Prediction of Employee Attrition

Lalita Takle (NetID: ltakle2), Mihir Sircar (NetID: msircar2), Rajat Kumar (NetID: rajat3)

05/09/2022

Introduction

The success of any organization largely depends on the performance of its employees. Employee Attrition is becoming a serious problem because of the increasing competition in the corporate world and it impacts all types of businesses, irrespective of geography, industry and size of the company. Employee attrition leads to significant costs for a business, including the cost of business disruption, hiring new staff and training new staff.

Now let's go to the importance of this study or how it will solve the existing problem setting. Identifying the specific reasons and factors which might lead to employee attrition would help the company management to take necessary measures well beforehand in effort towards retaining the maximum number of employees. The HR department will then be able to focus on improving the factors which are leading to Employee dissatisfaction, resulting in reducing companies' losses.

```
library(scales)
library(plotrix)
```

```
##
## Attaching package: 'plotrix'

## The following object is masked from 'package:scales':
##
##     rescale
```

```
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(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(ROSE)
```

```
## Warning: package 'ROSE' was built under R version 4.1.3

## Loaded ROSE 0.0-4
```

```
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.1.3
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.1.3

## randomForest 4.7-1

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

##
## Attaching package: 'randomForest'

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

```
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 4.1.3

## Loaded gbm 2.1.8
```

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

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

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(scales)
library(caret)
```

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

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

```
library(pROC)
```

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

##
## Attaching package: 'pROC'

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

```
library(e1071)
library(rminer)
```

```
## Warning: package 'rminer' was built under R version 4.1.3
```

Dataset

For exploring the HR analytics domain, we have used the IBM HR Analytics dataset from Kaggle. This is a fictional dataset created by IBM data scientists for analysis purposes.

Reading the data

```r
# make sure the file is in the same path as the rmd file.
data = read.csv('HR_Employee_Attrition.csv')
head(data)
```

```
##   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
## 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
```

Checking the dimensions of data

```
dim(data)
```

```
## [1] 1470   35
```

The dataset has 1,470 data points (rows) and 35 features (columns) describing each employee's background and characteristics. Here, attrition is the response variable which we are trying to predict.

```
names(data)
```

```
##  [1] "ï..Age"                   "Attrition"
##  [3] "BusinessTravel"           "DailyRate"
##  [5] "Department"               "DistanceFromHome"
##  [7] "Education"                "EducationField"
##  [9] "EmployeeCount"            "EmployeeNumber"
## [11] "EnvironmentSatisfaction"  "Gender"
## [13] "HourlyRate"               "JobInvolvement"
## [15] "JobLevel"                 "JobRole"
## [17] "JobSatisfaction"          "MaritalStatus"
## [19] "MonthlyIncome"            "MonthlyRate"
## [21] "NumCompaniesWorked"       "Over18"
## [23] "OverTime"                 "PercentSalaryHike"
## [25] "PerformanceRating"        "RelationshipSatisfaction"
## [27] "StandardHours"            "StockOptionLevel"
## [29] "TotalWorkingYears"        "TrainingTimesLastYear"
## [31] "WorkLifeBalance"          "YearsAtCompany"
## [33] "YearsInCurrentRole"       "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"
```

Renaming the Age column correctly.

```
names(data)[1] = 'Age'
```

```
str(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_Freque
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : chr  "Sales" "Research & Development" "Research & Development" "Researc
##  $ 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 Technician" "Re
##  $ 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 ...
```

There looks like 18 categorical and 17 numerical variables in the data set.

The categorical columns are currently character datatype. Let us convert them to factor type. Also, let us set Attrition variable to 1 and 0 instead of “Yes” and “No”.

```
cat_cols= c("BusinessTravel", "Department", "Education", "EducationField", "EnvironmentSatisfaction", "
data$Attrition = ifelse(data$Attrition=="Yes",1,0)
data[cat_cols] = lapply(data[cat_cols], factor)
```

Checking summary of data

```
summary(data)
```

```
##       Age        Attrition           BusinessTravel   DailyRate     
##  Min.   :18.00   0:1233    Non-Travel       : 150   Min.   : 102.0  
##  1st Qu.:30.00   1: 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   1:170    
```

```
##  Research & Development:961   1st Qu.: 2.000   2:282
##  Sales                 :446   Median : 7.000   3:572
##                               Mean   : 9.193   4:398
##                               3rd Qu.:14.000   5: 48
##                               Max.   :29.000
##
##           EducationField EmployeeCount EmployeeNumber   EnvironmentSatisfaction
##  Human Resources : 27    Min.   :1     Min.   :   1.0   1:284
##  Life Sciences   :606    1st Qu.:1     1st Qu.: 491.2   2:287
##  Marketing       :159    Median :1     Median :1020.5   3:453
##  Medical         :464    Mean   :1     Mean   :1024.9   4:446
##  Other           : 82    3rd Qu.:1     3rd Qu.:1555.8
##  Technical Degree:132    Max.   :1     Max.   :2068.0
##
##     Gender      HourlyRate     JobInvolvement JobLevel
##  Female:588   Min.   : 30.00   1: 83          1:543
##  Male  :882   1st Qu.: 48.00   2:375          2:534
##               Median : 66.00   3:868          3:218
##               Mean   : 65.89   4:144          4:106
##               3rd Qu.: 83.75                  5: 69
##               Max.   :100.00
##
##                       JobRole    JobSatisfaction  MaritalStatus MonthlyIncome
##  Sales Executive          :326   1:289           Divorced:327   Min.   : 1009
##  Research Scientist       :292   2:280           Married :673   1st Qu.: 2911
##  Laboratory Technician    :259   3:442           Single  :470   Median : 4919
##  Manufacturing Director   :145   4:459                          Mean   : 6503
##  Healthcare Representative:131                                  3rd Qu.: 8379
##  Manager                  :102                                  Max.   :19999
##  (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
##  3:1244            1:276                    Min.   :80    0:631
##  4: 226            2:303                    1st Qu.:80    1:596
##                    3:459                    Median :80    2:158
##                    4:432                    Mean   :80    3: 85
##                                             3rd Qu.:80
##                                             Max.   :80
##
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
##  Min.   : 0.00     Min.   :0.000         1: 80           Min.   : 0.000
##  1st Qu.: 6.00     1st Qu.:2.000         2:344           1st Qu.: 3.000
##  Median :10.00     Median :3.000         3:893           Median : 5.000
##  Mean   :11.28     Mean   :2.799         4:153           Mean   : 7.008
##  3rd Qu.:15.00     3rd Qu.:3.000                         3rd Qu.: 9.000
##  Max.   :40.00     Max.   :6.000                         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      
## 
```

Employee Count is equal to 1 for all observation which can not generate useful value for this sample data.

Over 18 is equal to 'Y', which means employee is not less than 18 years old.

Similarly,Standard Hours is equal to 80 for all observations and hence is not useful for classification.

Employee Number is simply an ID associated with each employee and is also not useful for classification. So let us disregard these 4 variables from the further analyses.

```
data = data[-c(9,10,22,27)]
```

Let us check for NA and duplicate values in the dataset.

```
apply(is.na(data), 2, sum)
```

```
##                      Age                Attrition           BusinessTravel 
##                        0                        0                        0 
##                DailyRate               Department         DistanceFromHome 
##                        0                        0                        0 
##                Education           EducationField  EnvironmentSatisfaction 
##                        0                        0                        0 
##                   Gender               HourlyRate           JobInvolvement 
##                        0                        0                        0 
##                 JobLevel                  JobRole          JobSatisfaction 
##                        0                        0                        0 
##            MaritalStatus            MonthlyIncome              MonthlyRate 
##                        0                        0                        0 
##       NumCompaniesWorked                 OverTime        PercentSalaryHike 
##                        0                        0                        0 
##        PerformanceRating RelationshipSatisfaction         StockOptionLevel 
##                        0                        0                        0 
##        TotalWorkingYears    TrainingTimesLastYear          WorkLifeBalance 
##                        0                        0                        0 
##           YearsAtCompany       YearsInCurrentRole  YearsSinceLastPromotion 
##                        0                        0                        0 
##     YearsWithCurrManager 
##                        0
```

```
sum(is.na(duplicated(data)))
```

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

Thankfully, the data has no NA and duplicate values.

In this analysis we would answering few research questions related to Employee Attrition.

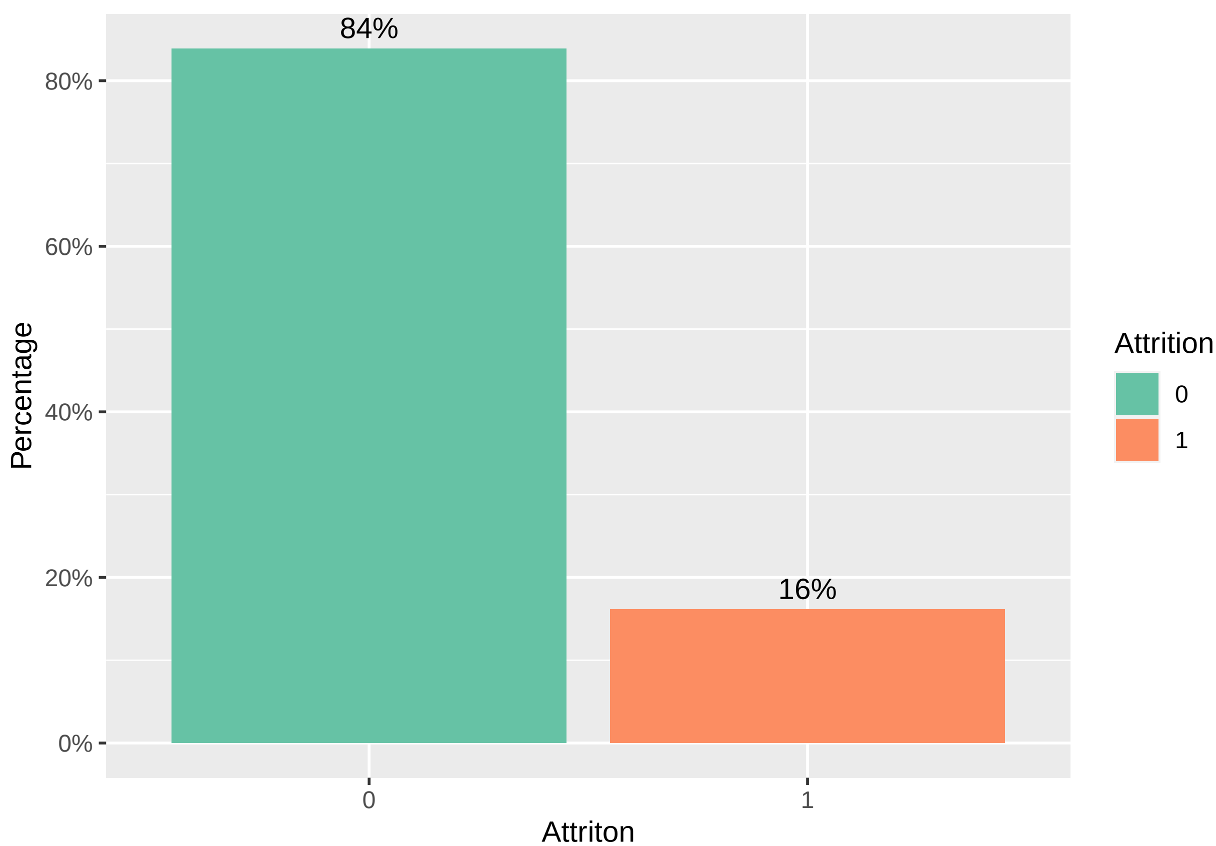
They are mentioned later on in this document.

Exploratory Data Analysis for this dataset to provide a initial intution on the dataset.

Let us first check the proportion of the response variable that is Attrition.

```
# Plotting the count of the attribution attribute

ggplot(data, aes(Attrition)) +
  geom_bar(position = "dodge", aes(y=(..count..)/sum(..count..), fill=Attrition)) +
  scale_y_continuous(labels=scales::percent) +
  ylab("Percentage") +
  xlab("Attriton") +
  geom_text(aes(label = scales::percent((..count..)/sum(..count..)), y=(..count..)/sum(..count..)), sta
  scale_fill_brewer(palette="Set2")
```

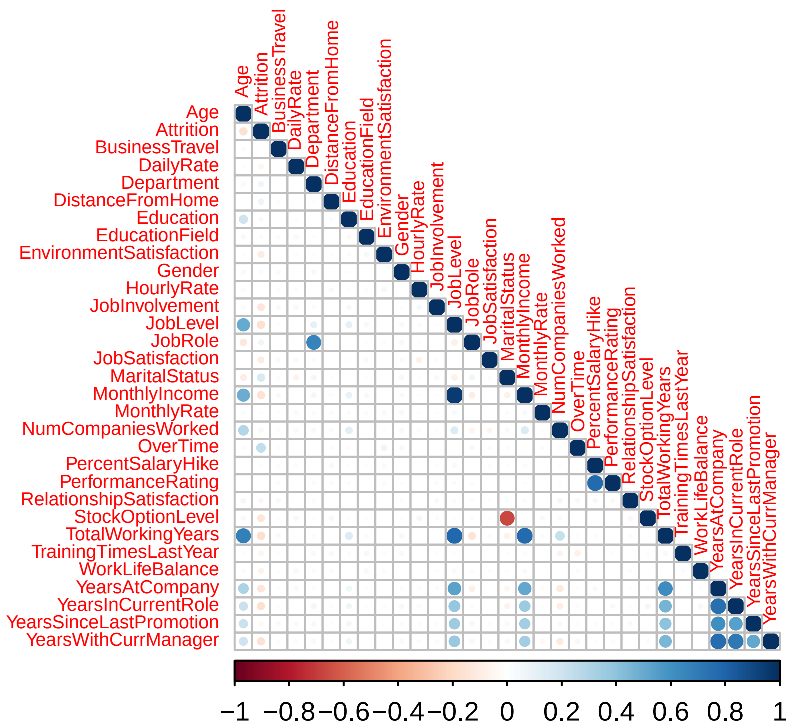


Upon checking for the proportion of values in our response variable, i.e. Attrition we realized that the data is severely imbalanced. This means that even without training any model, if we predict all the responses as '0', still we will get an accuracy of 83.88% (1233*100/1470). We consider this as our 'base model' for future reference. However, this model would have a poor performance if the test set has majorly '1' as the response variable.

We are oversampling this data set while developing the model.

Let us know first check correlation among different variables.

```
data_cor=data
for(i in 1:ncol(data_cor)){
data_cor[,i]<-as.integer(data_cor[,i])
}
corrplot(cor(data_cor), type = 'lower', tl.cex = 0.6)
```



Looking at the above plots we can conclude the following : -

1. Age variable is correlated with TotalWorkingYears
2. TotalWorkingYears correlated with MonthlyIncome
3. YearsWithCurrManager also correlated with YearsAtCompany
4. YearsWithCurrManger correlated with YearsInCurrentRole
5. YearsInCurrentRole correlated with YearsAtCompany
6. TotalWorkingYears correlated with JobLevel

We would definitely need some of the above predictors while making predictions.

Let us now proceed to have a quick glance at the other predictors of the data set.

Understanding Department Predictor.

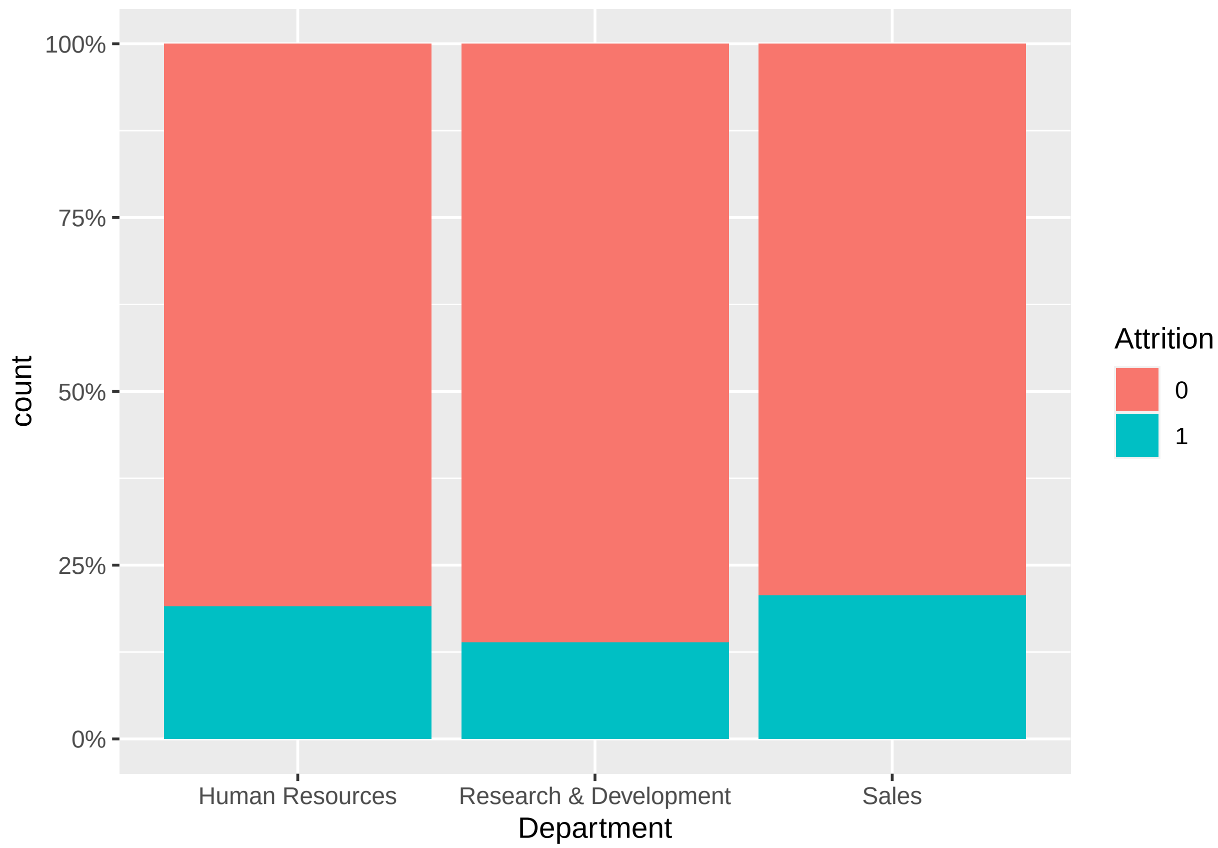
```r
summary(data$Department)
```

```
##        Human Resources Research & Development                  Sales
##                     63                    961                    446
```

```r
table(data$Department, data$Attrition)
```

```
##
##                            0   1
##   Human Resources         51  12
##   Research & Development 828 133
##   Sales                  354  92
```

```r
dept_plot = ggplot(data,aes(Department,fill=Attrition))+geom_bar(position="fill")+scale_y_continuous(la
dept_plot
```



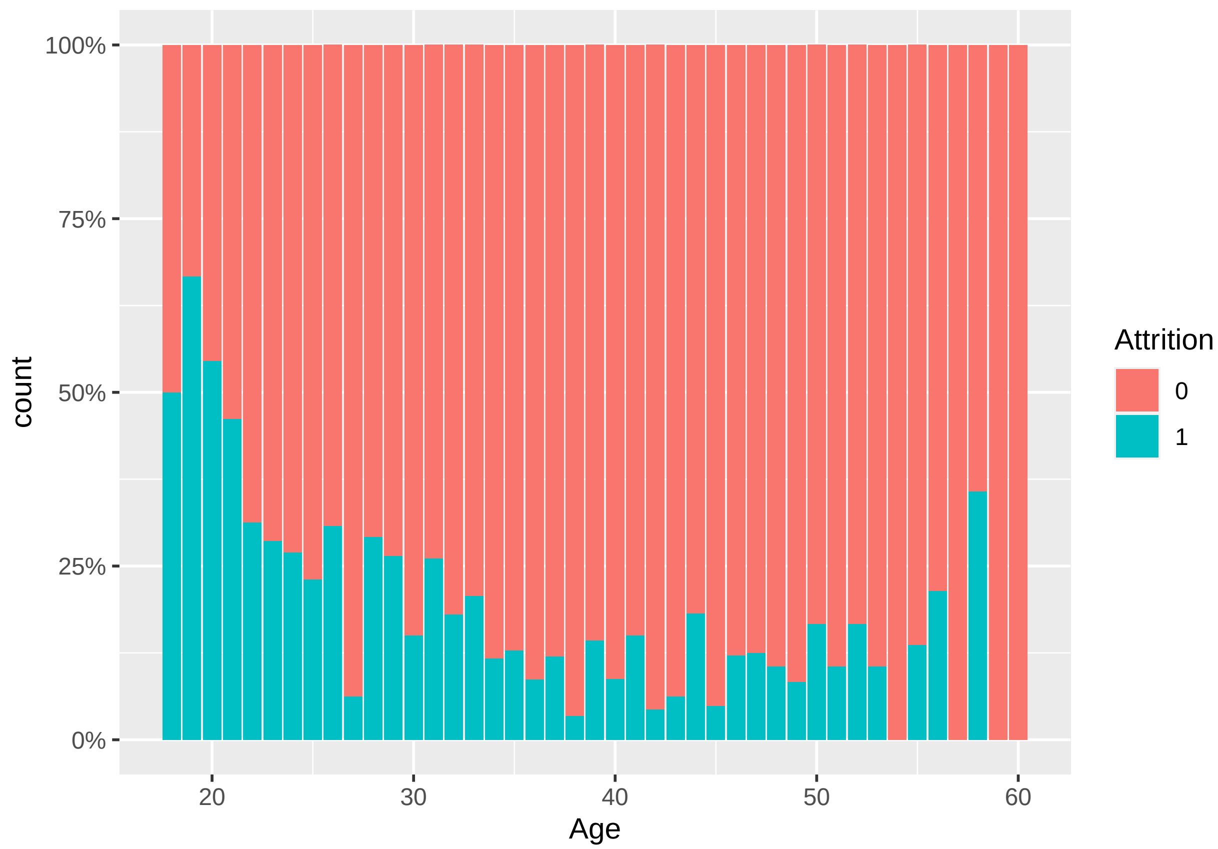
The proportion of Attrition is similar in the Sales and Human resources department. However, in case of R&D, there is a comparatively less proportion of Attrition. The possible reason for this might be because of the fact that getting accustomed to a company's R&D department can be very tedious and hence people in this department do not prefer switching their jobs.

