Employee Attrition Prediction

Nkonzo Sithole

2021-07-27

1. **Introduction**

Employees are the most important aset to organisations, hiring and retention of top talent is an extremely challenging task that requires capital, time and skills.

Employee attrition generally has negatively impact to many companies. Companies must have an HR strategy about hiring and retention, I have personally observer that many companies have internal surveys to check where they can improve to assist to prepare or avoid for such loss.

For example, studies found that staff churn is correlated with both demographic information as well as behavioral activities, satisfaction, etc.

I will be looking to predictors that must be taken into consideration by companies. Machine learning models or techniques can give better prediction on employee attrition, as by nature they mathematically model the correlation between factors and attrition outcome and maximize

In this study, (https://towardsdatascience.com/employee-retention-using-machine-learning-e7193e84bec4), they were looking at the cause of such leaving. I will use the data from https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

libraries

```
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("ggplot2")
library("RColorBrewer")
library("plotrix")
library("forcats")
library("ggplot2")
library("caret")
```

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

```
library("corrplot")
```

```
## corrplot 0.90 loaded
```

```
library("corrgram")
```

```
##
## Attaching package: ’corrgram’

## The following object is masked from ’package:lattice’:
##
##     panel.fill
```

```
library("gridExtra")
```

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

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

```
library("grid")
```

2. **Data exploration**

The experiments will be conducted on a data set of employees. The data set is publicly available and can be found at https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.

```
data <- read.csv("WA_Fn-UseC_-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
```

Summary

```
summary(data)
```

```
##      ï..Age       Attrition          BusinessTravel       DailyRate
##  Min.   :18.00   Length:1470        Length:1470        Min.   : 102.0
##  1st Qu.:30.00   Class :character   Class :character   1st Qu.: 465.0
##  Median :36.00   Mode  :character   Mode  :character   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     EducationField
##  Length:1470        Min.   : 1.000   Min.   :1.000   Length:1470
##  Class :character   1st Qu.: 2.000   1st Qu.:2.000   Class :character
##  Mode  :character   Median : 7.000   Median :3.000   Mode  :character
##                     Mean   : 9.193   Mean   :2.913
##                     3rd Qu.:14.000   3rd Qu.:4.000
##                     Max.   :29.000   Max.   :5.000
##  EmployeeCount EmployeeNumber   EnvironmentSatisfaction    Gender
##  Min.   :1     Min.   :   1.0   Min.   :1.000           Length:1470
##  1st Qu.:1     1st Qu.: 491.2   1st Qu.:2.000           Class :character
##  Median :1     Median :1020.5   Median :3.000           Mode  :character
##  Mean   :1     Mean   :1024.9   Mean   :2.722
##  3rd Qu.:1     3rd Qu.:1555.8   3rd Qu.:4.000
##  Max.   :1     Max.   :2068.0   Max.   :4.000
##    HourlyRate     JobInvolvement    JobLevel       JobRole
##  Min.   : 30.00   Min.   :1.00   Min.   :1.000   Length:1470
##  1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000   Class :character
##  Median : 66.00   Median :3.00   Median :2.000   Mode  :character
##  Mean   : 65.89   Mean   :2.73   Mean   :2.064
##  3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000
##  Max.   :100.00   Max.   :4.00   Max.   :5.000
##  JobSatisfaction MaritalStatus      MonthlyIncome    MonthlyRate
##  Min.   :1.000   Length:1470        Min.   : 1009   Min.   : 2094
##  1st Qu.:2.000   Class :character   1st Qu.: 2911   1st Qu.: 8047
##  Median :3.000   Mode  :character   Median : 4919   Median :14236
##  Mean   :2.729                      Mean   : 6503   Mean   :14313
##  3rd Qu.:4.000                      3rd Qu.: 8379   3rd Qu.:20462
##  Max.   :4.000                      Max.   :19999   Max.   :26999
##  NumCompaniesWorked    Over18            OverTime         PercentSalaryHike
##  Min.   :0.000      Length:1470        Length:1470        Min.   :11.00
##  1st Qu.:1.000      Class :character   Class :character   1st Qu.:12.00
##  Median :2.000      Mode  :character   Mode  :character   Median :14.00
##  Mean   :2.693                                            Mean   :15.21
##  3rd Qu.:4.000                                            3rd Qu.:18.00
##  Max.   :9.000                                            Max.   :25.00
##  PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Min.   :3.000     Min.   :1.000            Min.   :80    Min.   :0.0000
##  1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000
##  Median :3.000     Median :3.000            Median :80    Median :1.0000
##  Mean   :3.154     Mean   :2.712            Mean   :80    Mean   :0.7939
##  3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000
##  Max.   :4.000     Max.   :4.000            Max.   :80    Max.   :3.0000
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
##  Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
##  1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
##  Median :10.00     Median :3.000         Median :3.000   Median : 5.000
##  Mean   :11.28     Mean   :2.799         Mean   :2.761   Mean   : 7.008
##  3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000
##  Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
##  1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
##  Median : 3.000     Median : 1.000          Median : 3.000
##  Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
##  3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
##  Max.   :18.000     Max.   :15.000          Max.   :17.000
```

