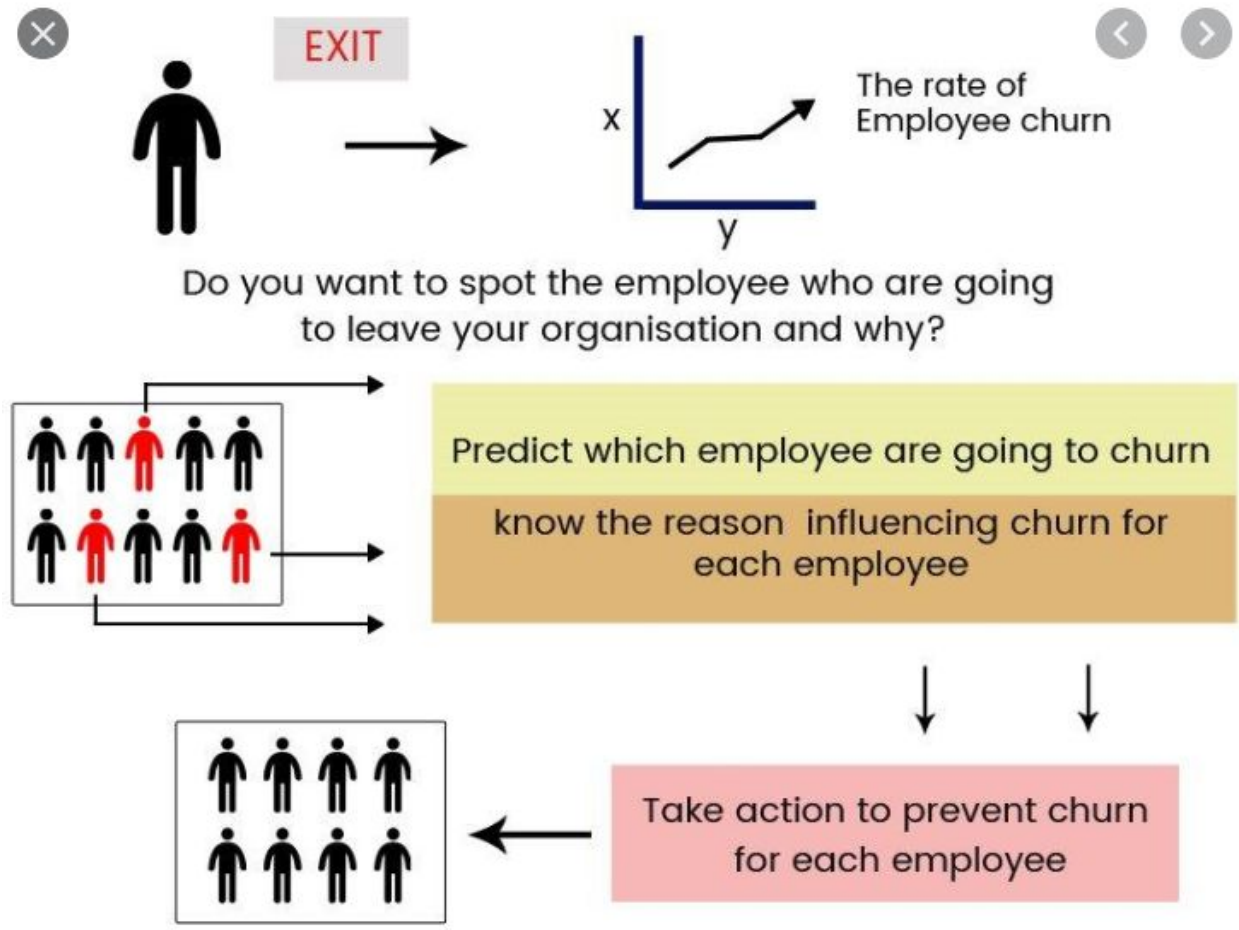
HR Analytics- Prediction of Employee attrition

Report by Soumonos Mukherjee

Dataset- IBM HR analytics Dataset on Employee Attrition



Introduction

The Dataset is collected from Kaggle. It's created by Data Scientists at IBM for HR analytics practice and purpose. The Dataset has a total of 34 predictor variables and 1 Target variable. We have described the Dataset here below.