Understanding Age Predictor.

```
summary(data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   30.00   36.00   36.92   43.00   60.00
```

```
age_plot = ggplot(data,aes(Age,fill=Attrition))+geom_bar(position="fill")+scale_y_continuous(labels = pe
age_plot
```



As we can clearly see in the graph, young employees tend to switch their jobs. However, people who show a commitment to the company by working for several years find stability within the organization and hence do not change their jobs frequently.

Besides the variables mentioned so far, we tried to look for any other factors that had significant effect on the Attrition variable.

Understanding Job Role Predictor

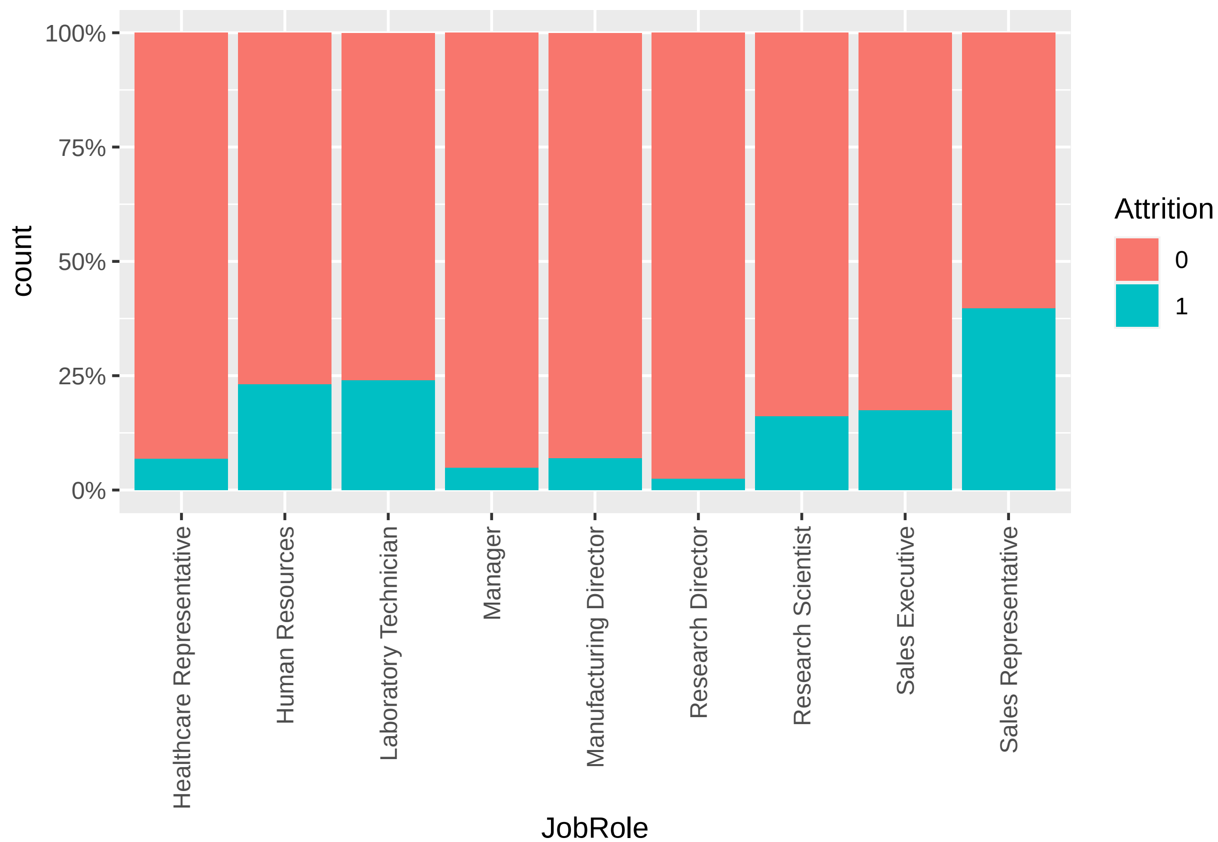
```
summary(data$JobRole)
```

```
## Healthcare Representative           Human Resources     Laboratory Technician
##                       131                        52                       259
##                   Manager    Manufacturing Director         Research Director
##                       102                       145                        80
##        Research Scientist           Sales Executive      Sales Representative
##                       292                       326                        83
```

```
table(data$JobRole, data$Attrition)
```

```
##
##                               0   1
##   Healthcare Representative 122   9
##   Human Resources            40  12
##   Laboratory Technician     197  62
##   Manager                    97   5
##   Manufacturing Director    135  10
##   Research Director          78   2
##   Research Scientist        245  47
##   Sales Executive           269  57
##   Sales Representative       50  33
```

```
jrole_plot = ggplot(data,aes(JobRole,fill=Attrition))+geom_bar(position="fill")+scale_y_continuous(label
jrole_plot
```



We found out that for the JobRole variable, SalesRepresentative had the maximum proportion of Attrition.

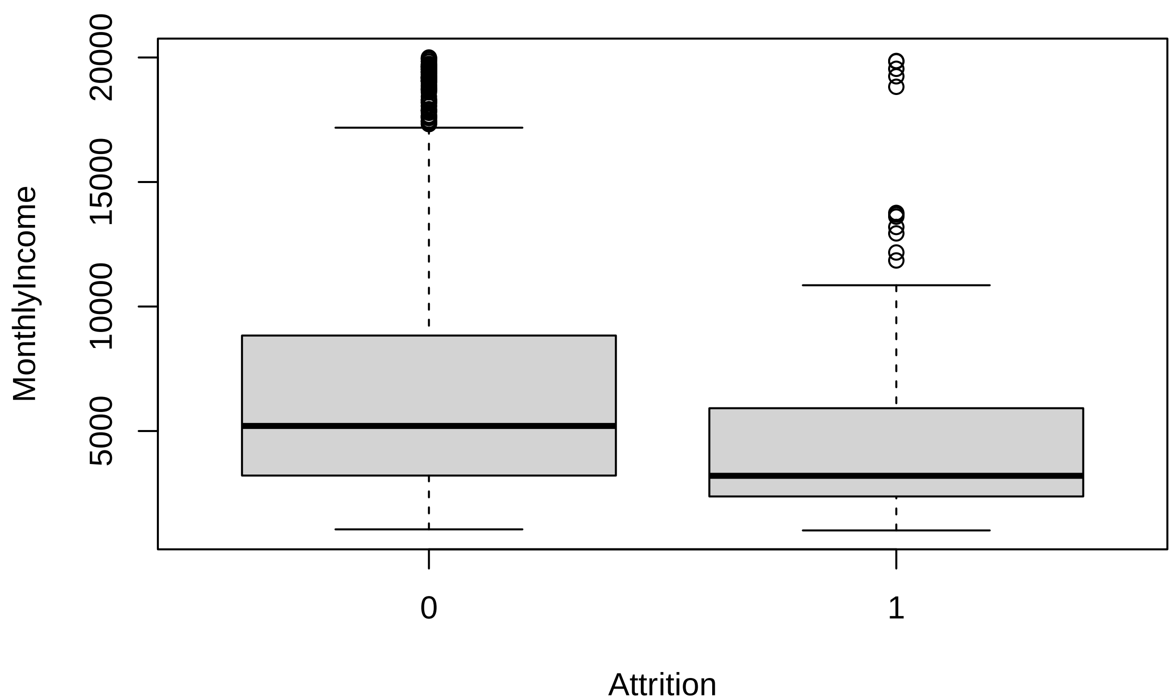
The possible explanation for this can be that Sales jobs are generally incentive-based and have less ties with the company. Hence, it might be easier for people to switch jobs for better incentives.

Understanding Monthly Income

```
summary(data$MonthlyIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1009    2911    4919    6503    8379   19999
```

```
boxplot(MonthlyIncome~Attrition, data = data)
```



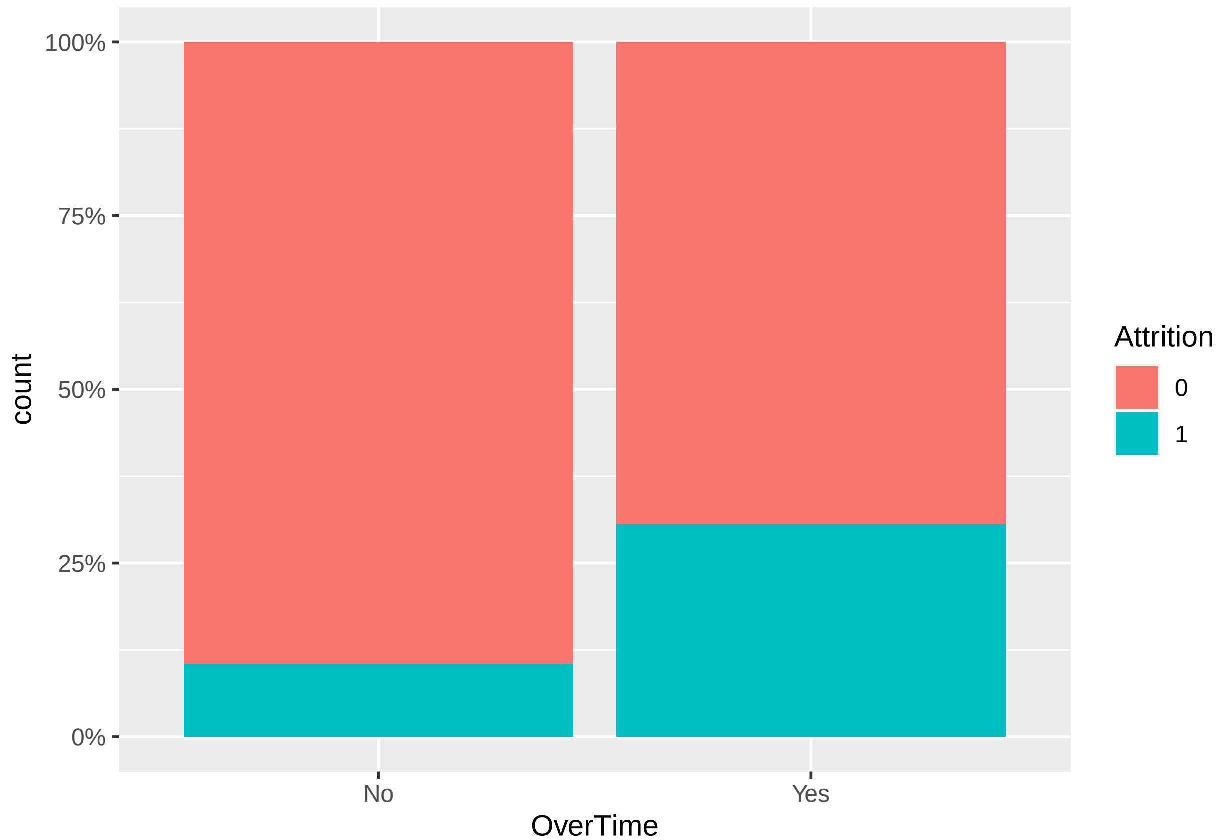
Next, we found out that MonthlyIncome also had considerable effect on Attrition. Majority of Employees in the Attrition group have a monthly income of less than 5000$.

Understanding Overtime Predictor

```
summary(data$OverTime)
```

```
##   No  Yes
## 1054  416
```

```
overtime_plot = ggplot(data,aes(OverTime,fill=Attrition))+geom_bar(position="fill")+scale_y_continuous(l
overtime_plot
```



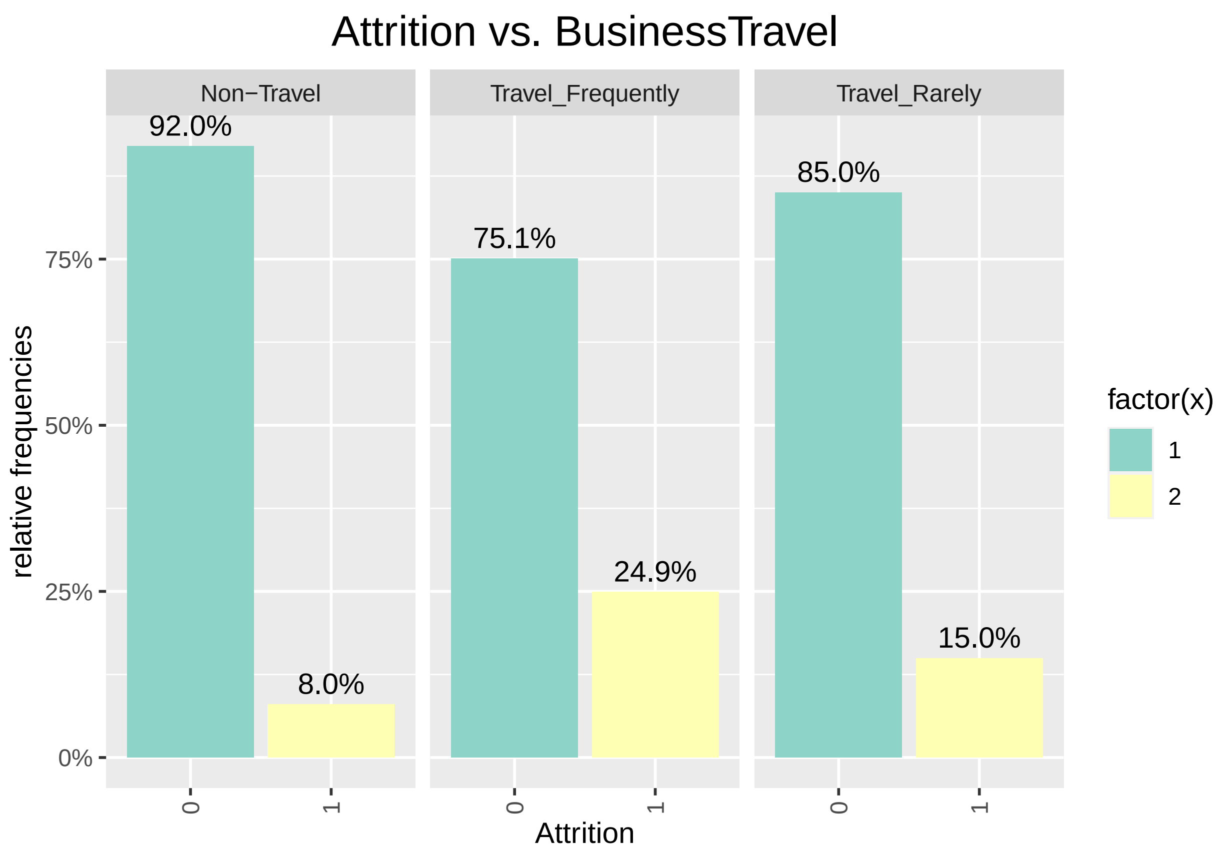
We also observed that, the proportion of Attrition among the people working overtime is more than that of people who do not work Overtime.

Understanding BusinessTravel Predictor

```
summary(data$BusinessTravel)
```

```
##        Non-Travel Travel_Frequently     Travel_Rarely
##               150               277              1043
```

```
ggplot(data,aes(x=Attrition,group=BusinessTravel))+
  geom_bar(aes(y=..prop..,fill=factor(..x..)),stat="count")+
  facet_grid(~BusinessTravel)+
  labs(x="Attrition",y="Percentage",title="Attrition vs. BusinessTravel")+
  theme(axis.text.x=element_text(angle=90,vjust=0.5),plot.title=element_text(size=16,hjust=0.5))+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
  scale_y_continuous(labels=scales::percent) +
  ylab("relative frequencies") +
  scale_fill_brewer(palette="Set3")
```



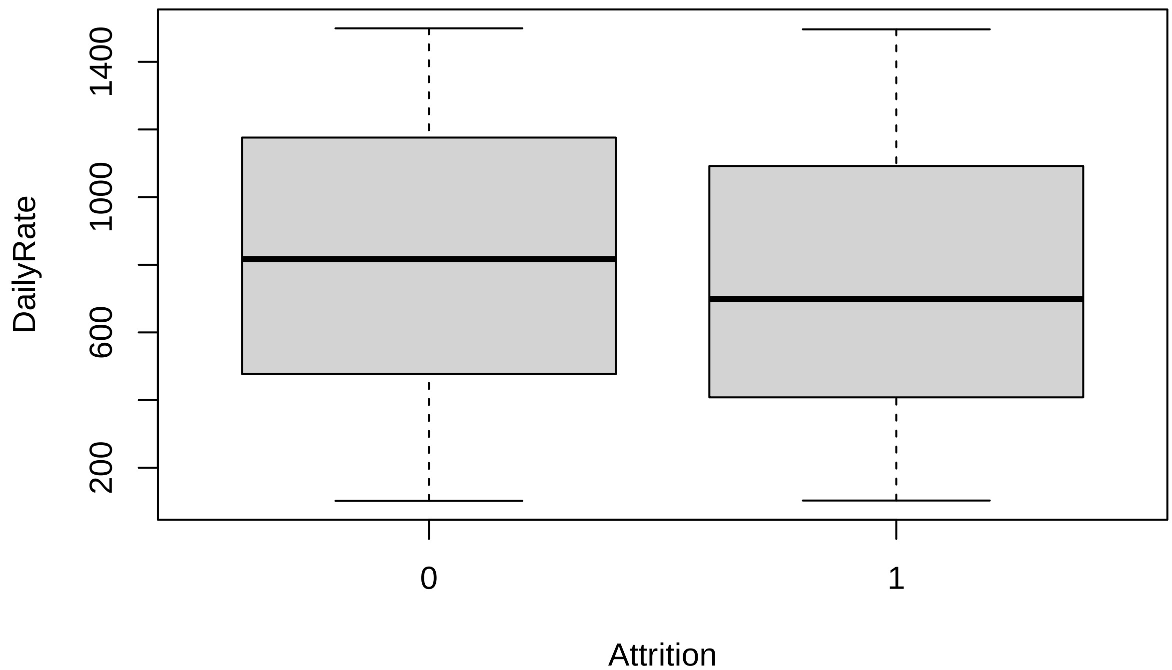
Here, we can see that the proportion of Attrition among the Frequent travelers is greater than that of those who do not travel frequently.

Understanding DailyRate Predictor

```
summary(data$DailyRate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   102.0   465.0   802.0   802.5  1157.0  1499.0
```

```
boxplot(DailyRate~Attrition, data = data)
```



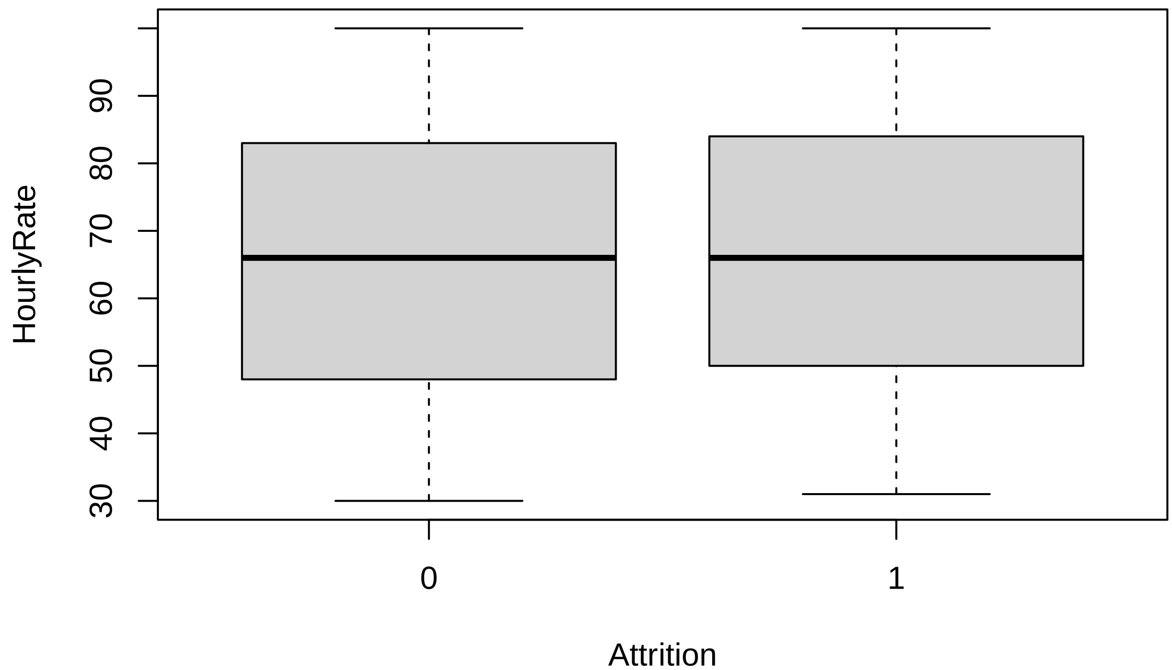
There is no significant effect of Daily Rate seen on Attrition.

Understanding Hourly Rate Predictor

```
summary(data$HourlyRate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   30.00   48.00   66.00   65.89   83.75  100.00
```

```
boxplot(HourlyRate~Attrition, data = data)
```



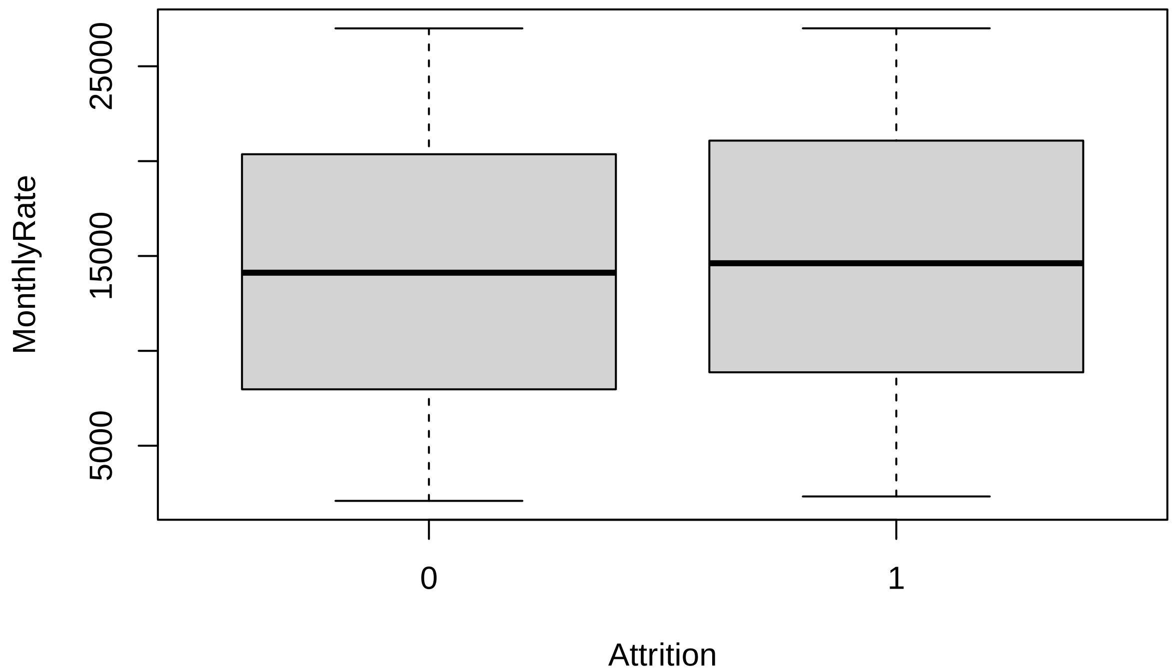
Not much relation is observed between HourlyRate and the Attrition.

