dbl> 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1…
## $ EducationFieldMarketing          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ EducationFieldMedical            <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0…
## $ EducationFieldOther              <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ EducationFieldTechnical.Degree   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ JobRoleHealthcare.Representative <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0…
## $ JobRoleHuman.Resources           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ JobRoleLaboratory.Technician     <dbl> 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0…
## $ JobRoleManager                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ JobRoleManufacturing.Director    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0…
## $ JobRoleResearch.Director         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ JobRoleResearch.Scientist        <dbl> 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1…
## $ JobRoleSales.Executive           <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ JobRoleSales.Representative      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ MaritalStatusDivorced            <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1…
## $ MaritalStatusMarried             <dbl> 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0…
## $ MaritalStatusSingle              <dbl> 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0…
## $ OverTime                         <dbl> 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0…
## $ Age                              <dbl> 0.44619856, 1.32191535, 0.00834016, -…
## $ DistanceFromHome                 <dbl> -1.01056544, -0.14709966, -0.88721318…
## $ Education                        <dbl> -0.89138490, -1.86779013, -0.89138490…
## $ EnvironmentSatisfaction          <dbl> -0.6603060, 0.2545383, 1.1693826, 1.1…
## $ JobInvolvement                   <dbl> 0.379543, -1.025818, -1.025818, 0.379…
## $ JobSatisfaction                  <dbl> 1.1528613, -0.6606284, 0.2461164, 0.2…
## $ MonthlyIncome                    <dbl> -0.1083127, -0.2916193, -0.9373347, -…
## $ NumCompaniesWorked               <dbl> 2.1244130, -0.6778187, 1.3237753, -0.…
## $ PerformanceRating                <dbl> -0.426085, 2.345353, -0.426085, -0.42…
## $ RelationshipSatisfaction         <dbl> -1.5836393, 1.1910327, -0.6587487, 0.…
## $ StockOptionLevel                 <dbl> -0.9316973, 0.2419060, -0.9316973, -0…
## $ TrainingTimesLastYear            <dbl> -2.1712429, 0.1556541, 0.1556541, 0.1…
## $ WorkLifeBalance                  <dbl> -2.4929720, 0.3379811, 0.3379811, 0.3…
## $ YearsSinceLastPromotion          <dbl> -0.67891464, -0.36858985, -0.67891464…
## $ YearsWithCurrManager             <dbl> 0.2457504, 0.8062671, -1.1555415, -1.…
```

5. Classification - Modelling

```
# To achieve reproducible model; set the random seed number
set.seed(100)

# Data is split into training and test set in a 80:20 ratio
TrainingIndex <- createDataPartition(HR_data2$Attrition, p=0.8, list = FALSE)
TrainingSet <- HR_data2[TrainingIndex,] # Training Set
TestingSet <- HR_data2[-TrainingIndex,] # Test Set
```

```
#Model fitting
model <- glm(Attrition ~.,family=binomial(link='logit'),data = TrainingSet )
summary(model)
```