Objective and Goal:

- Detailed Exploratory analysis of the Data.
- Data Cleaning and preprocessing.
- Statistical significance tests
- Predictive modelling with different algorithms
- Optimizing the algorithms by tuning parameters and hyperparameters to improve the results

Challenges:

There are certain Dataset specific challenges while dealing with this.

1. There are some redundant columns which substantially contribute nothing to the classification model . Rather they mislead the model. They need to be identified and deleted.
2. Results received from Statistical tests are often inconclusive. In predictive analytics, we can build hypotheses and test the same but they can not be highly regarded.
3. The Target variable (Attrition) is not very balanced. We need to pay attention to evaluation metrics other than accuracy in this kind of specific issue. As the model may have the tendency to just predict a single class all the time. That will yield a good accuracy but it will not be a good model.

Mitigations:

1. Columns with single class (all values are assigned to just 1 class) will be deleted.
2. The apparent numerical columns may be signifying categorical behavior, they will be converted to categorical (factor in R).
3. Statistical tests will be performed for better understanding of Data but they will not be strictly followed as Feature selectors unless they have conclusive implications.
4. We will consider metrics like Confusion Matrix and Cohen's Kappa to know how better our model is than a random predictor. We shall find a balance in the trade-off between the Accuracy and Kappa values.

Formulas:

Accuracy= mean(Predicted_values == Actual_values)

Kappa= 1- (1-P(predicted)/1-P(actual))

Data Cleaning and Preprocessing (With R&R-Studio):

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------- tidyver
se 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------ tidyverse_con
flicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(psych)
```

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

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

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

```
library(ROCR)
library(ggplot2)
```

```
df <- read.csv("HR-Attrition.csv")
summary(df)
```

```
##      i..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
```

Converting the categorical variables (described formerly as numerical) into Factors and dropping redundant columns:

```
names<- c("Education","Attrition","EnvironmentSatisfaction","JobInvolvement","JobLevel","JobSatisfaction","NumCom
paniesWorked","PerformanceRating",
  "RelationshipSatisfaction","StockOptionLevel","TrainingTimesLastYear","WorkLifeBalance")
df[,names]<- lapply(df[,names],factor)
summary(df)
```

```
##      ï..Age      Attrition  BusinessTravel       DailyRate
##  Min.   :18.00   No :1233   Length:1470        Min.   : 102.0
##  1st Qu.:30.00   Yes: 237   Class :character   1st Qu.: 465.0
##  Median :36.00              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     EmployeeCount
##  Length:1470        Min.   : 1.000   1:170     Length:1470        Min.   :1
##  Class :character   1st Qu.: 2.000   2:282     Class :character   1st Qu.:1
##  Mode  :character   Median : 7.000   3:572     Mode  :character   Median :1
##                     Mean   : 9.193   4:398                        Mean   :1
##                     3rd Qu.:14.000   5: 48                        3rd Qu.:1
##                     Max.   :29.000                                Max.   :1
##
##  EmployeeNumber   EnvironmentSatisfaction    Gender            HourlyRate
##  Min.   :   1.0   1:284                   Length:1470        Min.   : 30.00
##  1st Qu.: 491.2   2:287                   Class :character   1st Qu.: 48.00
##  Median :1020.5   3:453                   Mode  :character   Median : 66.00
##  Mean   :1024.9   4:446                                      Mean   : 65.89
##  3rd Qu.:1555.8                                              3rd Qu.: 83.75
##  Max.   :2068.0                                              Max.   :100.00
##
##  JobInvolvement JobLevel   JobRole          JobSatisfaction MaritalStatus
##  1: 83          1:543    Length:1470        1:289           Length:1470
##  2:375          2:534    Class :character   2:280           Class :character
##  3:868          3:218    Mode  :character   3:442           Mode  :character
##  4:144          4:106                       4:459
##                 5: 69
##
##
##  MonthlyIncome    MonthlyRate    NumCompaniesWorked    Over18
##  Min.   : 1009   Min.   : 2094   1      :521        Length:1470
##  1st Qu.: 2911   1st Qu.: 8047   0      :197        Class :character
##  Median : 4919   Median :14236   3      :159        Mode  :character
##  Mean   : 6503   Mean   :14313   2      :146
##  3rd Qu.: 8379   3rd Qu.:20462   4      :139
##  Max.   :19999   Max.   :26999   7      : 74
##                                  (Other):234
##    OverTime         PercentSalaryHike PerformanceRating
##  Length:1470        Min.   :11.00     3:1244
##  Class :character   1st Qu.:12.00     4: 226
##  Mode  :character   Median :14.00
##                     Mean   :15.21
##                     3rd Qu.:18.00
##                     Max.   :25.00
##
```