Fix Age column

```
colnames(data)[1] <- "Age"
```

Dataset is made up of the following rows and columns

```
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
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : chr  "Sales" "Research & Development" "Research & Devel
##  $ 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 Scie
##  $ 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
##  $ 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
##  $ 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 ...
```

```r
cat("Data Set has ",dim(data)[1], " Rows and ", dim(data)[2], " Columns" )
```

```
## Data Set has  1470  Rows and  35  Columns
```

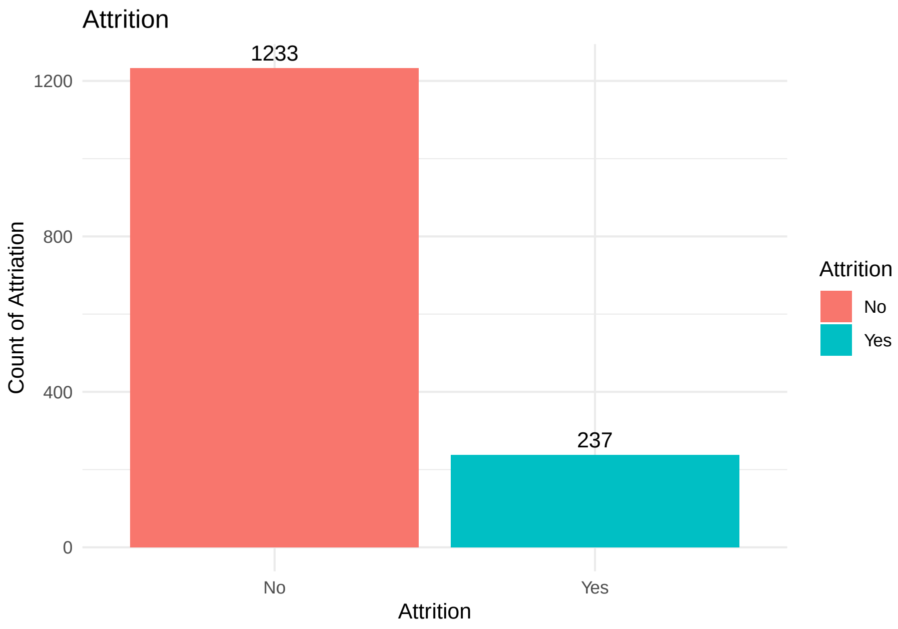
Checking missing and duplicate values

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

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

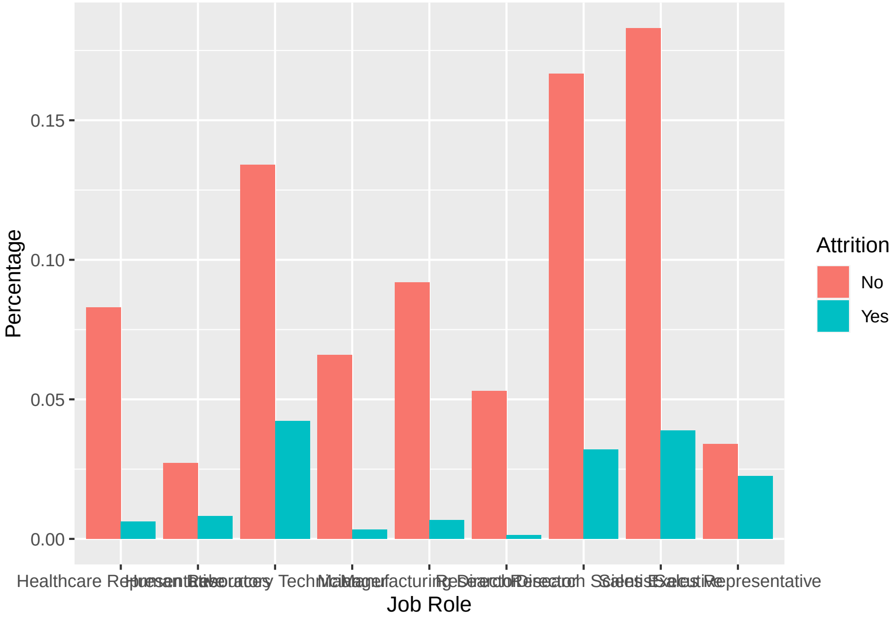
3. Data Visualization

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



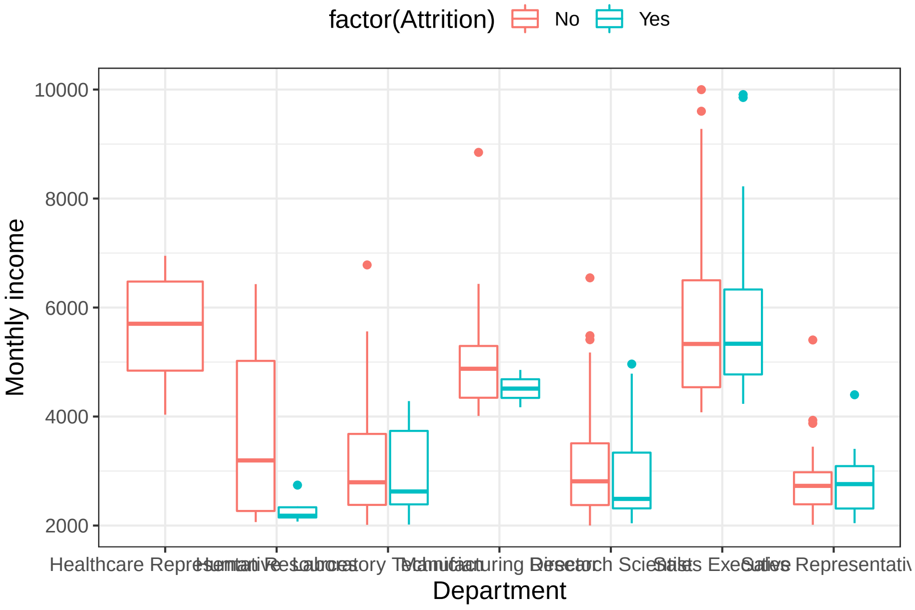
b) Checking employees status(attrition) per job title

```
ggplot(data, aes(JobRole, fill=Attrition)) +
  geom_bar(aes(y=(..count..)/sum(..count..)), position="dodge") +
  xlab("Job Role") +
  ylab("Percentage")
```