```
##
## Call:
## glm(formula = Attrition ~ ., family = binomial(link = "logit"),
##     data = TrainingSet)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0356  -0.4613  -0.2312  -0.0766   3.5945
##
## Coefficients: (5 not defined because of singularities)
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -0.66241    0.51878  -1.277  0.20165
## BusinessTravelNon.Travel          -0.86982    0.41824  -2.080  0.03755 *
## BusinessTravelTravel_Frequently    0.80515    0.24604   3.272  0.00107 **
## BusinessTravelTravel_Rarely             NA         NA      NA       NA
## DepartmentHuman.Resources        -13.33593  682.09278  -0.020  0.98440
## DepartmentResearch...Development   1.72803    1.27577   1.355  0.17558
## DepartmentSales                         NA         NA      NA       NA
## EducationFieldHuman.Resources      1.18525    1.04279   1.137  0.25570
## EducationFieldLife.Sciences       -0.81368    0.34865  -2.334  0.01961 *
## EducationFieldMarketing           -0.70680    0.46096  -1.533  0.12520
## EducationFieldMedical             -1.02297    0.35848  -2.854  0.00432 **
## EducationFieldOther               -1.29757    0.58295  -2.226  0.02602 *
## EducationFieldTechnical.Degree          NA         NA      NA       NA
## JobRoleHealthcare.Representative  -4.40468    1.44292  -3.053  0.00227 **
## JobRoleHuman.Resources            12.25560  682.09251   0.018  0.98566
## JobRoleLaboratory.Technician      -2.51441    1.32971  -1.891  0.05863 .
## JobRoleManager                    -2.40495    1.39399  -1.725  0.08449 .
## JobRoleManufacturing.Director     -3.55433    1.42154  -2.500  0.01241 *
## JobRoleResearch.Director          -5.36094    1.88490  -2.844  0.00445 **
## JobRoleResearch.Scientist         -3.40236    1.33761  -2.544  0.01097 *
## JobRoleSales.Executive            -1.07118    0.45032  -2.379  0.01737 *
## JobRoleSales.Representative             NA         NA      NA       NA
## MaritalStatusDivorced             -1.15375    0.40058  -2.880  0.00397 **
## MaritalStatusMarried              -0.85460    0.29564  -2.891  0.00384 **
## MaritalStatusSingle                     NA         NA      NA       NA
## OverTime                           2.24965    0.22788   9.872  < 2e-16 ***
## Age                               -0.45087    0.12806  -3.521  0.00043 ***
## DistanceFromHome                   0.32186    0.10022   3.212  0.00132 **
## Education                          0.19894    0.10580   1.880  0.06005 .
## EnvironmentSatisfaction           -0.55965    0.10675  -5.243 1.58e-07 ***
## JobInvolvement                    -0.39147    0.10002  -3.914 9.08e-05 ***
## JobSatisfaction                   -0.48350    0.10325  -4.683 2.83e-06 ***
## MonthlyIncome                     -0.36458    0.29363  -1.242  0.21437
## NumCompaniesWorked                 0.43758    0.10914   4.009 6.09e-05 ***
## PerformanceRating                 -0.04862    0.10846  -0.448  0.65396
## RelationshipSatisfaction          -0.26494    0.10430  -2.540  0.01108 *
## StockOptionLevel                  -0.23216    0.16239  -1.430  0.15281
## TrainingTimesLastYear             -0.20770    0.10854  -1.914  0.05568 .
## WorkLifeBalance                   -0.24826    0.10158  -2.444  0.01453 *
## YearsSinceLastPromotion            0.73246    0.14898   4.916 8.81e-07 ***
## YearsWithCurrManager              -0.71666    0.16302  -4.396 1.10e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1006.61  on 1175  degrees of freedom
## Residual deviance:  642.35  on 1140  degrees of freedom
## AIC: 714.35
##
## Number of Fisher Scoring iterations: 15
```

Model Intepretation

The coefficients indicates the average change in log odds of attrition. For instance, every unit increase in OverTime is associated with an average increase of 2.2840 in the log odds of Attrition. The p-values in the output also give us an idea of how effective each predictor variable is at predicting the probability of Attrition:

Model Evaluation

While linear models performance is measured by R^2, that of logistic regression is measured by a metric called McFadden's R^2. The value ranges from 0 to 1, in practice values over 0.40 indicates a good model fit.

We can compute McFadden's R^2 for our model using the pR2 function from the pscl package. A rule of thumb that is quite helpful is that a McFadden's pseudo R^2 ranging from 0.2 to 0.4 indicates very good model fit

```
pscl::pR2(model)["McFadden"]
```

```
## fitting null model for pseudo-r2
```

```
##  McFadden
## 0.3618759
```

A value of 0.3647089 indicates the model fits the data quite well and has high predictive power.

Confusion Matrix

The confusion matrix table in R helps in matching the predictions against actual values. It includes two dimensions, among them one will indicate the predicted values and another one will represent the actual values.