Understanding Monthly Rate Predictor

```
summary(data$MonthlyRate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2094    8047   14236   14313   20462   26999
```

```
boxplot(MonthlyRate~Attrition, data = data)
```



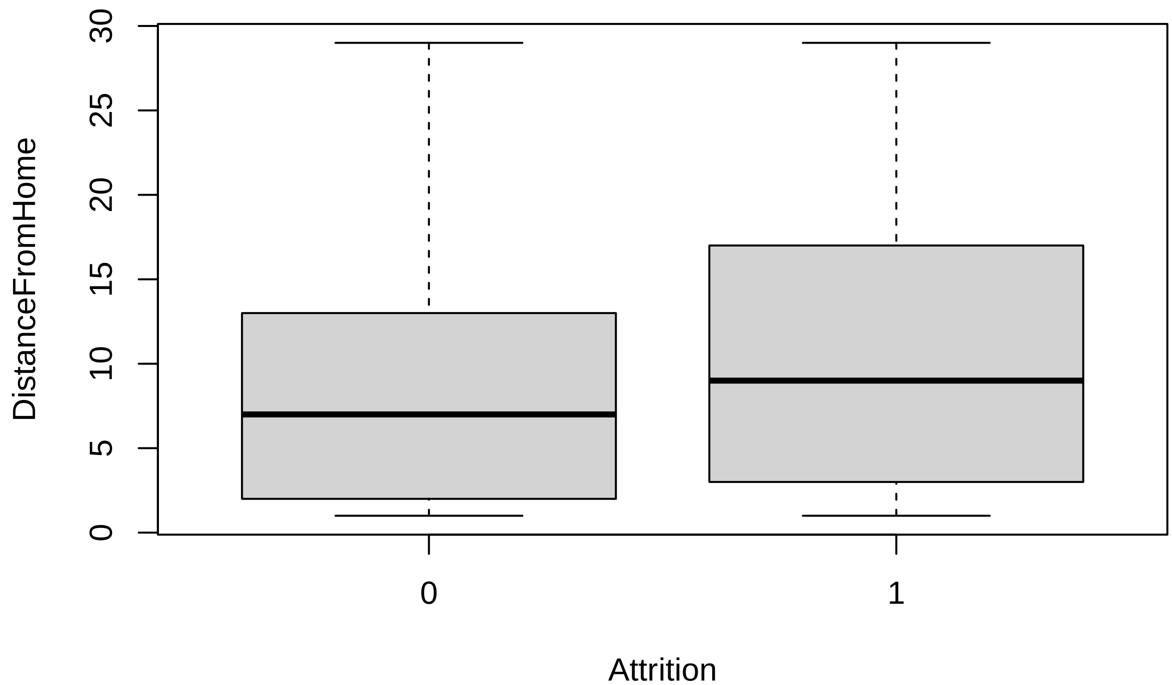
Not much relation is observed between Attrition and MonthlyRate.

Understanding DistanceFromHome - Distance from home in kms¶

```
summary(data$DistanceFromHome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   7.000   9.193  14.000  29.000
```

```
boxplot(DistanceFromHome~Attrition, data = data)
```



There is no significant effect of Distance From Home seen on Attrition.

Understanding Education - Education Level 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'

```
summary(data$Education)
```

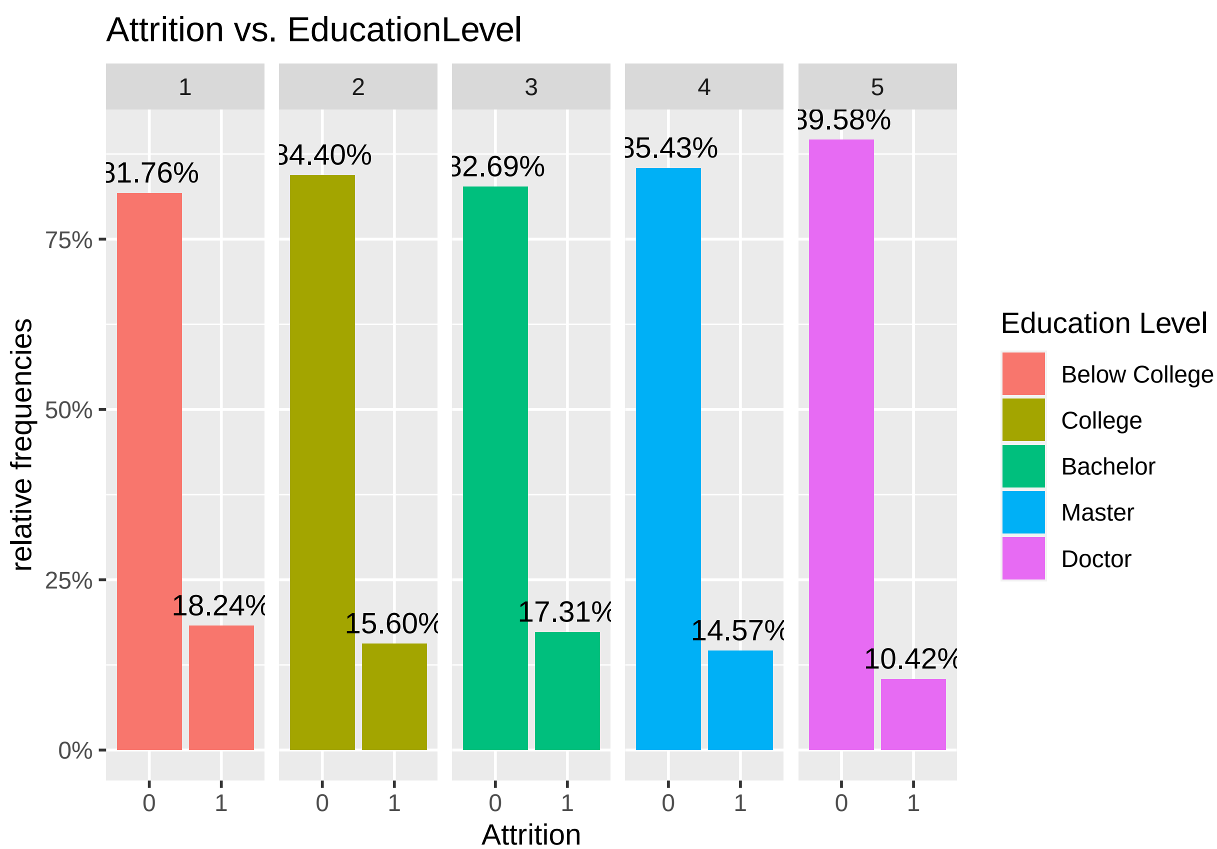
```
##   1   2   3   4   5
## 170 282 572 398  48
```

```
table(data$Education, data$Attrition)
```

```
##
##       0   1
##   1 139  31
##   2 238  44
##   3 473  99
##   4 340  58
##   5  43   5
```

```
ggplot(data,aes(x=Attrition,group=Education))+
  geom_bar(aes(y=..prop..,fill=factor(..group..)),stat="count")+
```

```
  facet_grid(~Education)+
  labs(x="Attrition",y="Percentage",title="Attrition vs. EducationLevel")+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
  scale_y_continuous(labels=scales::percent) +
  ylab("relative frequencies") +
  scale_fill_discrete(name="Education Level", label=c("Below College", "College", "Bachelor", "Master",
```



There is no significant effect of Education Level predictor seen on Attrition.

Understanding Education Field - Field of education

```
summary(data$EducationField)
```

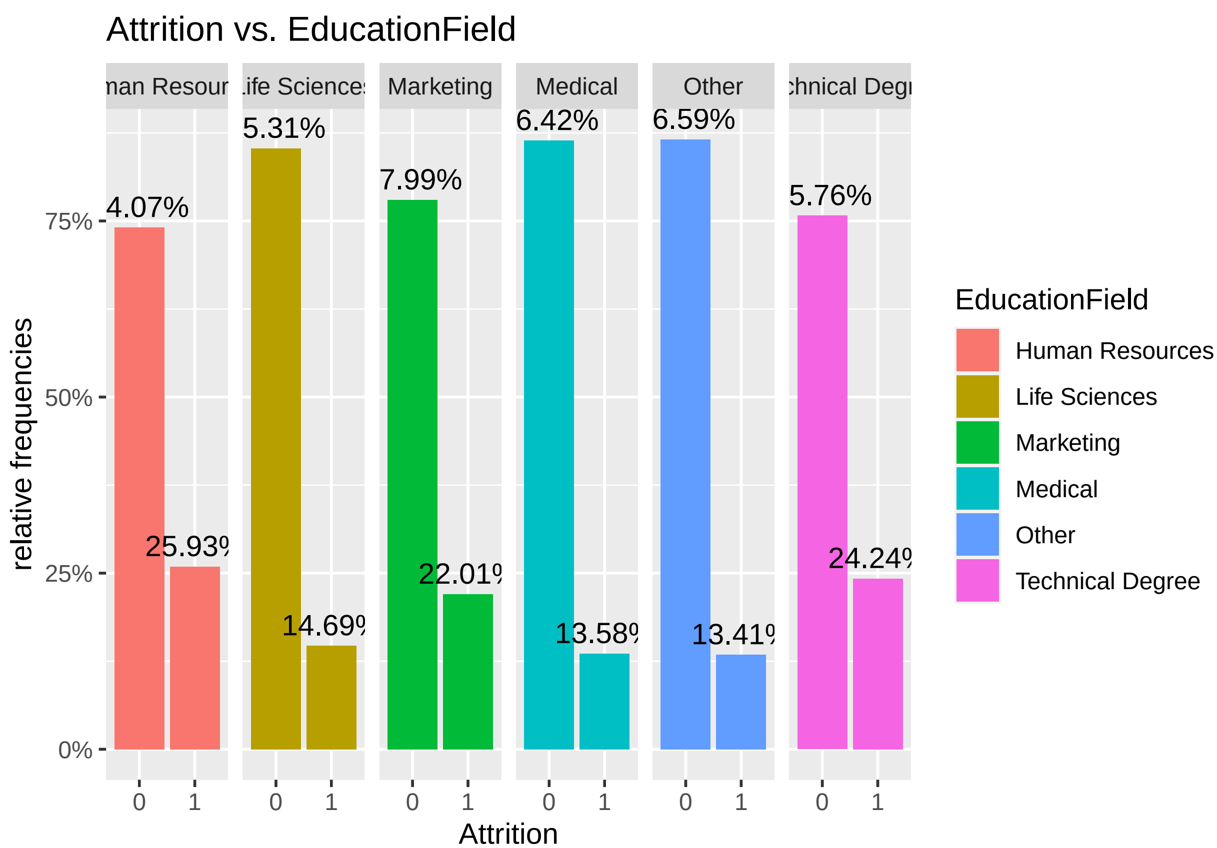
```
##  Human Resources     Life Sciences         Marketing           Medical
##               27               606               159               464
##            Other Technical Degree
##               82               132
```

```
table(data$EducationField, data$Attrition)
```

```
##
##                        0    1
```

```
##    Human Resources   20   7
##    Life Sciences    517  89
##    Marketing        124  35
##    Medical          401  63
##    Other             71  11
##    Technical Degree 100  32
```

```
ggplot(data,aes(x=Attrition,group=EducationField))+
  geom_bar(aes(y=..prop..,fill=factor(..group..)),stat="count")+
  facet_grid(~EducationField)+
  labs(x="Attrition",y="Percentage",title="Attrition vs. EducationField")+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
  scale_y_continuous(labels=scales::percent) +
  ylab("relative frequencies") +
  scale_fill_discrete(name="EducationField ", label=c("Human Resources", "Life Sciences", "Marketing",
```



It is observed that HR, Marketing and Technical degrees have higher Attrition proportions when compared to other Educational Fields.

Understanding Environment Satisfaction predictor

```
summary(data$EnvironmentSatisfaction)
```

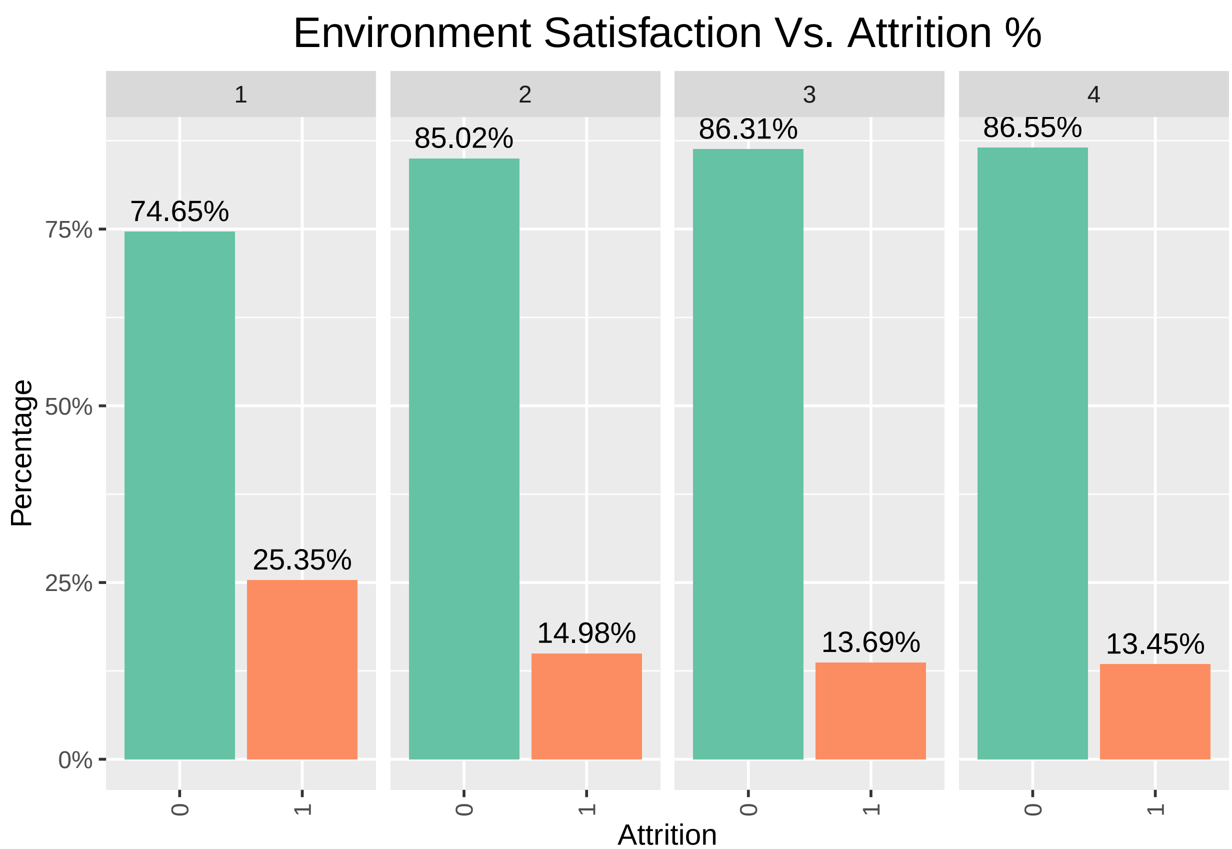
```
##    1    2    3    4
```

```
## 284 287 453 446
```

```
table(data$EnvironmentSatisfaction, data$Attrition)
```

```
##
##       0   1
##   1 212  72
##   2 244  43
##   3 391  62
##   4 386  60
```

```
ggplot(data,aes(x=Attrition,group=EnvironmentSatisfaction), ordered=T)+
  geom_bar(aes(y=..prop..,fill=factor(..x..)),stat="count")+
  facet_grid(~EnvironmentSatisfaction)+
  scale_y_continuous(labels=scales::percent) +
  theme(axis.text.x=element_text(angle=90,vjust=0.5),legend.position="none",plot.title=element_text(siz
  labs(x="Attrition",y="Percentage",title="Environment Satisfaction Vs. Attrition %")+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
  scale_fill_brewer(palette="Set2")
```



Employees with lower Environment Satisfaction tend to leave their jobs.

Understanding JobSatisfaction Predictor

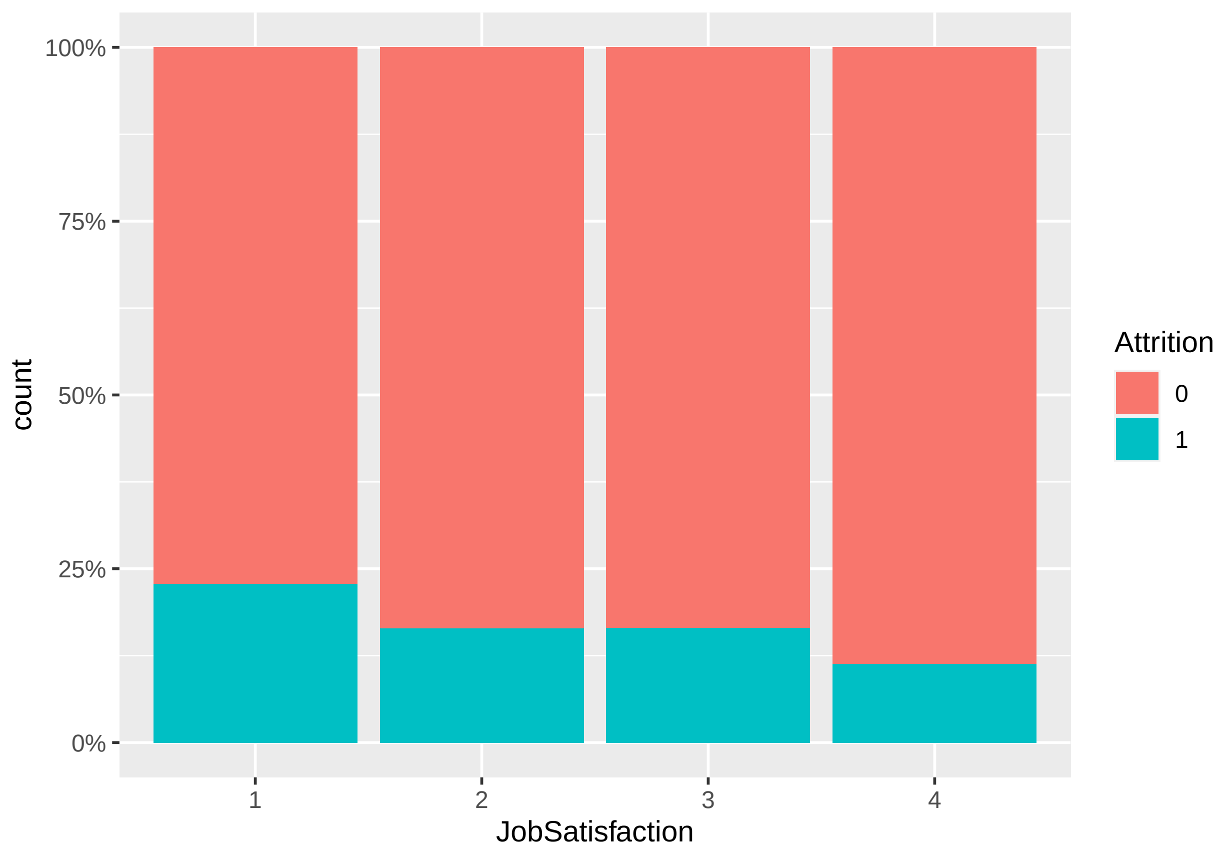
```
summary(data$JobSatisfaction)
```

```
##   1   2   3   4
## 289 280 442 459
```

```
table(data$JobSatisfaction, data$Attrition)
```

```
##
##       0   1
##   1 223  66
##   2 234  46
##   3 369  73
##   4 407  52
```

```
jsatisfaction_plot = ggplot(data,aes(JobSatisfaction,fill=Attrition))+geom_bar(position="fill")+scale_y_
jsatisfaction_plot
```



Similar to the Environment Satisfaction, lower Job Satisfaction make employees leave their jobs.

Understanding Gender Predictor

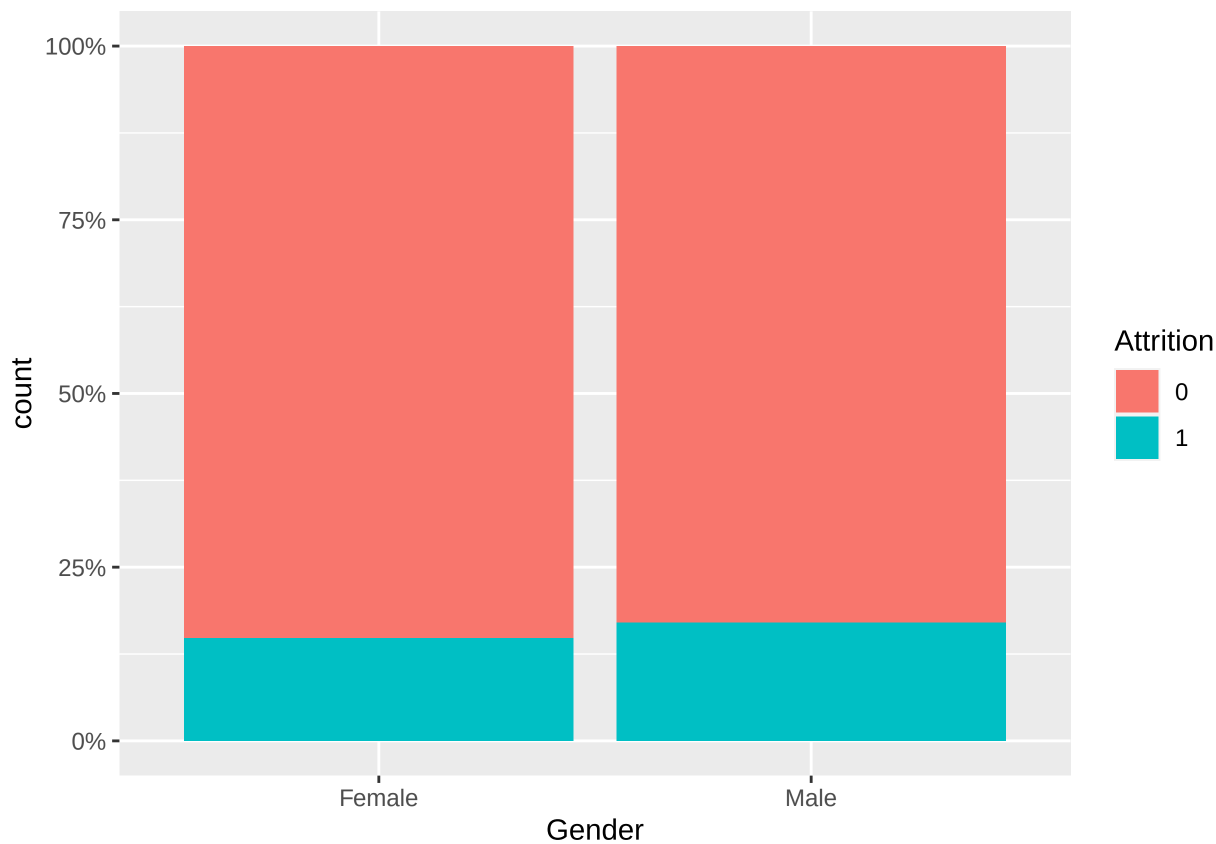
```
summary(data$Gender)
```

```
## Female   Male
##    588    882
```

```
table(data$Gender, data$Attrition)
```

```
##
##            0   1
##   Female 501  87
##   Male   732 150
```

```
gender_plot = ggplot(data,aes(Gender,fill=Attrition))+geom_bar(position="fill")+scale_y_continuous(label
gender_plot
```



There doesn't seem much relationship between Gender and Attrition.