```
##
##  TrainingTimesLastYear WorkLifeBalance YearsAtCompany   YearsInCurrentRole
##  0: 54                 1: 80           Min.   : 0.000   Min.   : 0.000
##  1: 71                 2:344           1st Qu.: 3.000   1st Qu.: 2.000
##  2:547                 3:893           Median : 5.000   Median : 3.000
##  3:491                 4:153           Mean   : 7.008   Mean   : 4.229
##  4:123                                 3rd Qu.: 9.000   3rd Qu.: 7.000
##  5:119                                 Max.   :40.000   Max.   :18.000
##  6: 65
##  YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000          Min.   : 0.000
##  1st Qu.: 0.000          1st Qu.: 2.000
##  Median : 1.000          Median : 3.000
##  Mean   : 2.188          Mean   : 4.123
##  3rd Qu.: 3.000          3rd Qu.: 7.000
##  Max.   :15.000          Max.   :17.000
##
```

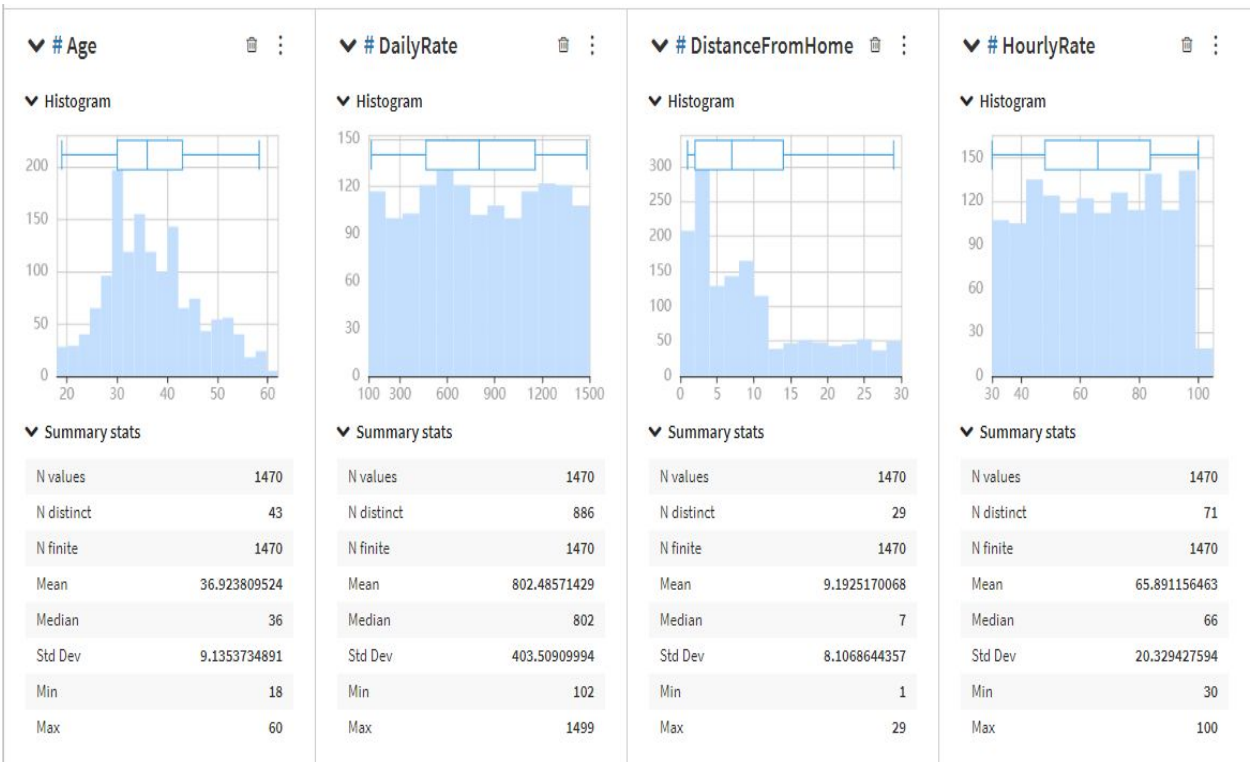
```
dfn<- select(df,-c("StandardHours","Over18","EmployeeNumber","EmployeeCount"))
```

Exploratory Analysis (With DataIku DS Studio)

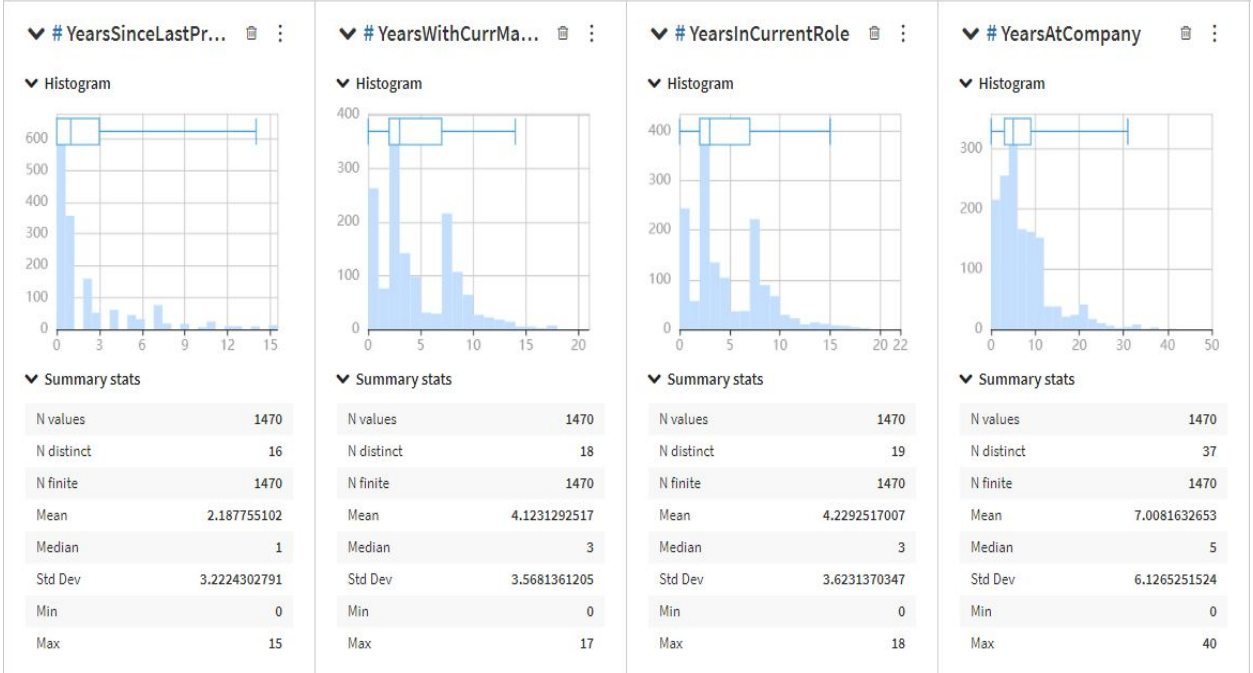
Univariate Analysis of Continuous Variables:

Charts and Metrics: **Box plots, Histograms, Summary stats (Distribution Data)**

Variables: **Age, Daily Rate, Distance from Home, Hourly Rate**



Variables: **Years since last promotion, Years with current manager, Years in current Role, Years at Company**