```
# Prediction on TestingSet
prediction <- predict(model, TestingSet, type = "response")
```

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

```
head(prediction)
```

```
##          5          7          9         14         22         25
## 0.23414477 0.15677665 0.02209677 0.02449423 0.62525367 0.08203140
```

```
#Assigning probabilities - If prediction exceeds threshold of 0.5, 1 else 0
prediction <- ifelse(prediction >0.5, 1, 0)
head(prediction)
```

```
##  5  7  9 14 22 25
##  0  0  0  0  1  0
```

```
#Computing confusion matrix values
confusionMatrix(factor(TestingSet$Attrition),factor(prediction), mode = 'everything', positive =
"0")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 225  12
##          1  34  23
##
##                Accuracy : 0.8435
##                  95% CI : (0.7969, 0.8831)
##     No Information Rate : 0.881
##     P-Value [Acc > NIR] : 0.97759
##
##                   Kappa : 0.4135
##
##  Mcnemar's Test P-Value : 0.00196
##
##             Sensitivity : 0.8687
##             Specificity : 0.6571
##          Pos Pred Value : 0.9494
##          Neg Pred Value : 0.4035
##               Precision : 0.9494
##                  Recall : 0.8687
##                      F1 : 0.9073
##              Prevalence : 0.8810
##          Detection Rate : 0.7653
##    Detection Prevalence : 0.8061
##       Balanced Accuracy : 0.7629
##
##        'Positive' Class : 0
##
```

Interpreting the measures in the confusion matrix:

- Accuracy 84.35% - The success rate or accuracy of the model is calculated by dividing total no. of correction predictions by total predictions (TP + TN/TP+TN+FP+FN)

- Sensitivity 86.87% - Also known as recall or True Positive Rate (TPR), sensitivity measures how often the model is correct when it predicts employee attrition TPR = (TP/TP+FN)
- Specificity 65.71% - This is the opposite of sensitivity and it measures how often the model is correct when predicts employees retention. The closer this number is to 0, the better. TNR = (TN/TN+FP)
- Precision 94.94% - Precision measures how well the model correctly predicts attrition. Precision = TP/TP+FP

6. Feature Importance

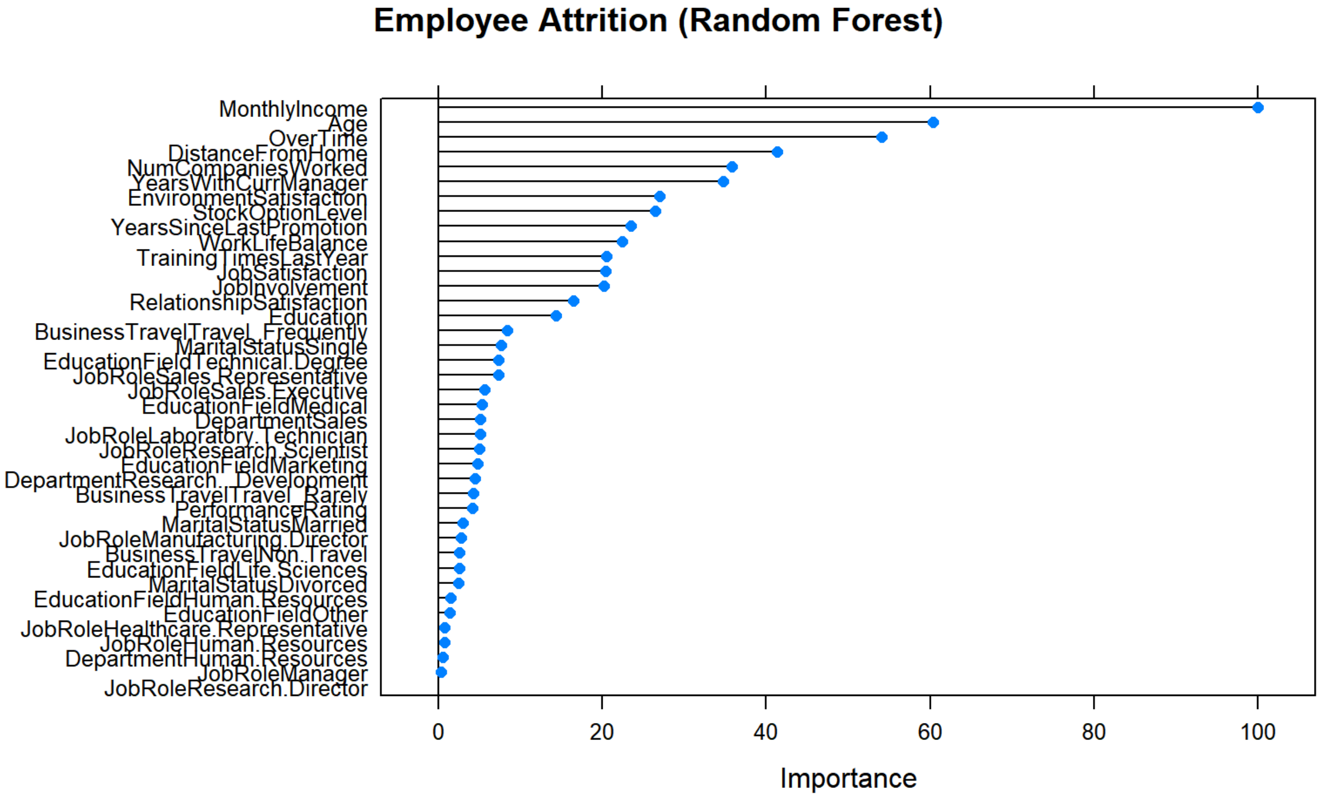
This section of focuses on ranking all the features in order of importance using the random forest algorithm in caret. A higher score means that the specific feature has a larger effect on the model in predicting the target label Attrition.

Feature importance exercise is critical because it makes it easier to identify variables to be dropped in order to reduce complexity of the model. Also, it is a straightforward way of communicating your model performance to other stakeholders.