Understanding JobInvolvement Predictor

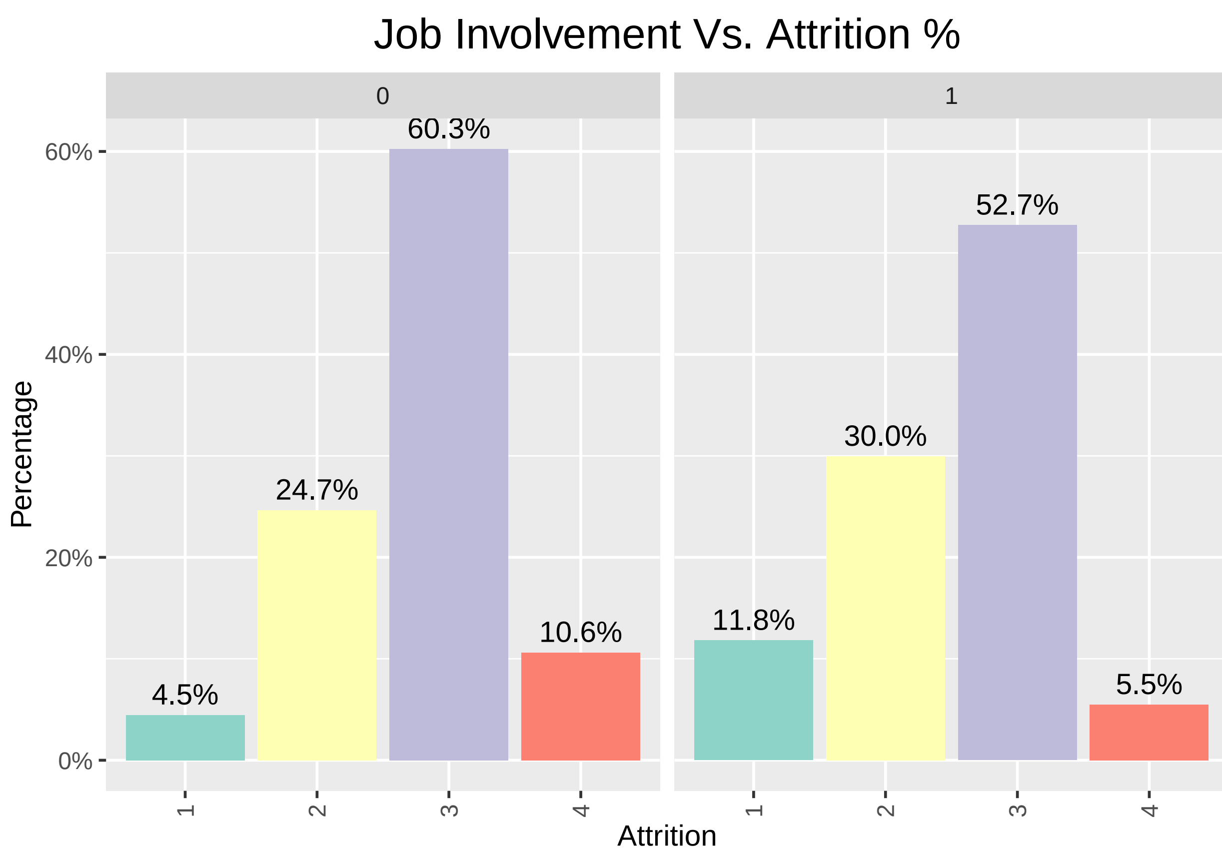
```r
summary(data$JobInvolvement)
```

```
##   1   2   3   4
##  83 375 868 144
```

```r
table(data$JobInvolvement)
```

```
##
##   1   2   3   4
##  83 375 868 144
```

```r
ggplot(data,aes(x=JobInvolvement,group=Attrition))+
  geom_bar(aes(y=..prop..,fill=factor(..x..)),stat="count")+
  facet_grid(~Attrition)+
  scale_y_continuous(labels=scales::percent) +
  theme(axis.text.x=element_text(angle=90,vjust=0.5),legend.position="none",plot.title=element_text(size
  labs(x="Attrition",y="Percentage",title="Job Involvement Vs. Attrition %")+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
  scale_fill_brewer(palette="Set3")
```



It is observed that as the JobInvolvement decreases, the proportion of Attrition also decreases. This suggests that employees with higher JobInvolvement tend to leave their jobs much easily.

Understanding Job Level Predictor

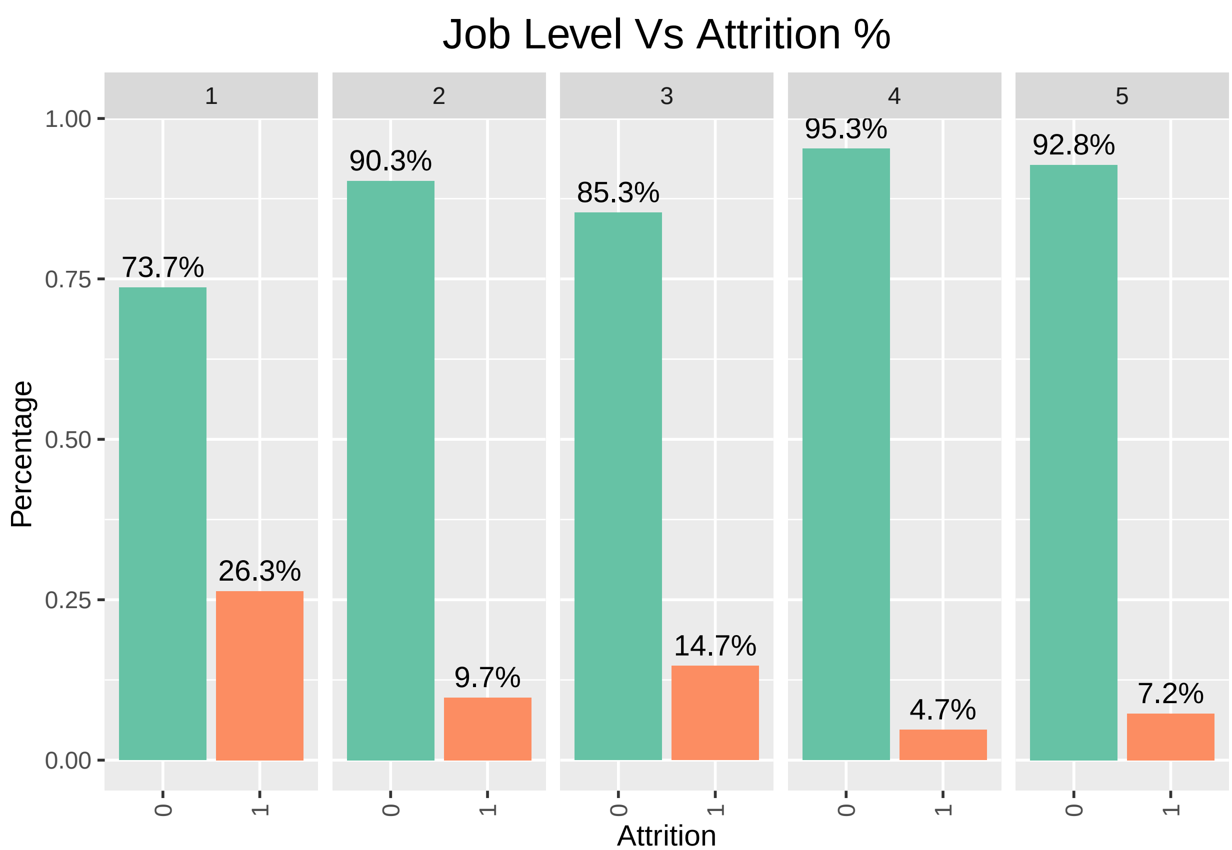
```
summary(data$JobLevel)
```

```
##   1   2   3   4   5
## 543 534 218 106  69
```

```
table(data$JobLevel, data$Attrition)
```

```
##
##       0   1
##   1 400 143
##   2 482  52
##   3 186  32
##   4 101   5
##   5  64   5
```

```
ggplot(data,aes(x=Attrition,group=JobLevel))+
  geom_bar(aes(y=..prop..,fill=factor(..x..)),stat="count")+
  facet_grid(~JobLevel)+
  theme(axis.text.x=element_text(angle=90,vjust=0.5),legend.position="none",plot.title=element_text(size
  labs(x="Attrition",y="Percentage",title="Job Level Vs Attrition %")+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
  scale_fill_brewer(palette="Set2")
```



Higher Job levels tend to have lower Attrition.

Understanding MaritalStatus Predictor.

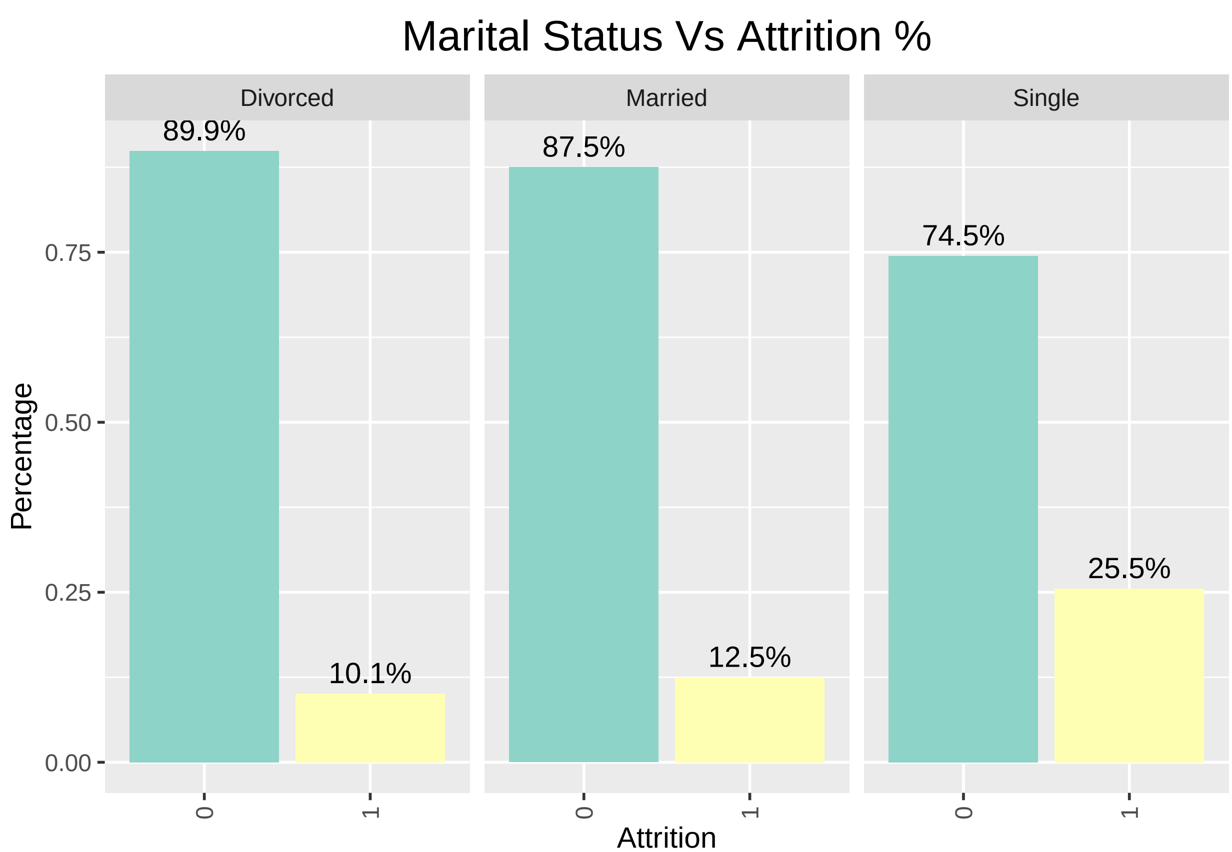
```
summary(data$MaritalStatus)
```

```
## Divorced  Married   Single
##      327      673      470
```

```
table(data$MaritalStatus, data$Attrition)
```

```
##
##              0   1
##   Divorced 294  33
##   Married  589  84
##   Single   350 120
```

```
ggplot(data,aes(x=Attrition,group=MaritalStatus))+
  geom_bar(aes(y=..prop..,fill=factor(..x..)),stat="count")+
  facet_grid(~MaritalStatus)+
  theme(axis.text.x=element_text(angle=90,vjust=0.5),legend.position="none",plot.title=element_text(siz
  labs(x="Attrition",y="Percentage",title="Marital Status Vs Attrition %")+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
  scale_fill_brewer(palette="Set3")
```



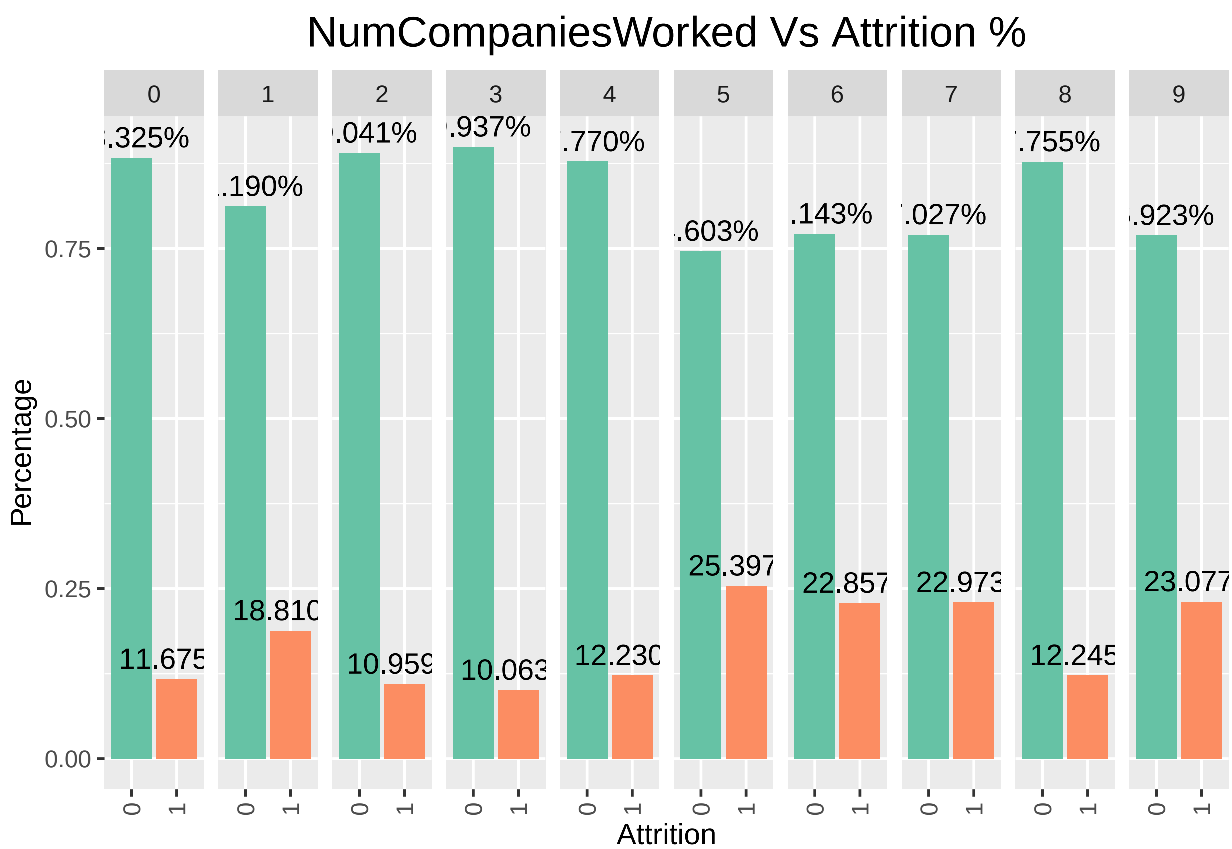
It is observed that Single employees tend to have higher proportion of Attrition when compared to Married or Divorced employees. This might be because they are willing to take risks.

Understanding NumCompaniesWorked Predictor.

```
summary(data$NumCompaniesWorked)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.693   4.000   9.000
```

```
ggplot(data,aes(x=Attrition,group=NumCompaniesWorked))+
  geom_bar(aes(y=..prop..,fill=factor(..x..)),stat="count")+
  facet_grid(~NumCompaniesWorked)+
  theme(axis.text.x=element_text(angle=90,vjust=0.5),legend.position="none",plot.title=element_text(siz
  labs(x="Attrition",y="Percentage",title="NumCompaniesWorked Vs Attrition %")+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
  scale_fill_brewer(palette="Set2")
```



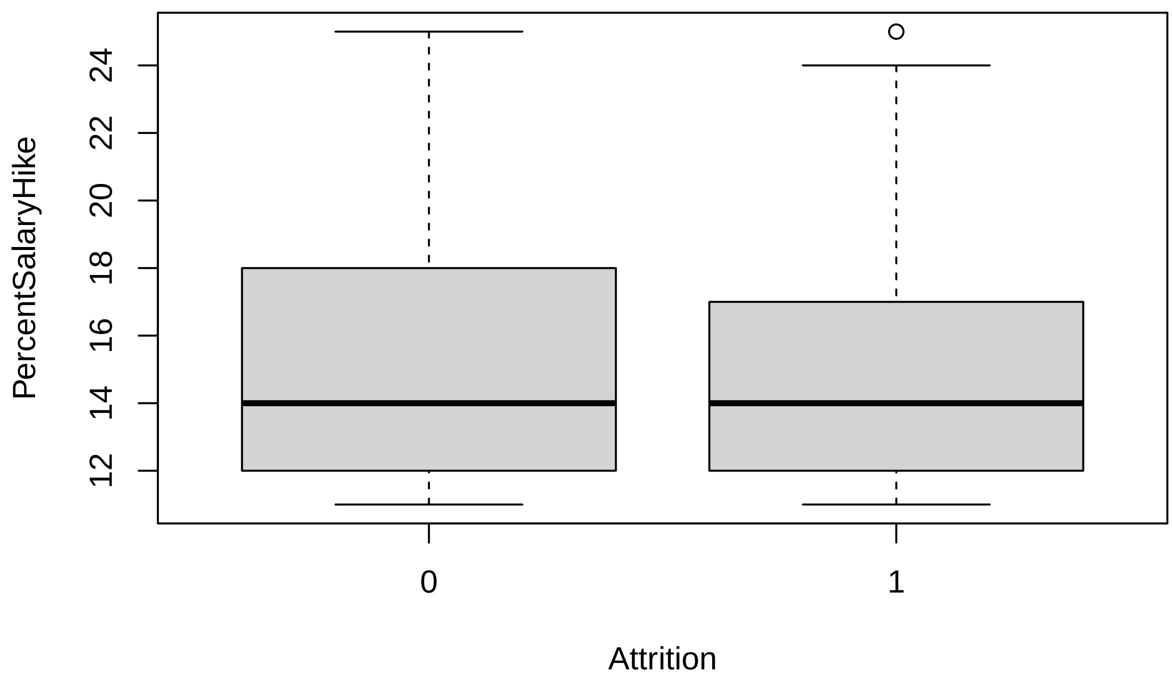
Attrition is higher when an employee has worked with 5 or more companies.