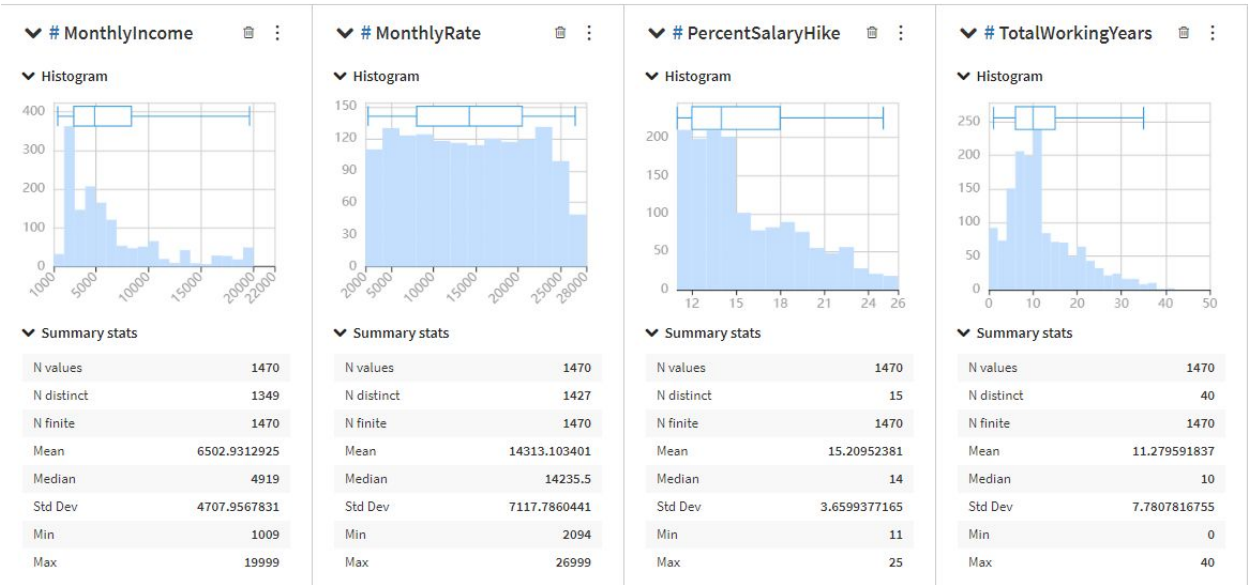
# YearsSinceLastPr...	
N values	1470
N distinct	16
N finite	1470
Mean	2.187755102
Median	1
Std Dev	3.2224302791
Min	0
Max	15

# YearsWithCurrMa...	
N values	1470
N distinct	18
N finite	1470
Mean	4.1231292517
Median	3
Std Dev	3.5681361205
Min	0
Max	17

# YearsInCurrentRole	
N values	1470
N distinct	19
N finite	1470
Mean	4.2292517007
Median	3
Std Dev	3.6231370347
Min	0
Max	18

# YearsAtCompany	
N values	1470
N distinct	37
N finite	1470
Mean	7.0081632653
Median	5
Std Dev	6.1265251524
Min	0
Max	40

Variables: **Monthly Income, Monthly Rate, Percentage of Salary Hike, Total Working years**



# MonthlyIncome	
N values	1470
N distinct	1349