```
set.seed(355)
rf <- train(Attrition ~., data = TrainingSet, method = "rf")
rf
```

```
## Random Forest
##
## 1176 samples
##   40 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1176, 1176, 1176, 1176, 1176, 1176, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##    2    0.3249222  0.2472995  0.2285590
##   21    0.3154840  0.2377607  0.2070822
##   40    0.3210502  0.2121344  0.2062005
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 21.
```

```
varimp_RF <- varImp(rf)
plot(varimp_RF, main = "Employee Attrition (Random Forest)")
```



7. Conclusion

To conclude, we have seen the entire process where we started with importing the dataset, getting to know the dataset at a high level, carrying out EDA (univariate & multivariate) and then moving on to data pre processing and then finally building models to predict the classification.

Every model comes with parameters which can be used to tune the models to obtain higher accuracy and specific results. e.g. in some health cases where in we want to predict if a particular person is having cancer, we need to have a model which overall may have a less accuracy but it should have has very less false negatives i.e. a person may actually have cancer but our model predicts that he does not have it. In such cases it becomes hyper parameter tuning comes into picture and we can tweak the results using it. As, we had no such specific requirements and achieved the desired levels of accuracy and AUC values we have not used hyper parameter tuning here.

Conclusion Sixteen percent of employees left the company.

In the stacked bar charts, we saw employees who left were:

In Sales Traveled frequently Worked over time Had low job satisfaction Had low environment satisfaction Had bad work life balance Chi-square results revealed gender, education, and performance rating did not have a significant role in employee attrition.

From Chi-square tests and ANOVA, statistically significant variables that affected an employee’s decision to leave include:

Monthly income Distance from home Business travel Environment satisfaction Job involvement Job role Job satisfaction Over time Stock option level Work life balance GBM with downsampling performed the best in minimizing false negatives, which will prevent us from overlooking employees that will actually leave. According to the variable importance plot, monthly income and over time are critical in attrition. Other important variables are related to work history, and distance from the office.

To prevent attrition, the company could consider raising wages, foster a company culture that promotes work life balance, and allow remote work so employees don’t have long commutes. Remote work will also permit flexible schedules that will aid in work life balance issues.