Understanding PercentSalaryHike - Percent salary hike for last year Predictor

```r
summary(data$PercentSalaryHike)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   12.00   14.00   15.21   18.00   25.00
```

```r
boxplot(PercentSalaryHike~Attrition, data = data)
```



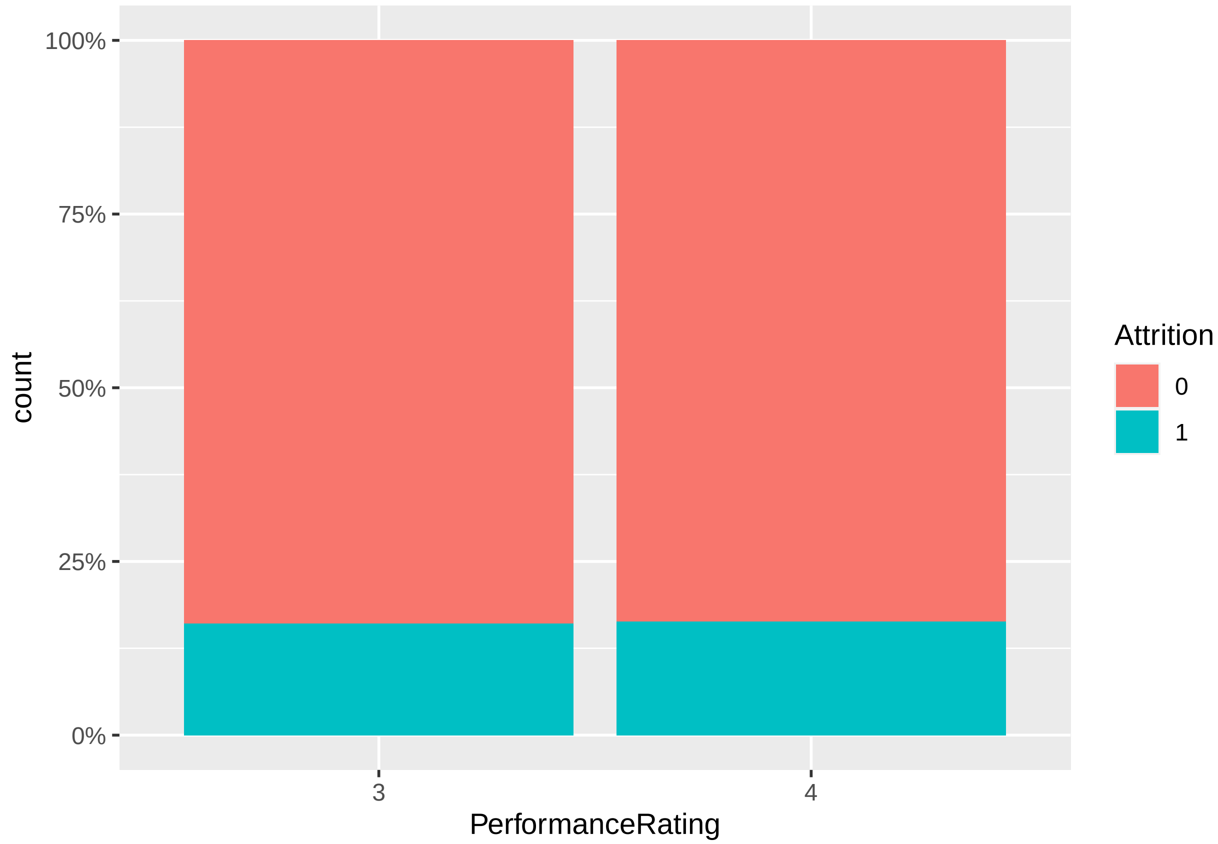
Not much relation is observed between PercentSalaryHike and Attrition.

Understanding PerformanceRating Predictor

```r
summary(data$PerformanceRating)
```

```
##    3    4
## 1244  226
```

```r
prating_plot = ggplot(data,aes(PerformanceRating,fill=Attrition))+geom_bar(position="fill")+scale_y_con
prating_plot
```



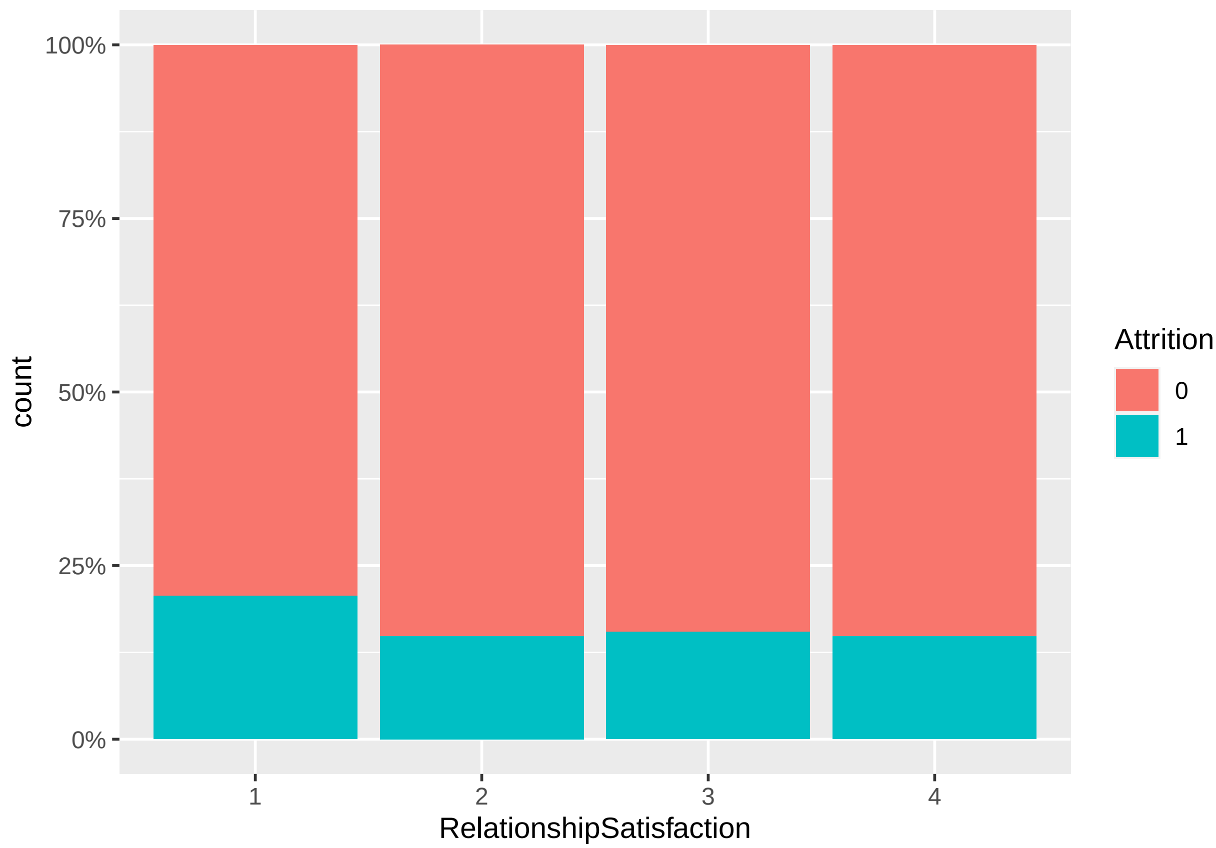
Performance rating also does not look like have any strong effect on Attrition.

Understanding Relationship Satisfaction Predictor

```
summary(data$RelationshipSatisfaction)
```

```
##   1   2   3   4
## 276 303 459 432
```

```
relsatisfaction_plot = ggplot(data,aes(RelationshipSatisfaction,fill=Attrition))+geom_bar(position="fil
relsatisfaction_plot
```



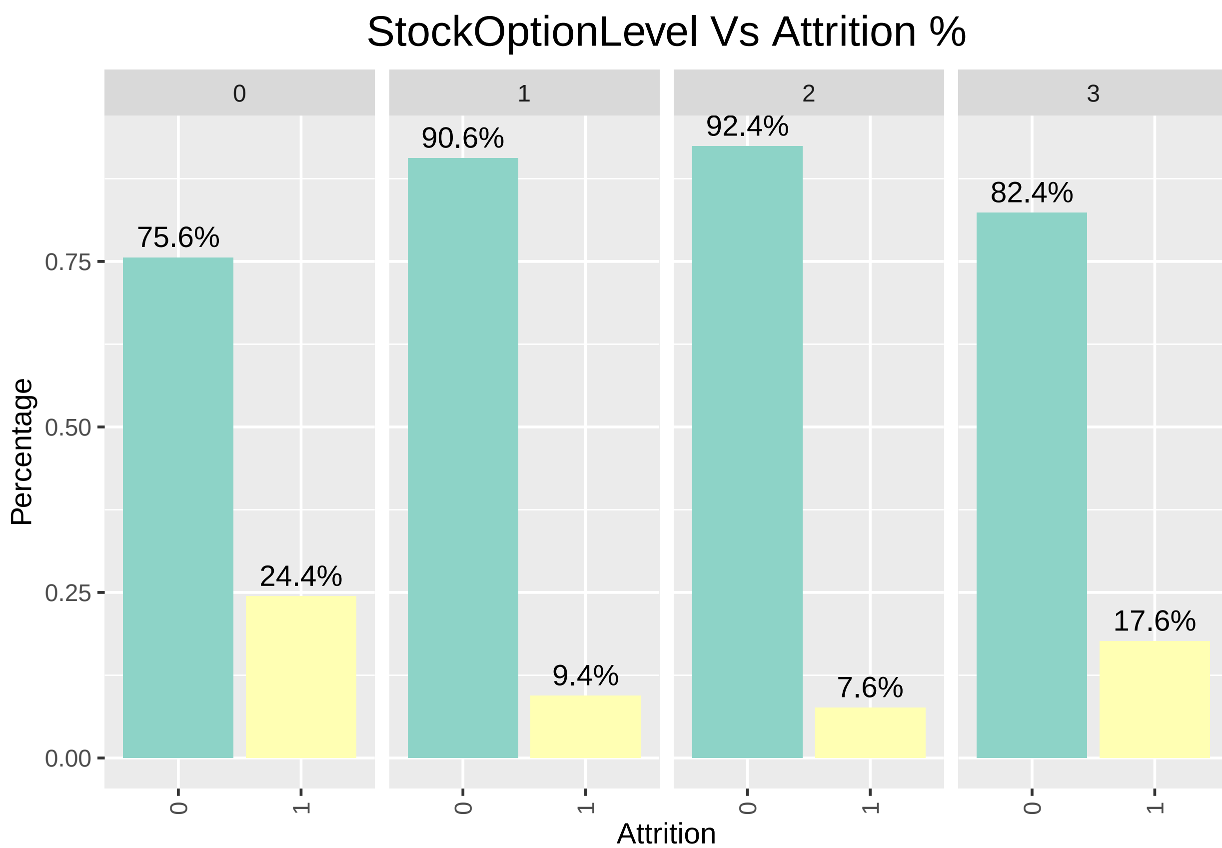
Not a great effect but there seems some connection of attrition with lower relationship satisfaction.

Understanding StockOptionLevel Predictor

```
summary(data$StockOptionLevel)
```

```
##   0   1   2   3
## 631 596 158  85
```

```
ggplot(data,aes(x=Attrition,group=StockOptionLevel))+
  geom_bar(aes(y=..prop..,fill=factor(..x..)),stat="count")+
  facet_grid(~StockOptionLevel)+
  theme(axis.text.x=element_text(angle=90,vjust=0.5),legend.position="none",plot.title=element_text(siz
  labs(x="Attrition",y="Percentage",title="StockOptionLevel Vs Attrition %")+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
  scale_fill_brewer(palette="Set3")
```



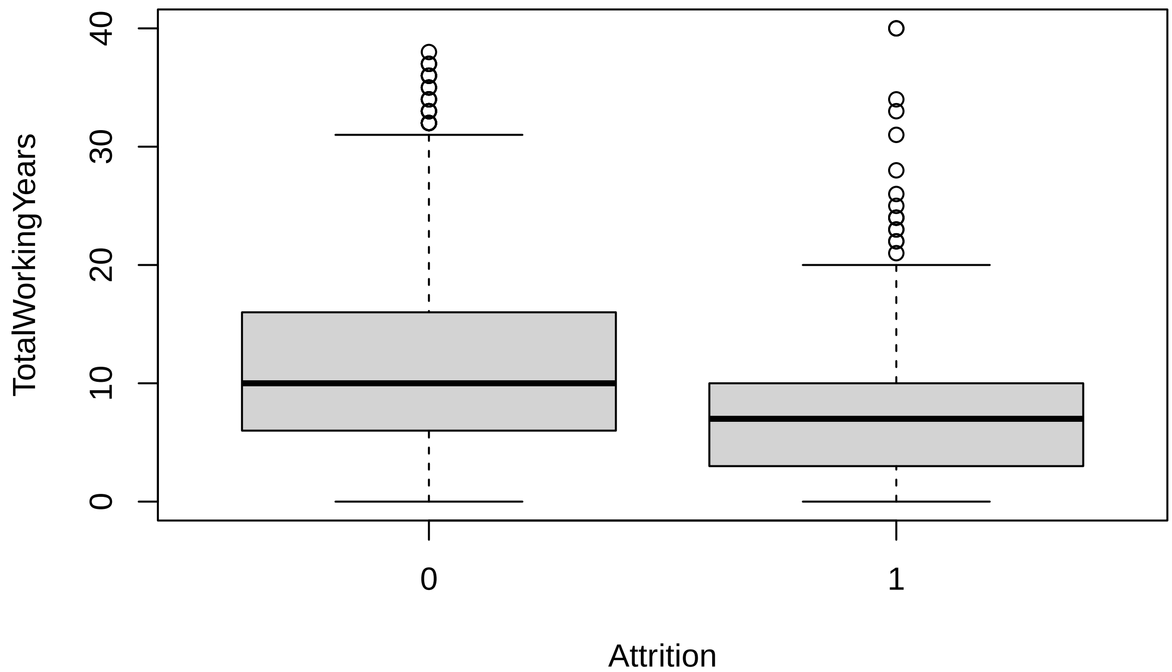
Higher attrition is observed for 0 stockOptionLevel.

Understanding TotalWorkingYears - Total number of years the employee has worked so far

```
summary(data$TotalWorkingYears)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    6.00   10.00   11.28   15.00   40.00
```

```
boxplot(TotalWorkingYears~Attrition, data = data)
```



Rate of Attrition is less in case there are long working years (preferably 10 or more).

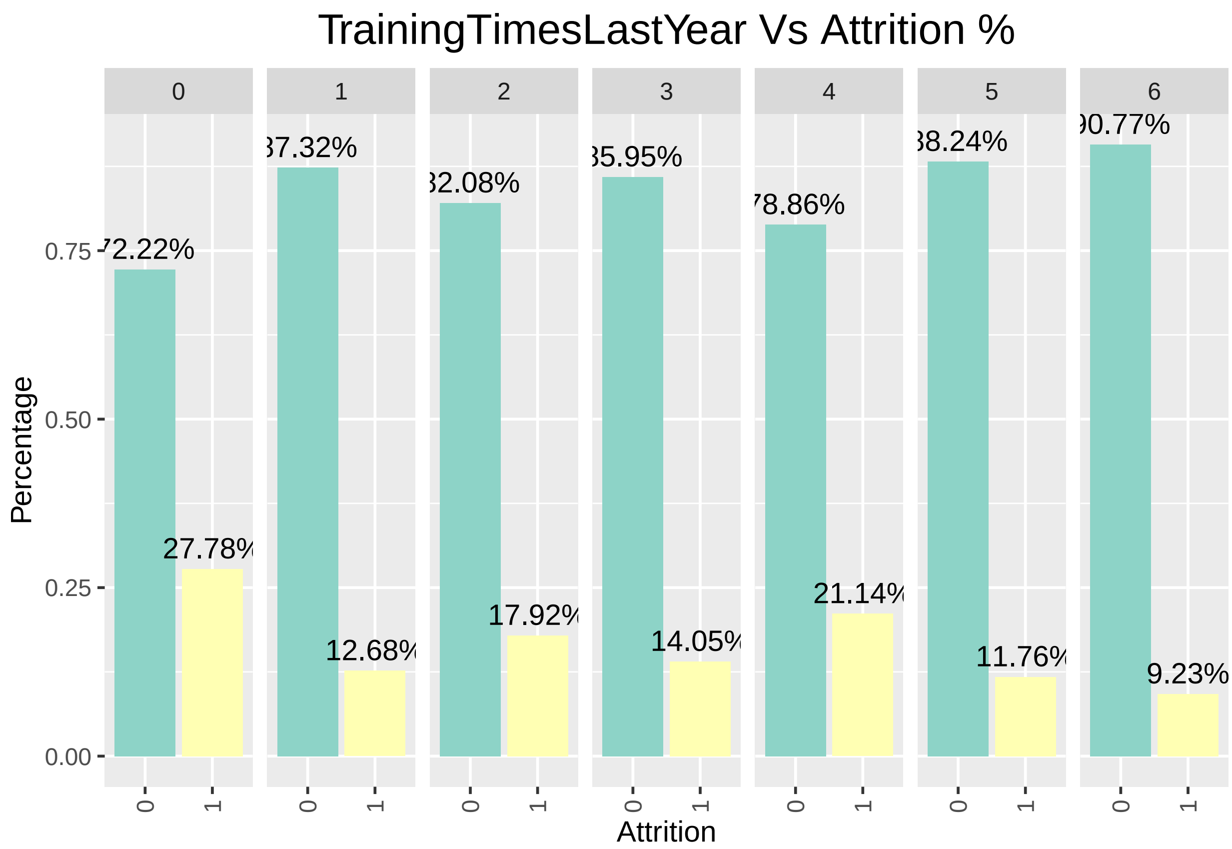
Understanding TrainingTimesLastYear - Number of times training was conducted for this employee last year.

```
table(data$TrainingTimesLastYear, data$Attrition)
```

```
##
##       0   1
##   0  39  15
##   1  62   9
##   2 449  98
##   3 422  69
##   4  97  26
##   5 105  14
##   6  59   6
```

```
ggplot(data,aes(x=Attrition,group=TrainingTimesLastYear))+
  geom_bar(aes(y=..prop..,fill=factor(..x..)),stat="count")+
  facet_grid(~TrainingTimesLastYear)+
  theme(axis.text.x=element_text(angle=90,vjust=0.5),legend.position="none",plot.title=element_text(siz
  labs(x="Attrition",y="Percentage",title="TrainingTimesLastYear Vs Attrition %")+
```

```
geom_text(aes(label = scales::percent(..prop..), y = ..prop.. ),stat= "count",vjust =-.5) +
scale_fill_brewer(palette="Set3")
```



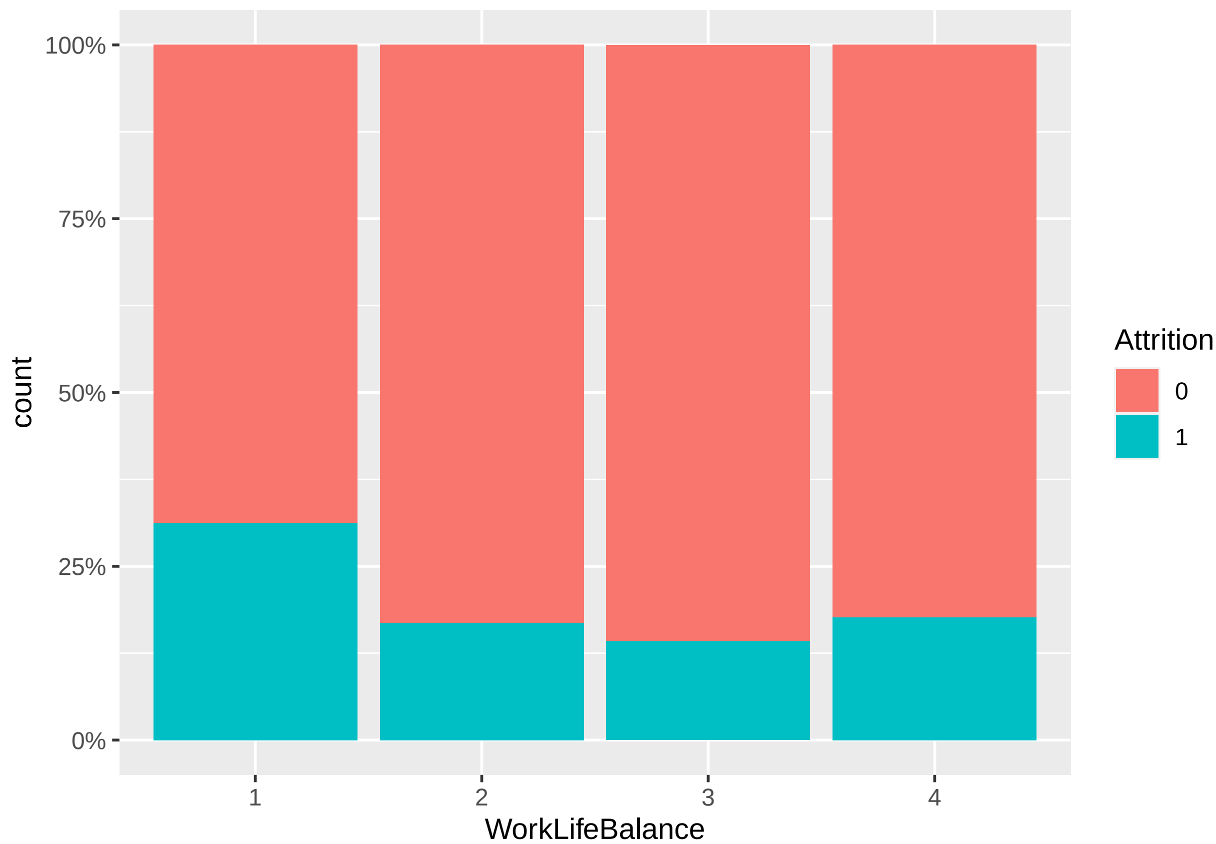
TrainingTimesLastYear predictor could effect Attrition.

Understanding WorkLifeBalance Predictor

```
summary(data$WorkLifeBalance)
```

```
##   1   2   3   4
##  80 344 893 153
```

```
worklifebalance_plot = ggplot(data,aes(WorkLifeBalance,fill=Attrition))+geom_bar(position="fill")+scale_
worklifebalance_plot
```



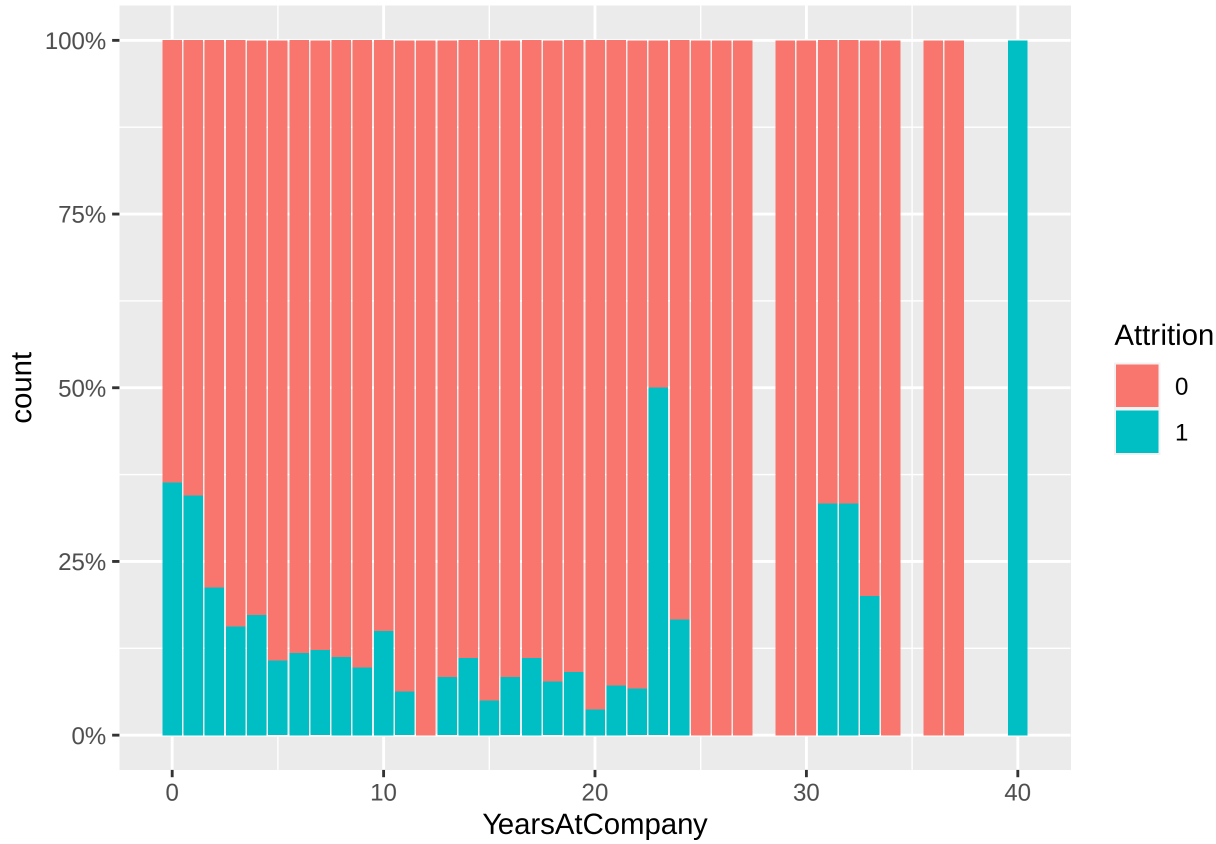
WorkLifeBalance could affect Attrition Rate.