N finite	1470
Mean	6502.9312925
Median	4919
Std Dev	4707.9567831
Min	1009
Max	19999

# MonthlyRate	
N values	1470
N distinct	1427
N finite	1470
Mean	14313.103401
Median	14235.5
Std Dev	7117.7860441
Min	2094
Max	26999

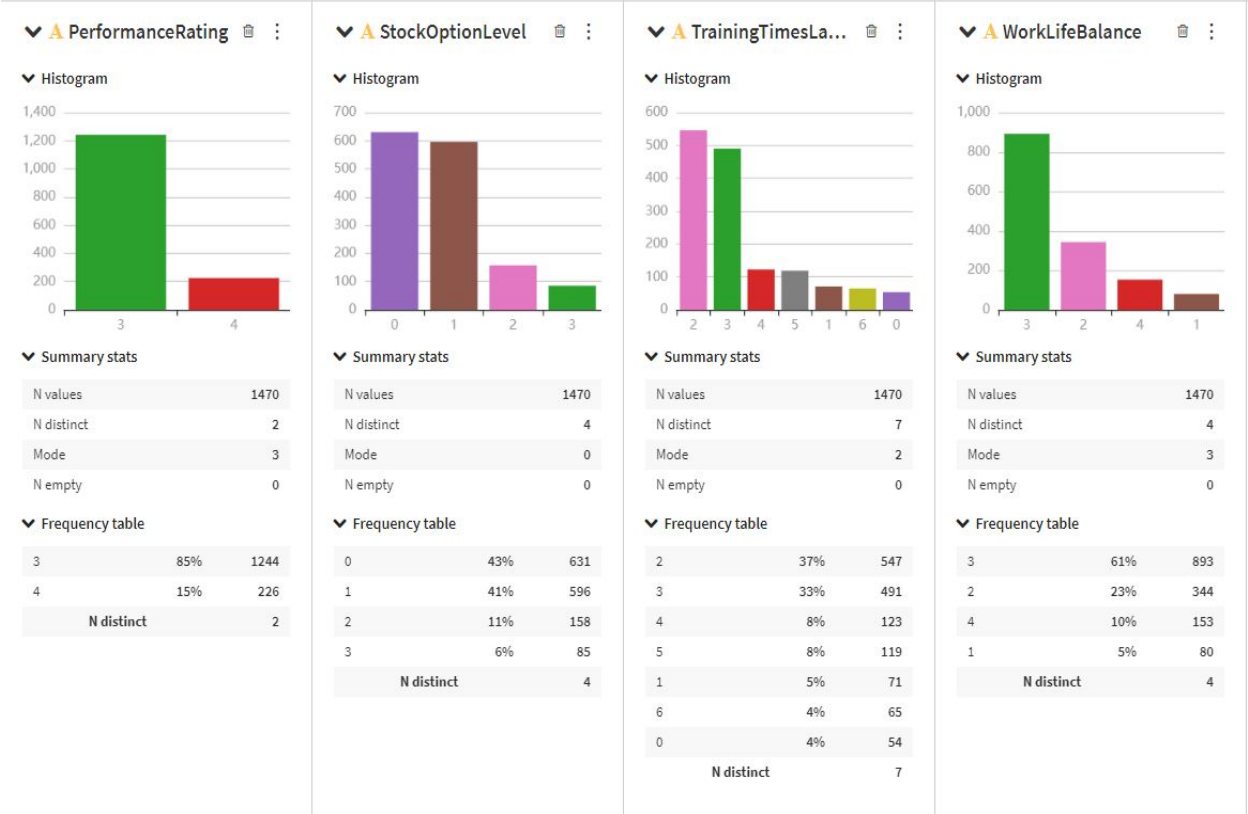
# PercentSalaryHike	
N values	1470
N distinct	15
N finite	1470
Mean	15.20952381
Median	14
Std Dev	3.6599377165
Min	11
Max	25

# TotalWorkingYears	
N values	1470
N distinct	40
N finite	1470
Mean	11.279591837
Median	10
Std Dev	7.7807816755
Min	0
Max	40

Univariate Analysis of Categorical Variables:

Metrics and Charts: **Bar Charts , Summary stats, Frequency Table (Value- Counts)**