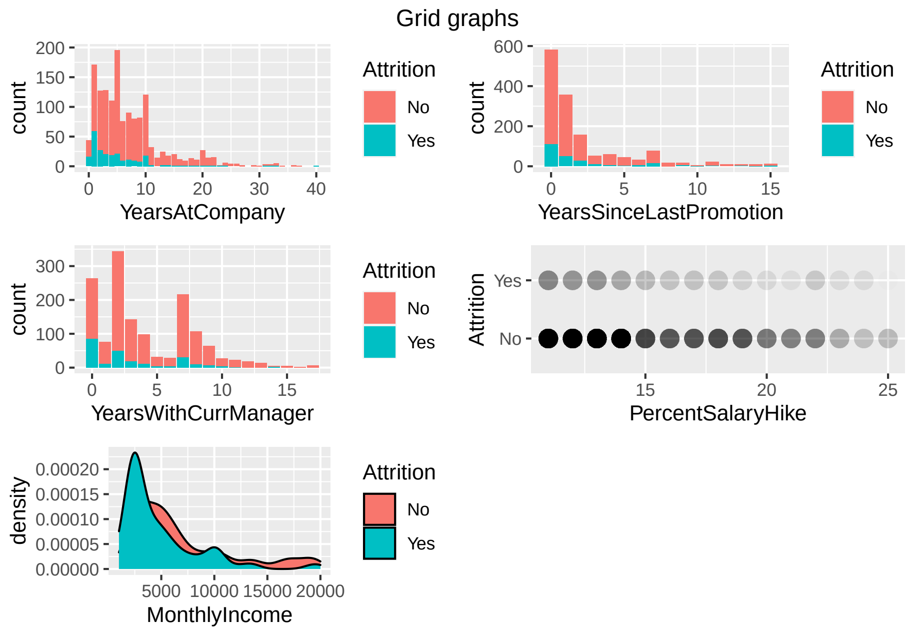
c) Income, jobRole, previous percentage salary hike and service years may affect decision for employees to leave.

```
ggplot(filter(data, (PercentSalaryHike >= 11) & (YearsAtCompany >= 2) & (YearsAtCompany
         aes(x=factor(JobRole), y=MonthlyIncome, color=factor(Attrition))) +
  geom_boxplot() +
  xlab("Department") +
  ylab("Monthly income") +
  scale_fill_discrete(guide=guide_legend(title="Attrition")) +
  theme_bw() +
  theme(text=element_text(size=13), legend.position="top")
```



d) Employees grid graph in relation with Years of service, Growth, Manager, Income and salary increase.

```
EmployeesYearOfService <- ggplot(data,aes(YearsAtCompany,fill = Attrition))+geom_bar()
EmployeesGrowth <- ggplot(data,aes(YearsSinceLastPromotion,fill = Attrition))+geom_bar(
EmployeesManager <- ggplot(data,aes(YearsWithCurrManager,fill = Attrition))+geom_bar()
EmployeeSalIncrease <- ggplot(data,aes(PercentSalaryHike,Attrition))+geom_point(size=4,
EmployeesIncome <- ggplot(data,aes(MonthlyIncome,fill=Attrition))+geom_density()
gr <- grid.arrange(EmployeesYearOfService,EmployeesGrowth,EmployeesManager,EmployeeSalI
```



```
gr
```

```
## TableGrob (4 x 2) "arrange": 6 grobs
##   z     cells    name               grob
## 1 1 (2-2,1-1) arrange     gtable[layout]
## 2 2 (2-2,2-2) arrange     gtable[layout]
## 3 3 (3-3,1-1) arrange     gtable[layout]
## 4 4 (3-3,2-2) arrange     gtable[layout]
## 5 5 (4-4,1-1) arrange     gtable[layout]
## 6 6 (1-1,1-2) arrange text[GRID.text.579]
```

data correlation

remove near zero variables

```
near_Zero_variables <- names(data[, nearZeroVar(data)]) %>% print()
```

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

```
data <- data %>% select(-one_of(near_Zero_variables))
```

corrgram(data, lower.panel = panel.shade, upper.panel = panel.pie, text.panel = panel.txt, main = "Corrgram of all numeric variables")

From this, I will use algorithms like rf or XGBoost to build a model that can predict in fact which employees are most likely to leave in the future