Understanding YearsAtCompany Predictor

```
summary(data$YearsAtCompany)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   5.000   7.008   9.000  40.000
```

```
yearsatcompany_plot = ggplot(data,aes(YearsAtCompany,fill=Attrition))+geom_bar(position="fill")+scale_y_
yearsatcompany_plot
```



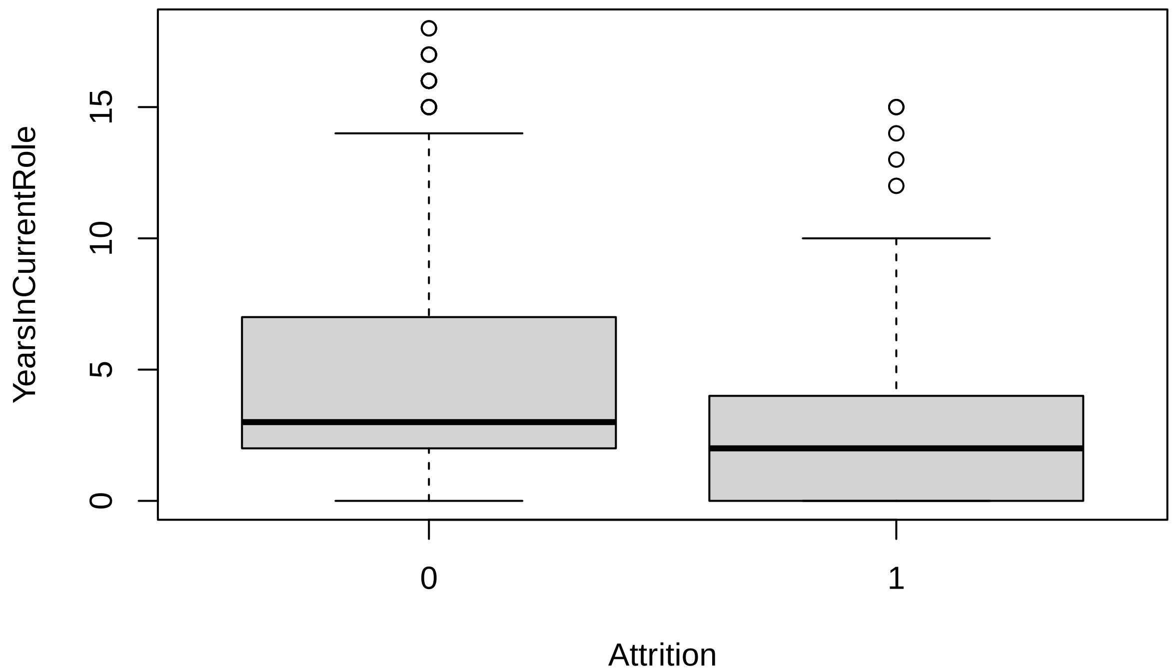
There is no significant relation between YearsAtCompany and Attrition.

Understanding YearsInCurrentRole Predictor

```
summary(data$YearsInCurrentRole)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   4.229   7.000  18.000
```

```
boxplot(YearsInCurrentRole~Attrition, data = data)
```



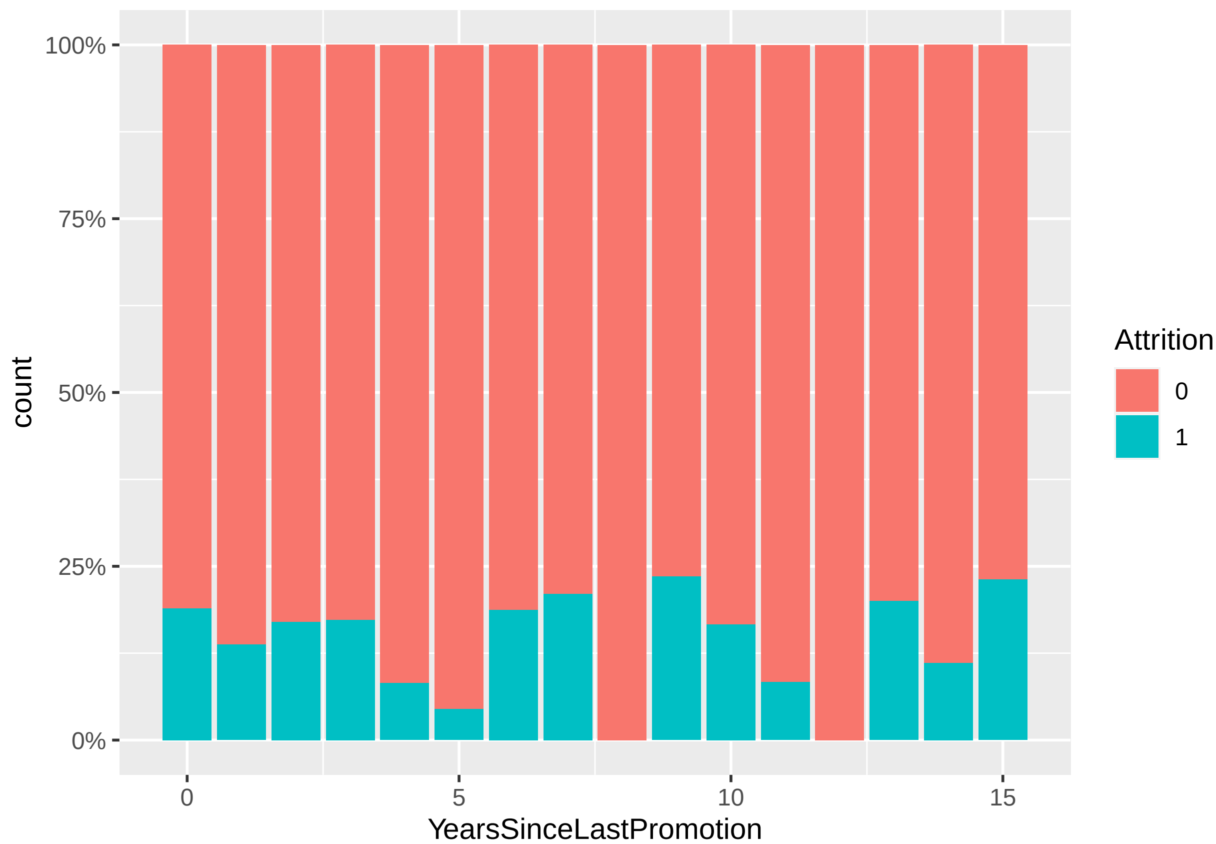
It is observed that lower attrition rates with Employees having Higher Years of Current Role in the company.

Understanding YearsSinceLastPromotion Predictor

```
summary(data$YearsSinceLastPromotion)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   2.188   3.000  15.000
```

```
yearssinceprom_plot = ggplot(data,aes(YearsSinceLastPromotion,fill=Attrition))+geom_bar(position="fill")
yearssinceprom_plot
```



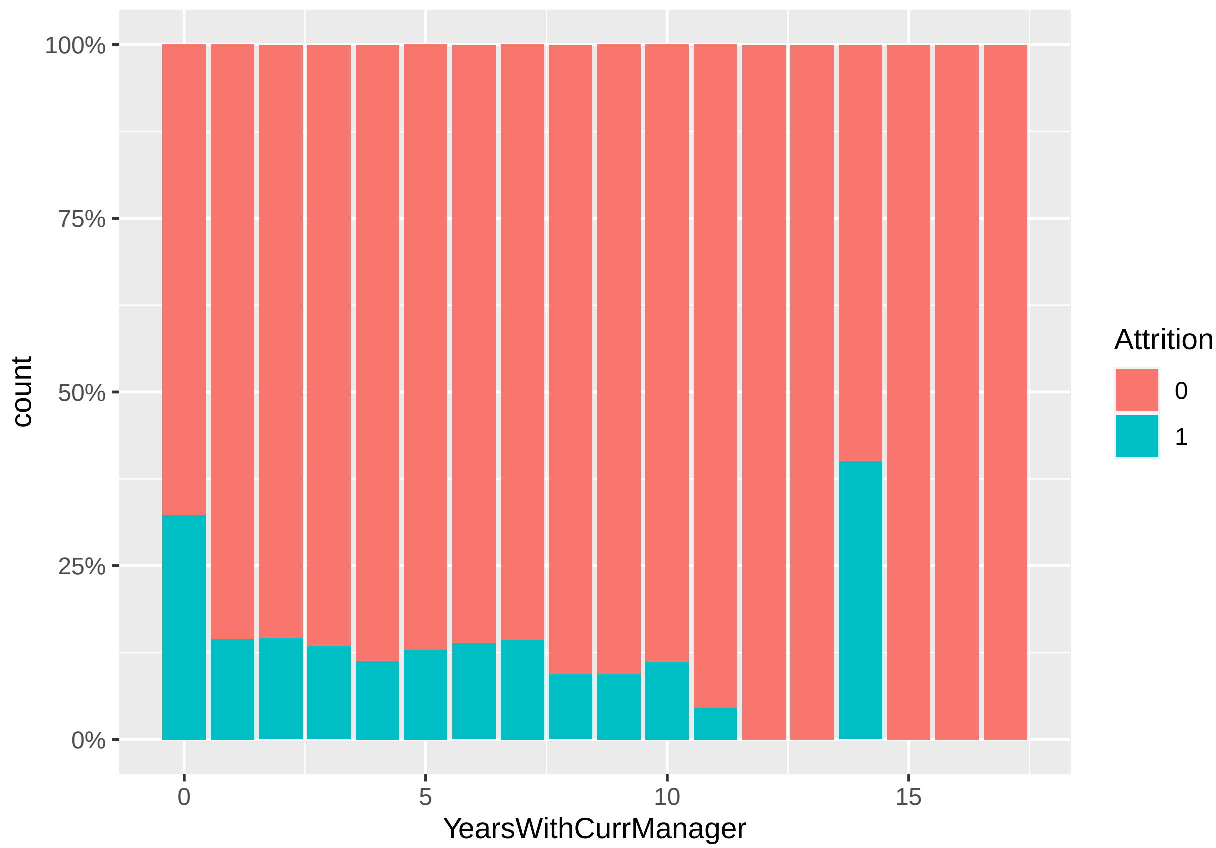
Not much significant relation is observed.

Understanding YearsWithCurrManager Predictor

```
summary(data$YearsWithCurrManager)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   2.000   3.000   4.123   7.000  17.000
```

```
yearswithmanager_plot = ggplot(data,aes(YearsWithCurrManager,fill=Attrition))+geom_bar(position="fill")+
yearswithmanager_plot
```



It is observed that as the YearswithCurrManager increases, the proportion of attrition decreases.

Based on EDA, - We found several features that have visible effect on the target variable: - **age**, **total_working_years**, **years_at_company**, **years_in_current_role** and **monthly_income** -numerical - **over_time**, **marital_status** and **job_role** - nominal categorical - **business_travel**, **job_level** and **stock_option_level** - ordinal categorical

- The profile of a worker who is most likely to be churned:

1. Young
2. Low salary
3. Working overtime
4. Single
5. Working as a sales rep or a lab tech
6. Has a low overall satisfaction level
7. Travels frequently
8. Has stock level set to 0

Model Building

As seen previously, there seems to be imbalance in the dataset, so we would be sampling the dataset using ovun.sample funcion which is a part of 'ROSE' package.

```
set.seed(1)
data_over = ovun.sample(Attrition~., data = data, method = "both", N = 1470)$data
print(table(data_over$Attrition))
```

```
##
##   0   1
## 774 696
```

Here, method='both' is a combination of over-sampling and under-sampling technique. The majority class i.e. '0' is under-sampled whereas the majority class i.e. '1' is over-sampled.

```
sapply(data_over,class)
```

```
##                      Age                Attrition           BusinessTravel
##                "integer"                 "factor"                 "factor"
##                DailyRate               Department         DistanceFromHome
##                "integer"                 "factor"                "integer"
##                Education           EducationField  EnvironmentSatisfaction
##                 "factor"                 "factor"                 "factor"
##                   Gender               HourlyRate           JobInvolvement
##                 "factor"                "integer"                 "factor"
##                 JobLevel                  JobRole          JobSatisfaction
##                 "factor"                 "factor"                 "factor"
##            MaritalStatus            MonthlyIncome              MonthlyRate
##                 "factor"                "integer"                "integer"
##       NumCompaniesWorked                 OverTime        PercentSalaryHike
##                "integer"                 "factor"                "integer"
##        PerformanceRating RelationshipSatisfaction         StockOptionLevel
##                 "factor"                 "factor"                 "factor"
##        TotalWorkingYears    TrainingTimesLastYear          WorkLifeBalance
##                "integer"                "integer"                 "factor"
##           YearsAtCompany       YearsInCurrentRole  YearsSinceLastPromotion
##                "integer"                "integer"                "integer"
##     YearsWithCurrManager
##                "integer"
```

Splitting the data in Test and Train data.

We have divided our sample in 70:30 ratio for train and test set.

```
set.seed(1)
split = sort(sample(nrow(data_over), nrow(data_over)*.7))
train=data_over[split,]
test=data_over[-split,]
```

Implementing Logistic Regression Model

```
set.seed(1)
glm.fit = glm(Attrition~.,data=train,family=binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Attrition ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6082  -0.5453  -0.0696   0.5144   3.2032
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -6.443e+00  5.354e+02  -0.012 0.990399
## Age                              -4.277e-02  1.445e-02  -2.959 0.003082 **
## BusinessTravelTravel_Frequently   3.081e+00  4.770e-01   6.459 1.05e-10 ***
## BusinessTravelTravel_Rarely       1.834e+00  4.407e-01   4.161 3.17e-05 ***
## DailyRate                        -9.709e-04  2.497e-04  -3.888 0.000101 ***
## DepartmentResearch & Development  1.211e+01  5.354e+02   0.023 0.981961
## DepartmentSales                   1.074e+01  5.354e+02   0.020 0.983992
## DistanceFromHome                  7.814e-02  1.258e-02   6.212 5.25e-10 ***
## Education2                       -2.261e-01  3.837e-01  -0.589 0.555563
## Education3                       -7.874e-02  3.251e-01  -0.242 0.808613
## Education4                        2.648e-01  3.548e-01   0.746 0.455463
## Education5                       -1.184e-01  6.264e-01  -0.189 0.850051
## EducationFieldLife Sciences      -2.052e+00  9.541e-01  -2.150 0.031528 *
## EducationFieldMarketing          -1.250e+00  1.011e+00  -1.237 0.216024
## EducationFieldMedical            -1.793e+00  9.462e-01  -1.895 0.058085 .
## EducationFieldOther              -1.857e+00  1.055e+00  -1.761 0.078166 .
## EducationFieldTechnical Degree   -2.168e-01  9.971e-01  -0.217 0.827834
## EnvironmentSatisfaction2         -1.215e+00  3.000e-01  -4.050 5.11e-05 ***
## EnvironmentSatisfaction3         -1.127e+00  2.784e-01  -4.048 5.16e-05 ***
## EnvironmentSatisfaction4         -1.182e+00  2.780e-01  -4.252 2.12e-05 ***
## GenderMale                        3.704e-02  2.055e-01   0.180 0.856961
## HourlyRate                        2.036e-03  5.158e-03   0.395 0.693003
## JobInvolvement2                  -1.494e+00  4.169e-01  -3.584 0.000338 ***
## JobInvolvement3                  -1.420e+00  3.981e-01  -3.566 0.000363 ***
## JobInvolvement4                  -2.130e+00  5.008e-01  -4.252 2.12e-05 ***
## JobLevel2                        -9.290e-01  4.331e-01  -2.145 0.031968 *
## JobLevel3                         1.569e+00  7.850e-01   1.999 0.045603 *
## JobLevel4                         1.474e+00  1.186e+00   1.243 0.214021
## JobLevel5                         5.950e+00  1.612e+00   3.692 0.000223 ***
## JobRoleHuman Resources            1.253e+01  5.354e+02   0.023 0.981333
## JobRoleLaboratory Technician      1.799e+00  6.248e-01   2.879 0.003994 **
## JobRoleManager                    1.619e+00  9.384e-01   1.726 0.084431 .
## JobRoleManufacturing Director     1.448e+00  5.782e-01   2.505 0.012257 *
## JobRoleResearch Director         -2.177e+00  1.159e+00  -1.879 0.060229 .
## JobRoleResearch Scientist         6.834e-01  6.235e-01   1.096 0.273076
## JobRoleSales Executive            3.676e+00  1.276e+00   2.881 0.003964 **
## JobRoleSales Representative       3.174e+00  1.310e+00   2.423 0.015411 *
## JobSatisfaction2                 -8.685e-01  3.186e-01  -2.726 0.006414 **
## JobSatisfaction3                 -1.139e+00  2.911e-01  -3.914 9.07e-05 ***
## JobSatisfaction4                 -1.258e+00  2.990e-01  -4.206 2.60e-05 ***
## MaritalStatusMarried              6.282e-01  2.989e-01   2.101 0.035616 *
## MaritalStatusSingle               8.934e-01  4.290e-01   2.083 0.037268 *
## MonthlyIncome                    -2.545e-04  9.797e-05  -2.598 0.009388 **
## MonthlyRate                      -3.728e-06  1.357e-05  -0.275 0.783569
```

```
## NumCompaniesWorked                1.941e-01  4.597e-02   4.224 2.40e-05 ***
## OverTimeYes                       1.744e+00  2.178e-01   8.006 1.18e-15 ***
## PercentSalaryHike                -2.175e-02  4.486e-02  -0.485 0.627704
## PerformanceRating4                4.082e-01  4.386e-01   0.931 0.351944
## RelationshipSatisfaction2        -6.801e-01  3.353e-01  -2.028 0.042542 *
## RelationshipSatisfaction3        -8.887e-01  2.751e-01  -3.230 0.001236 **
## RelationshipSatisfaction4        -7.988e-01  2.826e-01  -2.827 0.004700 **
## StockOptionLevel1                -9.784e-01  3.364e-01  -2.908 0.003635 **
## StockOptionLevel2                -1.139e+00  4.356e-01  -2.615 0.008925 **
## StockOptionLevel3                 7.495e-01  5.410e-01   1.385 0.165939
## TotalWorkingYears                -2.123e-02  2.962e-02  -0.717 0.473531
## TrainingTimesLastYear            -1.356e-01  7.550e-02  -1.796 0.072442 .
## WorkLifeBalance2                 -7.953e-01  4.228e-01  -1.881 0.059950 .
## WorkLifeBalance3                 -1.426e+00  4.060e-01  -3.512 0.000445 ***
## WorkLifeBalance4                 -7.312e-01  5.087e-01  -1.437 0.150604
## YearsAtCompany                    1.006e-01  3.897e-02   2.583 0.009807 **
## YearsInCurrentRole               -2.266e-01  5.280e-02  -4.292 1.77e-05 ***
## YearsSinceLastPromotion           2.039e-01  4.727e-02   4.313 1.61e-05 ***
## YearsWithCurrManager             -2.364e-01  5.227e-02  -4.523 6.09e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1419.47  on 1028  degrees of freedom
## Residual deviance:  756.85  on  966  degrees of freedom
## AIC: 882.85
##
## Number of Fisher Scoring iterations: 12
```

As you can see, the significant variables in predicting Attrition through Logistic regression are somewhat similar to our findings through EDA. Mainly factors such as Age, MonthlyIncome, JobRole, YearsAtCompany and OverTime have been seen to affect Attrition. This means, Employees mostly change their jobs for better pay in the early years of their career.

Predicting using fitted Logitsic Regression Model

```
glm.probs = predict(glm.fit, test, type="response")
glm.pred=rep(0,length(glm.probs))
glm.pred[glm.probs > 0.5] <- 1
table(glm.pred,test$Attrition)
```

```
##
## glm.pred   0   1
##        0 164  41
##        1  53 183
```

Model Statistics

```
confusionMatrix(as.factor(glm.pred), test$Attrition, mode = "prec_recall", positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 164  41
##          1  53 183
##
##                Accuracy : 0.7868
##                  95% CI : (0.7456, 0.8242)
##     No Information Rate : 0.5079
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.5732
##
##  Mcnemar's Test P-Value : 0.2566
##
##               Precision : 0.7754
##                  Recall : 0.8170
##                      F1 : 0.7957
##              Prevalence : 0.5079
##          Detection Rate : 0.4150
##    Detection Prevalence : 0.5351
##       Balanced Accuracy : 0.7864
##
##        'Positive' Class : 1
##
```

Thus, we have achieved 79% accuracy with Logistic Regression Model. The Precision, Recall and F1 score is also 79%.

Let us now implement Linear Discrimant Analysis (LDA) model.

```
set.seed(1)
lda.fit = lda(Attrition~.,data=train)
lda.fit
```

```
## Call:
## lda(Attrition ~ ., data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.5413022 0.4586978
##
## Group means:
##        Age BusinessTravelTravel_Frequently BusinessTravelTravel_Rarely
```

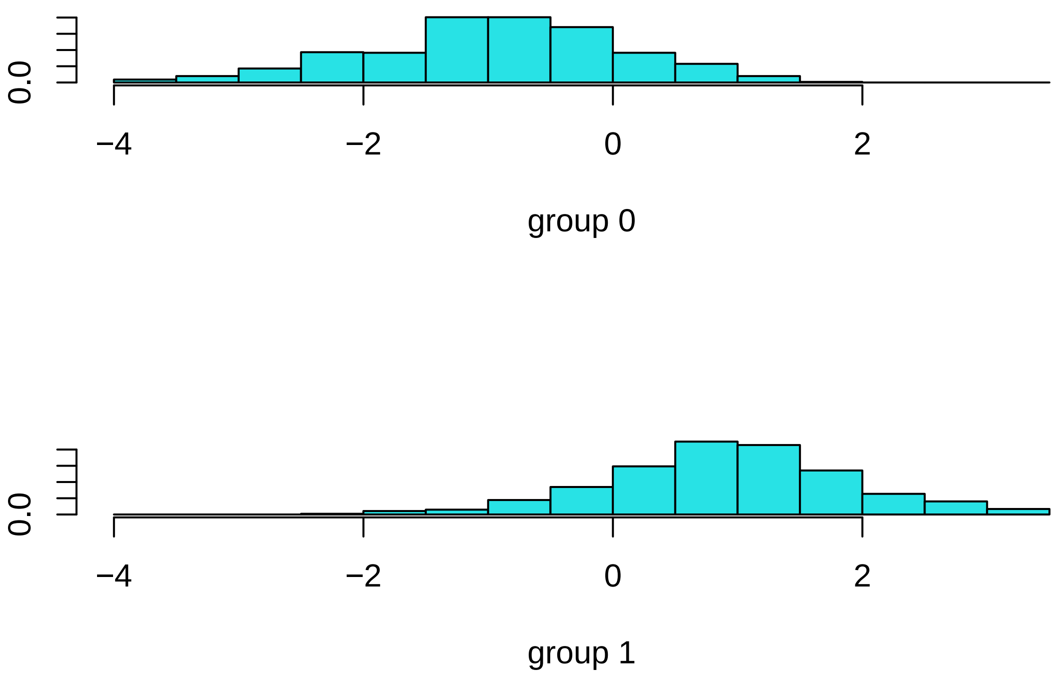
```
## 0 37.38959                      0.1633752                    0.7181329
## 1 33.50636                      0.3050847                    0.6546610
##   DailyRate DepartmentResearch & Development DepartmentSales DistanceFromHome
## 0  826.1221                        0.6768402       0.2836625         8.504488
## 1  708.7140                        0.5593220       0.3813559        11.612288
##   Education2 Education3 Education4 Education5 EducationFieldLife Sciences
## 0  0.2010772  0.3895871  0.2675045 0.03949731                  0.4578097
## 1  0.1567797  0.4385593  0.2542373 0.02754237                  0.3792373
##   EducationFieldMarketing EducationFieldMedical EducationFieldOther
## 0              0.09694794             0.3267504          0.03770197
## 1              0.16101695             0.2627119          0.03177966
##   EducationFieldTechnical Degree EnvironmentSatisfaction2
## 0                     0.06642729                0.2046679
## 1                     0.12076271                0.1927966
##   EnvironmentSatisfaction3 EnvironmentSatisfaction4 GenderMale HourlyRate
## 0                0.3105925                0.3087971  0.6391382   65.03411
## 1                0.2415254                0.2563559  0.6377119   64.98305
##   JobInvolvement2 JobInvolvement3 JobInvolvement4 JobLevel2 JobLevel3
## 0       0.2692998       0.5583483      0.13105925 0.4272890 0.1490126
## 1       0.2881356       0.5105932      0.08262712 0.2245763 0.1398305
##    JobLevel4  JobLevel5 JobRoleHuman Resources JobRoleLaboratory Technician
## 0 0.07001795 0.04667864             0.03770197                    0.1292639
## 1 0.03177966 0.01906780             0.05932203                    0.2372881
##   JobRoleManager JobRoleManufacturing Director JobRoleResearch Director
## 0     0.06463196                    0.11490126              0.061041293
## 1     0.02542373                    0.05508475              0.006355932
##   JobRoleResearch Scientist JobRoleSales Executive JobRoleSales Representative
## 0                 0.2154399              0.2082585                  0.04667864
## 1                 0.2139831              0.2436441                  0.13135593
##   JobSatisfaction2 JobSatisfaction3 JobSatisfaction4 MaritalStatusMarried
## 0        0.1849192        0.3213645        0.3339318            0.5134650
## 1        0.1864407        0.2902542        0.2224576            0.3495763
##   MaritalStatusSingle MonthlyIncome MonthlyRate NumCompaniesWorked OverTimeYes
## 0           0.2657092      6642.447    14020.20           2.511670   0.2746858
## 1           0.5105932      4761.947    14350.97           3.002119   0.5148305
##   PercentSalaryHike PerformanceRating4 RelationshipSatisfaction2
## 0          15.29803          0.1579892                 0.1849192
## 1          15.21398          0.1525424                 0.1779661
##   RelationshipSatisfaction3 RelationshipSatisfaction4 StockOptionLevel1
## 0                 0.3267504                 0.2818671         0.4578097
## 1                 0.3241525                 0.2775424         0.2245763
##   StockOptionLevel2 StockOptionLevel3 TotalWorkingYears TrainingTimesLastYear
## 0        0.11490126        0.04488330         11.716338              2.850987
## 1        0.04237288        0.08686441          8.383475              2.739407
##   WorkLifeBalance2 WorkLifeBalance3 WorkLifeBalance4 YearsAtCompany
## 0        0.2315978        0.6337522       0.08976661       7.535009
## 1        0.2563559        0.5444915       0.09745763       4.934322
##   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## 0           4.624776                2.184919             4.477558
## 1           2.665254                1.737288             2.707627
##
## Coefficients of linear discriminants:
##                                           LD1
## Age                             -2.081583e-02
```

```
## BusinessTravelTravel_Frequently   1.171729e+00
## BusinessTravelTravel_Rarely       5.979039e-01
## DailyRate                        -5.433801e-04
## DepartmentResearch & Development  1.427603e+00
## DepartmentSales                   8.988346e-01
## DistanceFromHome                  3.457423e-02
## Education2                       -1.753681e-01
## Education3                       -1.714698e-01
## Education4                        2.335744e-02
## Education5                       -1.447154e-01
## EducationFieldLife Sciences      -7.609260e-01
## EducationFieldMarketing          -3.583975e-01
## EducationFieldMedical            -6.768198e-01
## EducationFieldOther              -8.111108e-01
## EducationFieldTechnical Degree   -1.189130e-01
## EnvironmentSatisfaction2         -5.958159e-01
## EnvironmentSatisfaction3         -6.473596e-01
## EnvironmentSatisfaction4         -5.250601e-01
## GenderMale                       -2.439039e-03
## HourlyRate                        6.003944e-05
## JobInvolvement2                  -6.596193e-01
## JobInvolvement3                  -6.509699e-01
## JobInvolvement4                  -9.641117e-01
## JobLevel2                        -5.589839e-01
## JobLevel3                         5.833072e-01
## JobLevel4                         6.598786e-01
## JobLevel5                         2.304364e+00
## JobRoleHuman Resources            1.227563e+00
## JobRoleLaboratory Technician      5.673820e-01
## JobRoleManager                    5.038768e-01
## JobRoleManufacturing Director     2.791937e-01
## JobRoleResearch Director         -6.893241e-01
## JobRoleResearch Scientist        -8.242104e-02
## JobRoleSales Executive            1.300095e+00
## JobRoleSales Representative       1.112473e+00
## JobSatisfaction2                 -4.457339e-01
## JobSatisfaction3                 -4.809500e-01
## JobSatisfaction4                 -6.985108e-01
## MaritalStatusMarried              4.800384e-02
## MaritalStatusSingle               1.077497e-01
## MonthlyIncome                    -1.197200e-04
## MonthlyRate                       8.289869e-07
## NumCompaniesWorked                1.090179e-01
## OverTimeYes                       8.244903e-01
## PercentSalaryHike                -2.157933e-02
## PerformanceRating4                3.252682e-01
## RelationshipSatisfaction2        -4.197600e-01
## RelationshipSatisfaction3        -4.675726e-01
## RelationshipSatisfaction4        -4.275032e-01
## StockOptionLevel1                -6.437769e-01
## StockOptionLevel2                -8.197401e-01
## StockOptionLevel3                 1.302657e-01
## TotalWorkingYears                -1.373494e-02
## TrainingTimesLastYear            -7.160043e-02
```

```
## WorkLifeBalance2                  -2.440770e-01
## WorkLifeBalance3                  -5.558325e-01
## WorkLifeBalance4                  -1.443895e-01
## YearsAtCompany                     5.134084e-02
## YearsInCurrentRole                -9.758016e-02
## YearsSinceLastPromotion            6.935801e-02
## YearsWithCurrManager              -9.328409e-02
```

Plotting the model

```
plot(lda.fit)
```



Prediction using LDA model

```
lda.pred=predict (lda.fit,test)
names(lda.pred)
```

```
## [1] "class"     "posterior" "x"
```

```
lda.class=lda.pred$class
```

Model Performance and Statistics.

```
confusionMatrix(as.factor(lda.class), test$Attrition, mode = "prec_recall", positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 163  33
##          1  54 191
##
##                Accuracy : 0.8027
##                  95% CI : (0.7625, 0.8389)
##     No Information Rate : 0.5079
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.6047
##
##  Mcnemar's Test P-Value : 0.03201
##
##               Precision : 0.7796
##                  Recall : 0.8527
##                      F1 : 0.8145
##              Prevalence : 0.5079
##          Detection Rate : 0.4331
##    Detection Prevalence : 0.5556
##       Balanced Accuracy : 0.8019
##
##        'Positive' Class : 1
##
```

The coefficients of linear discriminants output provide the linear combination of all the predictor variables that are used to perform the LDA decision rule. We plotted these discriminants. It is seen that these classes overlap to some extent and LDA performs poorly when compared to Logistic regression. We have also achieved a similar accuracy as of Logistic Regression for Linear Discriminant Analysis model i.e. 79%.

Implementing Decision Tree Model.

```
set.seed(1)
tree.ibm = tree(Attrition~., data=train)
summary(tree.ibm)
```

```
##
## Classification tree:
## tree(formula = Attrition ~ ., data = train)
## Variables actually used in tree construction:
##  [1] "StockOptionLevel"         "MonthlyIncome"
##  [3] "JobRole"                  "RelationshipSatisfaction"
##  [5] "DailyRate"                "Age"
```

```
##  [7] "DistanceFromHome"         "MonthlyRate"
##  [9] "PercentSalaryHike"        "JobSatisfaction"
## [11] "EnvironmentSatisfaction"  "TotalWorkingYears"
## [13] "YearsAtCompany"           "EducationField"
## [15] "TrainingTimesLastYear"    "OverTime"
## [17] "NumCompaniesWorked"       "JobLevel"
## Number of terminal nodes:  34
## Residual mean deviance:  0.6307 = 627.5 / 995
## Misclassification error rate: 0.1293 = 133 / 1029
```

As we can see, out of the 36 variables that we had, these 18 variables were actually used by the model for tree construction. The Residual mean deviance which is a measure of the error remaining in the tree after construction is 61.75%. We see that the 'Misclassification error rate' which is the proportion of observations that were predicted to fall in another class than they actually did is 13.4%

Prediction using Tree Model.

```r
tree.pred = predict(tree.ibm , test, type = "class")
confusionMatrix(tree.pred, test$Attrition, mode = "prec_recall", positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 179  59
##          1  38 165
##
##                Accuracy : 0.78
##                  95% CI : (0.7384, 0.8178)
##     No Information Rate : 0.5079
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.5606
##
##  Mcnemar's Test P-Value : 0.04229
##
##               Precision : 0.8128
##                  Recall : 0.7366
##                      F1 : 0.7728
##              Prevalence : 0.5079
##          Detection Rate : 0.3741
##    Detection Prevalence : 0.4603
##       Balanced Accuracy : 0.7807
##
##        'Positive' Class : 1
##
```

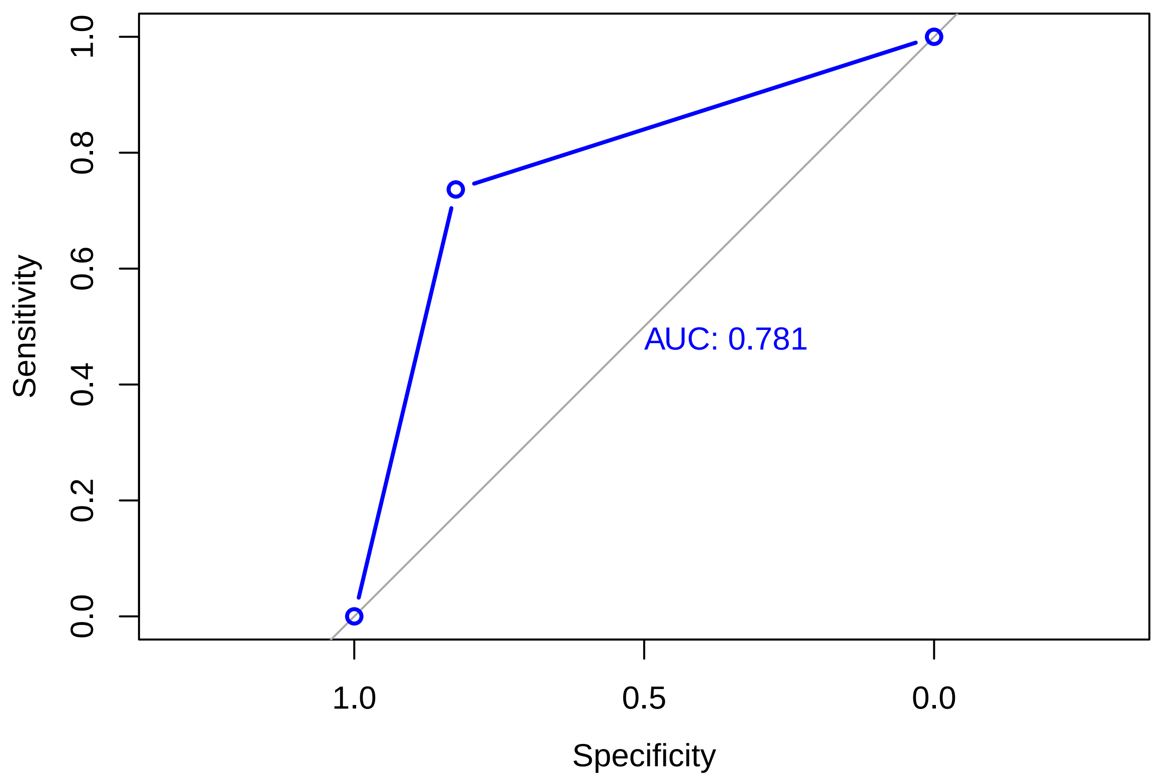
We then evaluated the performance of this model on the testing set. The model was able to achieve an accuracy of 79.14% on the test dataset with an F1 score of 80%. F1 score is calculated using a harmonic mean of precision and recall.

Roc Curve to determine the how well the fit is (model accuracy)

```
dtree.plot = plot.roc (as.numeric(test$Attrition), as.numeric(tree.pred),lwd=2, type="b", print.auc=TRUE
```

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```



AUC of 0.781 makes it a decent fit for this dataset. We will try fitting other models to check if we can improve on these results.

Since the performance of this model is relatively poor, so we tried implementing some o†her techniques to improve its performance in terms of accuracy, F1 score and AUC value.

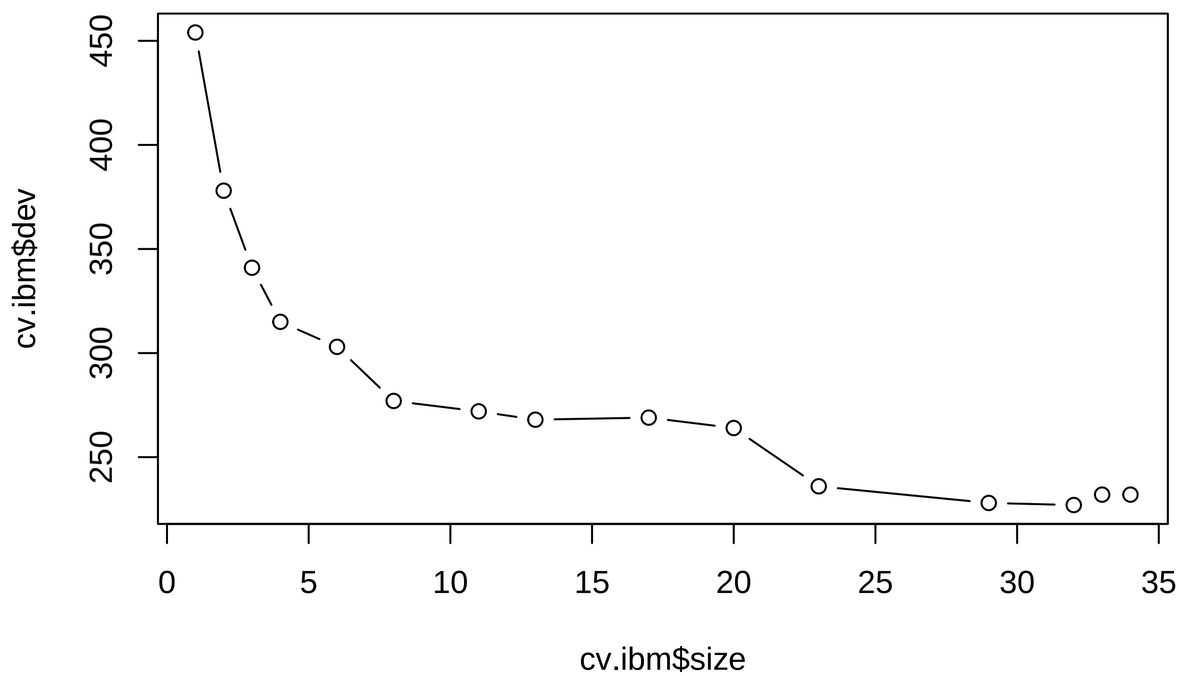
Firstly, we tried a pruning the tree to check whether we can achieve a better performance.

```
cv.ibm = cv.tree(tree.ibm , FUN = prune.misclass)
cv.ibm
```

```
## $size
##  [1] 34 33 32 29 23 20 17 13 11  8  6  4  3  2  1
##
## $dev
##  [1] 232 232 227 228 236 264 269 268 272 277 303 315 341 378 454
##
```

```
## $k
##  [1]       -Inf   0.000000   1.000000   1.666667   3.666667   4.666667
##  [7]   5.000000   5.500000   6.500000   7.333333  10.500000  11.000000
## [13]  20.000000  54.000000 108.000000
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```
plot(cv.ibm$size , cv.ibm$dev, type = "b")
```



Cv.ibm$dev corresponds to the number of cross-validation errors. The tree with 32 terminal nodes results in the least cross-validation errors. Hence, we consider 32 nodes while building the pruned model.

```
prune.ibm = prune.misclass(tree.ibm , best = 32)
tree.pred.prune = predict(prune.ibm , test, type = "class")
confusionMatrix(tree.pred.prune, test$Attrition, mode = "prec_recall", positive="1")
```

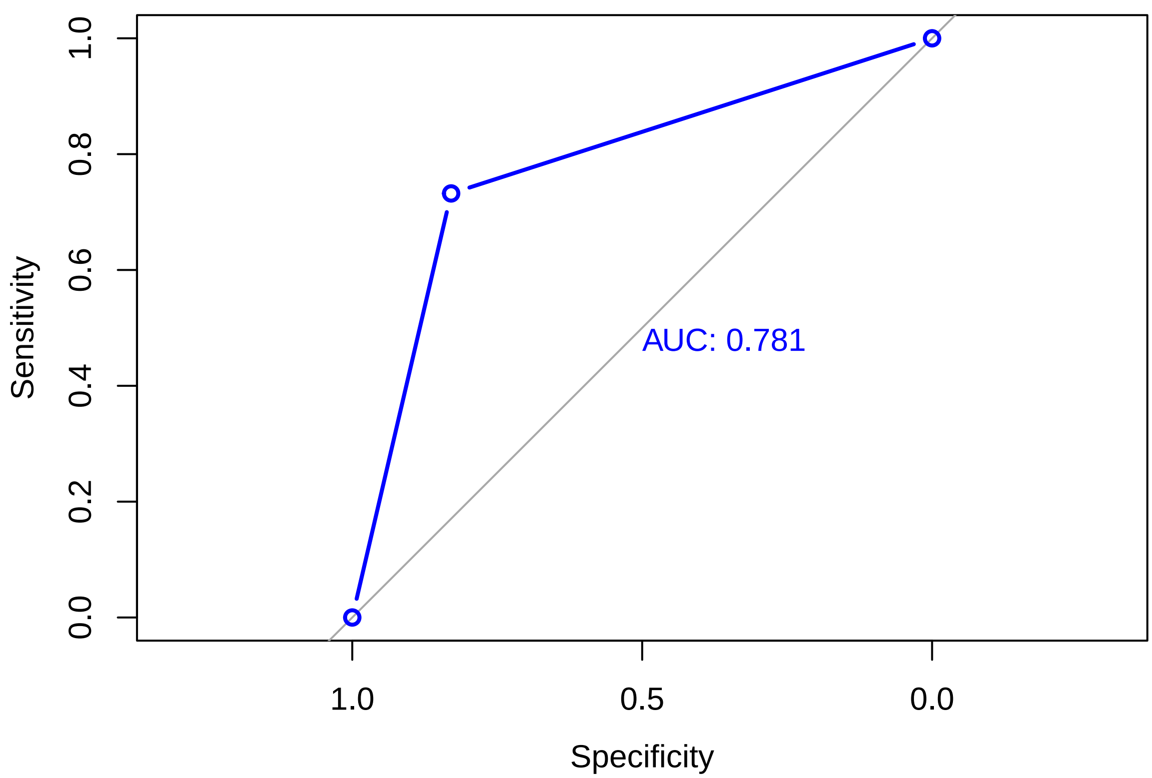
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 180  60
##          1  37 164
```

```
##
##                Accuracy : 0.78
##                  95% CI : (0.7384, 0.8178)
##     No Information Rate : 0.5079
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.5607
##
##  Mcnemar's Test P-Value : 0.0255
##
##               Precision : 0.8159
##                  Recall : 0.7321
##                      F1 : 0.7718
##              Prevalence : 0.5079
##          Detection Rate : 0.3719
##    Detection Prevalence : 0.4558
##       Balanced Accuracy : 0.7808
##
##        'Positive' Class : 1
##
```

```
dtree.pred.prune = plot.roc (as.numeric(test$Attrition), as.numeric(tree.pred.prune),lwd=2, type="b", p
```

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```



We do not see great improvement in the accuracy of the model with respect to the previous non-pruned model. Even though there is a slight improvement in the precision, there's a corresponding dip in the recall which leads to an overall low F1 score as compared to the previous model. Hence, we can conclude that pruning the tree is not that beneficial in this case. Not much difference is observed in AUC value as well.

Our next approach is to implement Bagging model.