4. **Data Preparation and Partitioning**

convert certain integer variable to factor variable.

```
factor_variables <- c("Education", "EnvironmentSatisfaction", "JobInvolvement", "JobLev
data[, factor_variables] <- lapply((data[, factor_variables]), as.factor)
data <- data %>% mutate_if(is.character, as.factor)
str(data)
```

```
## 'data.frame':    1470 obs. of  32 variables:
##  $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..:
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : Factor w/ 5 levels "1","2","3","4",..: 2 1 2 4 1 2 3 1
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4
##  $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : Factor w/ 4 levels "1","2","3","4": 2 3 4 4 1 4 3 4 4 3
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 3 2 2 3 3 3 4 3 2 3
##  $ JobLevel                : Factor w/ 5 levels "1","2","3","4",..: 2 2 1 1 1 1 1 1
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 7
##  $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4 1 3 3 3
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787
##  $ NumCompaniesWorked      : Factor w/ 10 levels "0","1","2","3",..: 9 2 7 2 10 1 5
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : Factor w/ 2 levels "3","4": 1 2 1 1 1 1 2 2 2 1 ...
##  $ RelationshipSatisfaction: Factor w/ 4 levels "1","2","3","4": 1 4 2 3 4 3 1 2 2 2
##  $ StockOptionLevel        : Factor w/ 4 levels "0","1","2","3": 1 2 1 1 2 1 4 2 1 3
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : 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 ...
```

Before modeling, first I use `set.seed(1)` and partition my data into train and test sets, which will be used to model and produce predictions. Then towards the end of this report, I will show the final model performance on the validation set.

```
set.seed(1)
train_index <- createDataPartition(data$Attrition , times =1, p = 0.7, list = FALSE)
train <- data[train_index,]
test <- data[-train_index,]
```

5. **Modeling, Tuning & Evaluation**

##training control to tune

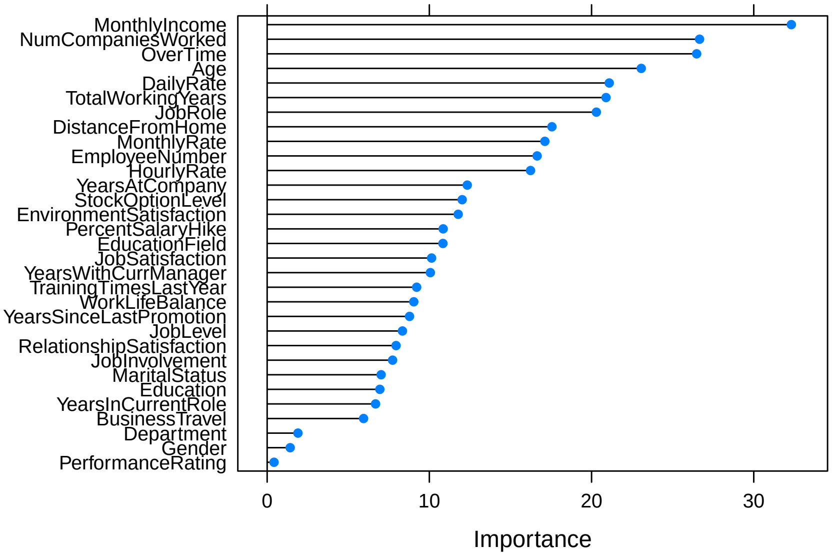
```
##random forest model
control <- trainControl(method="repeatedcv", number=3, repeats=1)
random_forest_model <- train(dplyr::select(data, -Attrition),
                             data$Attrition,
                             data=train,
                             method="rf",
                             preProcess="scale",
                             trControl=control)

prediction_rfm <- predict(random_forest_model, newdata=select(test, -Attrition))
confusionMatrix(prediction_rfm,reference=test$Attrition,positive="Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  369   0
##        Yes   0  71
##
##                Accuracy : 1
##                  95% CI : (0.9917, 1)
##     No Information Rate : 0.8386
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar’s Test P-Value : NA
##
##             Sensitivity : 1.0000
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##              Prevalence : 0.1614
##          Detection Rate : 0.1614
##    Detection Prevalence : 0.1614
##       Balanced Accuracy : 1.0000
##
##        'Positive' Class : Yes
##
```

```
imp <- varImp(random_forest_model, scale=FALSE)
```

6. **Conclusion** We can see that Salary has big impact in employees attrition

```
plot(imp)
```


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