```
set.seed (1)
bag.ibm = randomForest(Attrition~., data = train, mtry = ncol(train)-1, importance = TRUE)
```

Bagging is simply a special case of a random forest where the value of mtry is equal to the total number of all predictors.

```
yhat.bag = predict(bag.ibm , newdata = test)
confusionMatrix(yhat.bag, test$Attrition, mode = "prec_recall", positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 203  24
##          1  14 200
##
##                Accuracy : 0.9138
##                  95% CI : (0.8836, 0.9383)
##     No Information Rate : 0.5079
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8277
##
##  Mcnemar's Test P-Value : 0.1443
##
##               Precision : 0.9346
##                  Recall : 0.8929
##                      F1 : 0.9132
##              Prevalence : 0.5079
##          Detection Rate : 0.4535
##    Detection Prevalence : 0.4853
##       Balanced Accuracy : 0.9142
##
##        'Positive' Class : 1
##
```

We then evaluated the performance of this model on the testing set. The model was able to achieve an accuracy of 90.93% on the test dataset with an F1 score of 90.78%.

The performance of this model is better than all the previously implemented models.

Next, we consider the performance of random forest model.

```
set.seed (1)
bag.ibm2 = randomForest(Attrition~., data = train, mtry = sqrt((ncol(train)-1)), importance = TRUE)
```

By default, we use the value of mtry as the square root of the total number of predictors in case of random forest for classification.

```
yhat.bag2 = predict(bag.ibm2 , newdata = test)
confusionMatrix(yhat.bag2, test$Attrition, mode = "prec_recall", positive="1")
```

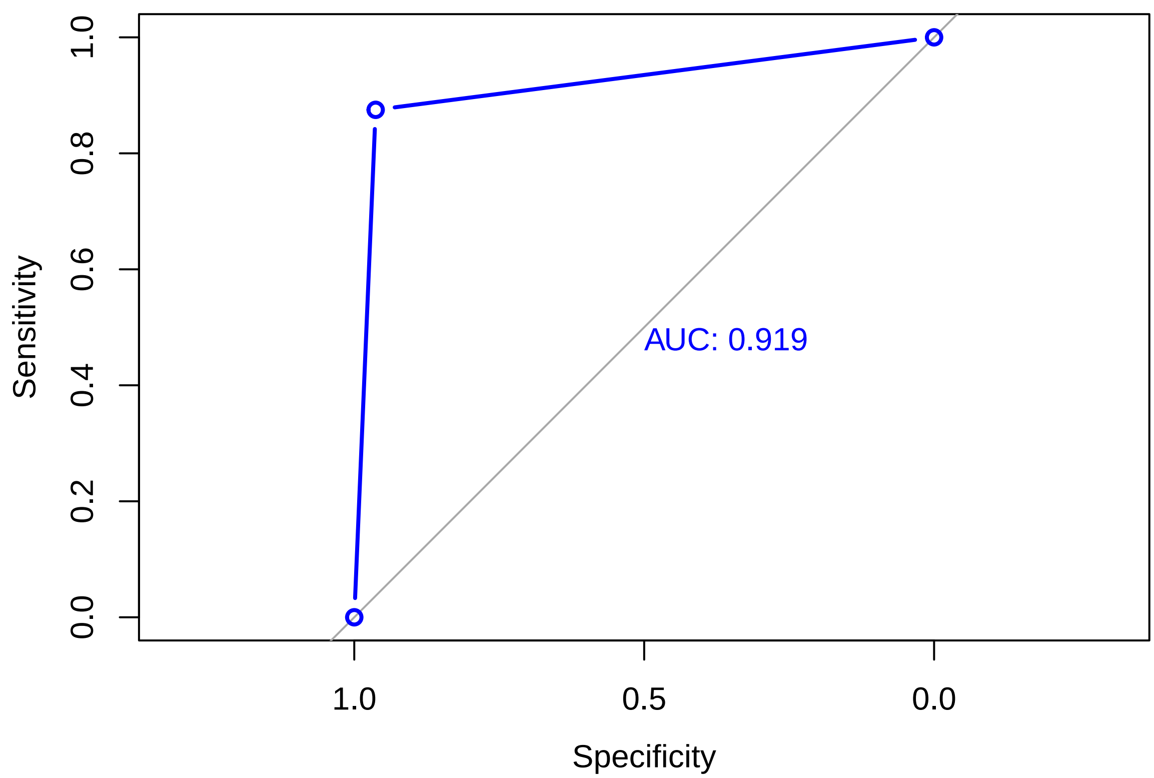
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 209  28
##          1   8 196
##
##                Accuracy : 0.9184
##                  95% CI : (0.8888, 0.9422)
##     No Information Rate : 0.5079
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8369
##
##  Mcnemar's Test P-Value : 0.001542
##
##               Precision : 0.9608
##                  Recall : 0.8750
##                      F1 : 0.9159
##              Prevalence : 0.5079
##          Detection Rate : 0.4444
##    Detection Prevalence : 0.4626
##       Balanced Accuracy : 0.9191
##
##        'Positive' Class : 1
##
```

Roc Plot to determine, how good the fit is.

```
rf.Plot = plot.roc (as.numeric(test$Attrition), as.numeric(yhat.bag2),lwd=2, type="b", print.auc=TRUE,c
```

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```



We then evaluated the performance of this model on the testing set. The model was able to achieve an accuracy and f1 score of around 91% on the test set which is one of the best rates we have got so far. AUC is observed as .914 making it a competitive fit.

```
importance(bag.ibm2)
```

```
##                                  0        1 MeanDecreaseAccuracy
## Age                      17.890526 32.43438             32.34510
## BusinessTravel           12.399408 20.16391             20.52055
## DailyRate                15.525404 31.36228             30.99810
## Department                6.844469 11.97805             12.27422
## DistanceFromHome         18.655435 29.98558             30.16736
## Education                 8.375052 25.49180             24.74226
## EducationField           14.475154 25.71369             24.96125
## EnvironmentSatisfaction  14.562928 29.39126             28.40887
## Gender                    5.835782 12.94855             12.84834
## HourlyRate               14.000443 31.88072             31.55444
## JobInvolvement            9.865191 22.32520             21.47961
## JobLevel                 10.734669 17.32406             18.15318
## JobRole                  21.080798 34.97145             34.92047
## JobSatisfaction          15.375766 28.38874             27.91289
## MaritalStatus            11.721746 18.69758             17.89097
## MonthlyIncome            16.757295 30.64315             31.44145
## MonthlyRate              14.881528 31.87469             31.83737
## NumCompaniesWorked       13.606815 23.76972             23.70093
## OverTime                 17.194806 22.37437             22.46641
```

```
## PercentSalaryHike        10.590402 30.16960             28.00018
## PerformanceRating         4.014605 11.00222             10.55441
## RelationshipSatisfaction 11.318147 26.23583             25.71627
## StockOptionLevel         18.904499 26.82558             26.82016
## TotalWorkingYears        14.852907 24.32693             25.79663
## TrainingTimesLastYear    10.321853 25.75684             26.24450
## WorkLifeBalance          11.555229 25.73331             25.23956
## YearsAtCompany           15.488666 22.94337             23.42862
## YearsInCurrentRole       14.836584 19.97720             21.07012
## YearsSinceLastPromotion   4.288314 22.91975             21.87705
## YearsWithCurrManager     16.501274 21.90507             22.82140
##                          MeanDecreaseGini
## Age                             32.338022
## BusinessTravel                   9.067684
## DailyRate                       25.536334
## Department                       4.112816
## DistanceFromHome                24.897491
## Education                       12.168152
## EducationField                  14.666204
## EnvironmentSatisfaction         15.982404
## Gender                           3.250059
## HourlyRate                      21.080278
## JobInvolvement                   9.578505
## JobLevel                        14.627105
## JobRole                         33.645039
## JobSatisfaction                 16.409389
## MaritalStatus                   11.431949
## MonthlyIncome                   35.153909
## MonthlyRate                     22.252166
## NumCompaniesWorked              15.269036
## OverTime                        14.555717
## PercentSalaryHike               15.480322
## PerformanceRating                2.127633
## RelationshipSatisfaction        12.539639
## StockOptionLevel                24.248435
## TotalWorkingYears               23.142667
## TrainingTimesLastYear           11.757409
## WorkLifeBalance                 12.033814
## YearsAtCompany                  25.654140
## YearsInCurrentRole              17.529043
## YearsSinceLastPromotion          9.390777
## YearsWithCurrManager            20.367301
```

We went on to see the features which impact the most to our response variable.

As observed in the EDA and previous analysis, here also we can see that Age, JobRole and MonthlyIncome have a high gini index which means that they have a high importance.

However, overtime and department do not play much role as opposed to the results from EDA.

Lastly we tried implementing the SVM Model in the dataset. As mentioned earlier, of having an imbalanced dataset, we tried implementing with both imbalanced and balanced dataset.

```
# Again reading the data set.
input_data = read.csv("HR_Employee_Attrition.csv")
```

```r
#Making necessary variables as factors
input_data$Attrition = as.factor(input_data$Attrition)
input_data$BusinessTravel = as.factor(input_data$BusinessTravel)
input_data$Department = as.factor(input_data$Department)
input_data$Gender = as.factor(input_data$Gender)
input_data$JobRole = as.factor(input_data$JobRole)
input_data$MaritalStatus = as.factor(input_data$MaritalStatus)
input_data$EducationField = as.factor(input_data$EducationField)
input_data$Education = as.factor(input_data$Education)
input_data$JobLevel = as.factor(input_data$JobLevel)
input_data$StockOptionLevel = as.factor(input_data$StockOptionLevel)
input_data$EnvironmentSatisfaction = as.factor(input_data$EnvironmentSatisfaction)
input_data$JobSatisfaction = as.factor(input_data$JobSatisfaction)
input_data$WorkLifeBalance = as.factor(input_data$WorkLifeBalance)
input_data$JobInvolvement = as.factor(input_data$JobInvolvement)
input_data$PerformanceRating = as.factor(input_data$PerformanceRating)
input_data$OverTime = as.factor(input_data$OverTime)
```

First implementing the SVM model on an imbalanced data set. As seen below an imbalance is of around 85%-15%.

```r
table(input_data$Attrition)
```

```
##
##   No  Yes
## 1233  237
```

```r
# SVM without oversampling

svmData = input_data

svmData$EmployeeCount = NULL #Every value is 1 so, we are dropping this variable
svmData$StandardHours = NULL #Every value is 8 so, we are dropping this variable
svmData$EmployeeNumber = NULL
svmData$Over18 = NULL


set.seed(1)
indexes = sample(1:nrow(svmData), size=0.8*nrow(svmData))
SVMtrain.Data = svmData[indexes,]
SVMtest.Data = svmData[-indexes,]
tuned = tune(svm,factor(Attrition)~.,data = SVMtrain.Data)
svm.model = svm(SVMtrain.Data$Attrition~., data=SVMtrain.Data
                 ,type="C-classification", gamma=tuned$best.model$gamma
                 ,cost=tuned$best.model$cost
                 ,kernel="radial")
svm.prd = predict(svm.model,newdata=SVMtest.Data)
confusionMatrix(svm.prd,SVMtest.Data$Attrition)
```

```
## Confusion Matrix and Statistics
##
##              Reference
```

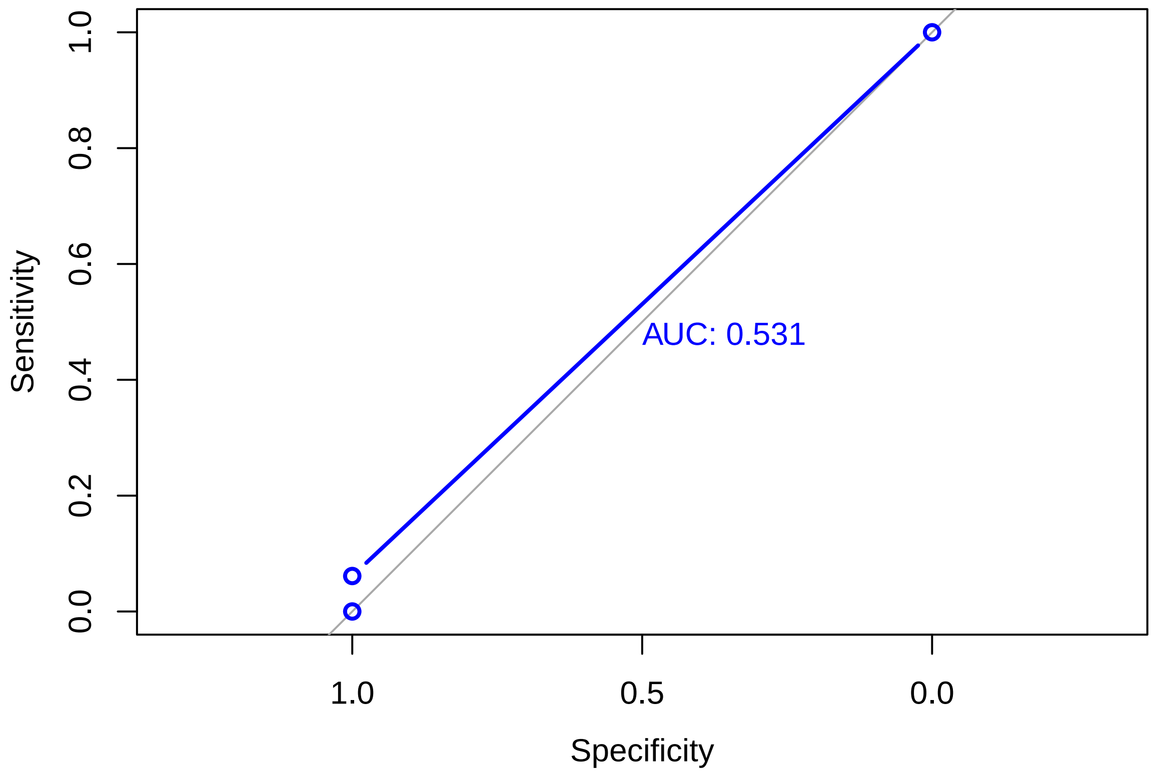
```
## Prediction  No Yes
##        No  245  46
##        Yes   0   3
##
##                Accuracy : 0.8435
##                  95% CI : (0.7969, 0.8831)
##     No Information Rate : 0.8333
##     P-Value [Acc > NIR] : 0.3534
##
##                   Kappa : 0.098
##
##  Mcnemar's Test P-Value : 3.247e-11
##
##             Sensitivity : 1.00000
##             Specificity : 0.06122
##          Pos Pred Value : 0.84192
##          Neg Pred Value : 1.00000
##              Prevalence : 0.83333
##          Detection Rate : 0.83333
##    Detection Prevalence : 0.98980
##       Balanced Accuracy : 0.53061
##
##        'Positive' Class : No
##
```

Looking at the above results at first glance, one might be tempted to say that this model could be a great fit. However, AUC value might have a different say on this. Let us check for that.

```
svm.plot = plot.roc (as.numeric(SVMtest.Data$Attrition), as.numeric(svm.prd),lwd=2, type="b", print.auc
```

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```



With AUC value of 0.531, this model seems to perform and hence one might easily conclude that accuracy should not be ultimate choice for model performance. It is quite evident to say as well that imbalance dataset has also played its part in this model.

So, we tried implementing the SVM after oversampling the data.

SVM with oversampling

```
svmData = data_over

svmData$EmployeeCount = NULL #Every value is 1 so, we are dropping this variable
svmData$StandardHours = NULL #Every value is 8 so, we are dropping this variable
svmData$EmployeeNumber = NULL
svmData$Over18 = NULL


set.seed(1)
indexes = sample(1:nrow(svmData), size=0.7*nrow(svmData))
SVMtrain.Data <- svmData[indexes,]
SVMtest.Data <- svmData[-indexes,]
tuned = tune(svm,factor(Attrition)~.,data = SVMtrain.Data)
svm.model <- svm(SVMtrain.Data$Attrition~., data=SVMtrain.Data
                 ,type="C-classification", gamma=tuned$best.model$gamma
                 ,cost=tuned$best.model$cost
```

```
                 ,kernel="radial")
svm.prd = predict(svm.model,newdata=SVMtest.Data)
confusionMatrix(svm.prd,SVMtest.Data$Attrition)
```

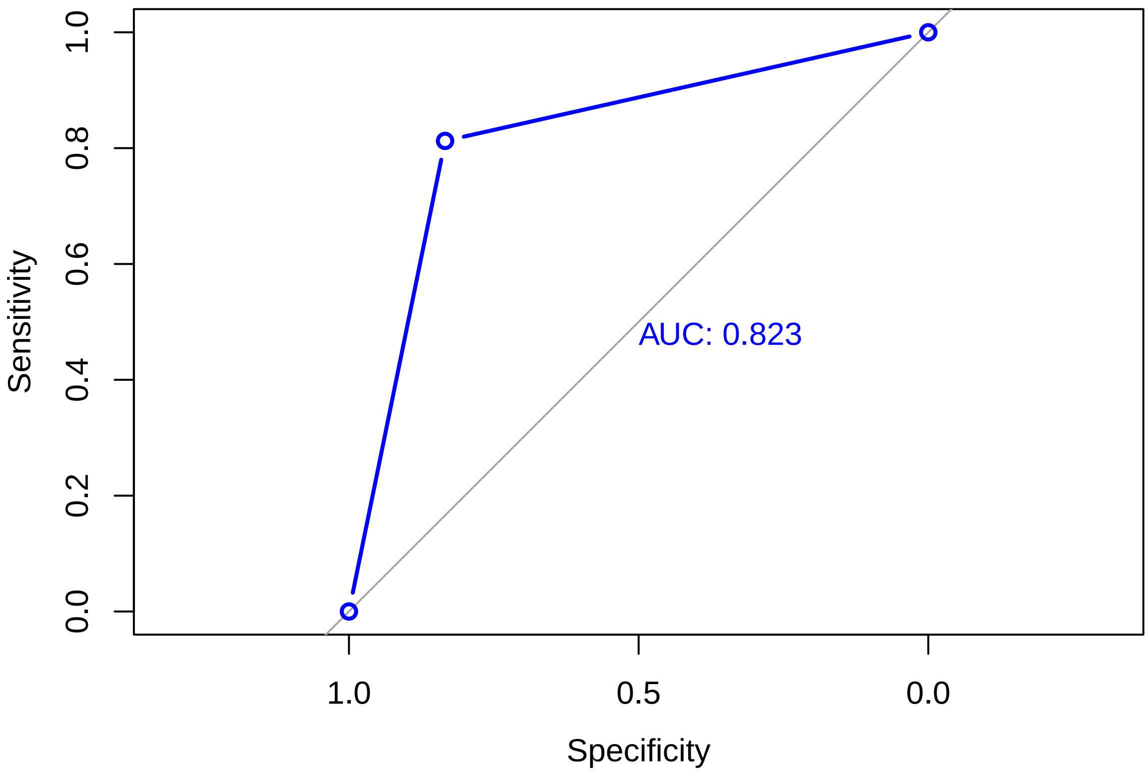
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 181  42
##          1  36 182
##
##                Accuracy : 0.8231
##                  95% CI : (0.7843, 0.8576)
##     No Information Rate : 0.5079
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6463
##
##  Mcnemar's Test P-Value : 0.5713
##
##             Sensitivity : 0.8341
##             Specificity : 0.8125
##          Pos Pred Value : 0.8117
##          Neg Pred Value : 0.8349
##              Prevalence : 0.4921
##          Detection Rate : 0.4104
##    Detection Prevalence : 0.5057
##       Balanced Accuracy : 0.8233
##
##        'Positive' Class : 0
##
```

Accuarcy, Sensitivity and Specificity looks good. Lets double check the AUC value for full confirmation.

```
svm.plot = plot.roc (as.numeric(SVMtest.Data$Attrition), as.numeric(svm.prd),lwd=2, type="b", print.auc
```

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```



With an AUC value of 0.805 the model looks a good fit for the dataset.

With this model we were specifically trying to convey the impact of the imbalanced dataset and how different performance measures can be used in determining the model performance. It can clearly seen that accuracy may be false indication of model performance.

After implementing the various models and EDA Analysis, we are in a better state to answer below research questions.

1) What proportion of the staff is leaving and where is it occurring?

As seen through analysis, rate of attrition is comparatively low in companies, though there are departments like Tech and Sales where Atrrition is relatively higher.

2) How does age affect attrition?

Age does affect attrition.

3) What other factors else contribute to the attrition?

Factors of Overtime, Monthly Income, Years at Company also affect attrition.

4)How well can we predict future attritions?

Based on various models, we are able to gain a prediction rate of 80 - 91 %, which is pretty satisfying. Out of all the models that we have tested, Random Forest has shown the best performance on the testing dataset and hence should be an ideal choice.

5)How can the organization reduce the rate of attrition inside the company?

It is slightly difficult to predict, looking in different department levels, but to generalize Monthly Income, Years at Company and Job satisfaction are few good factors to keep in check in order to reduce Attrition.

Lastly, there are many solutions to this problem in the kaggle, we tried differentiating from these published solution in terms by applying more models and keeping in check the imbalance dataset issue. We also looked beyond the accuracy as the sole parameter for model performance.

As far as our collaboration goes, We started with the initial assessment of the dataset, exploring together and finding the features that could affect our response variable. All three of us started exploring the data and its features to understand the relationships between the variables. After that, we had a discussion upon our findings and understood what all variables were completely irrelevant for our analysis and decided to exclude them.

Through EDA, we together looked onto the important aspects of various predictors. After the completion of EDA, we divided the modelling techniques to be implemented among each other. **Lalita** worked on Logistic Regression and Linear Discriminant Analysis, **Mihir** worked on Decision Trees and Random Forests and **Rajat** worked on SVM with and without sampling.

All 3 of us worked on comparison of models and understanding the best working model and finally collaborating the results along with the visualizations in our final report.

Finally, we observed that this project could be very useful for any organization in getting insights about the factors that might lead to Employee Attrition. In future we would like to create dashboards giving more user-friendly insights of factors affecting attrition. We would also like to test our prediction rate on some other real company dataset to see if we get similar results.