Variables: **Performance Rating, Stock option level, Training time last year, Work-Life Balance**



PerformanceRating

Histogram

Summary stats

N values	1470
N distinct	2
Mode	3
N empty	0

Frequency table

3	85%	1244
4	15%	226
N distinct		2

StockOptionLevel

Histogram

Summary stats

N values	1470
N distinct	4
Mode	0
N empty	0

Frequency table

0	43%	631
1	41%	596
2	11%	158
3	6%	85
N distinct		4

TrainingTimesLa...

Histogram

Summary stats

N values	1470
N distinct	7
Mode	2
N empty	0

Frequency table

2	37%	547
3	33%	491
4	8%	123
5	8%	119
1	5%	71
6	4%	65
0	4%	54
N distinct		7

WorkLifeBalance

Histogram

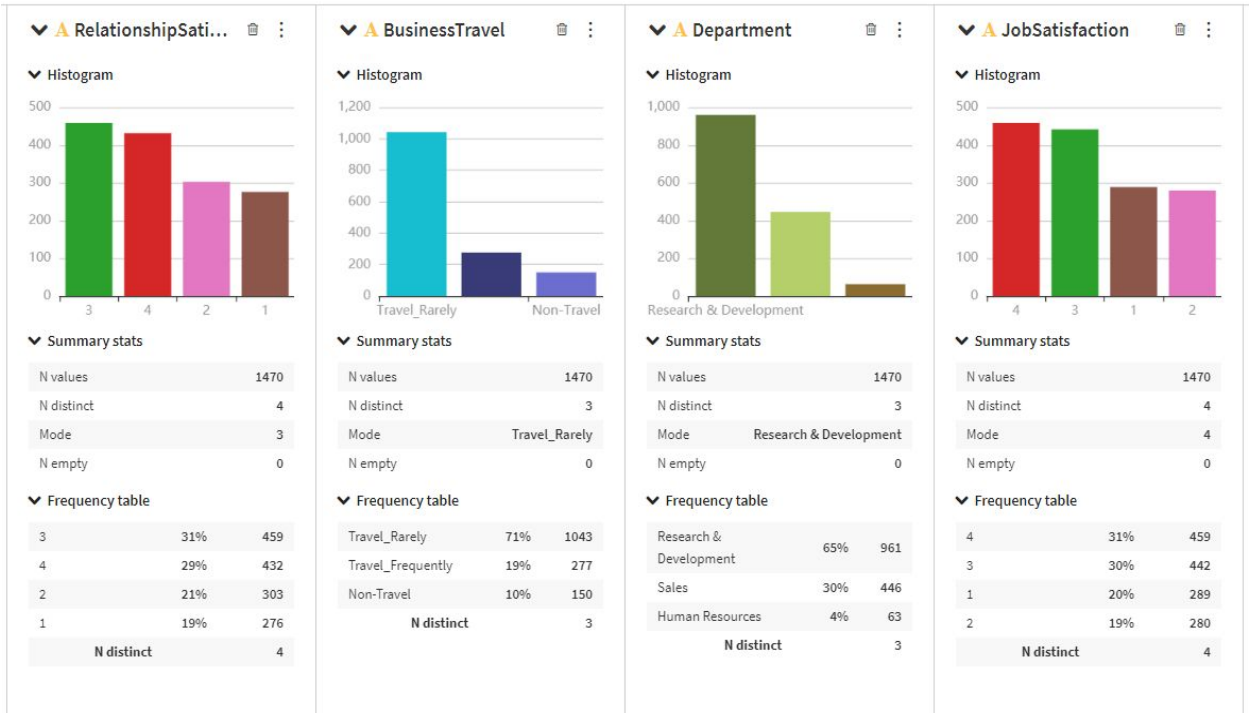
Summary stats

N values	1470
N distinct	4
Mode	3
N empty	0

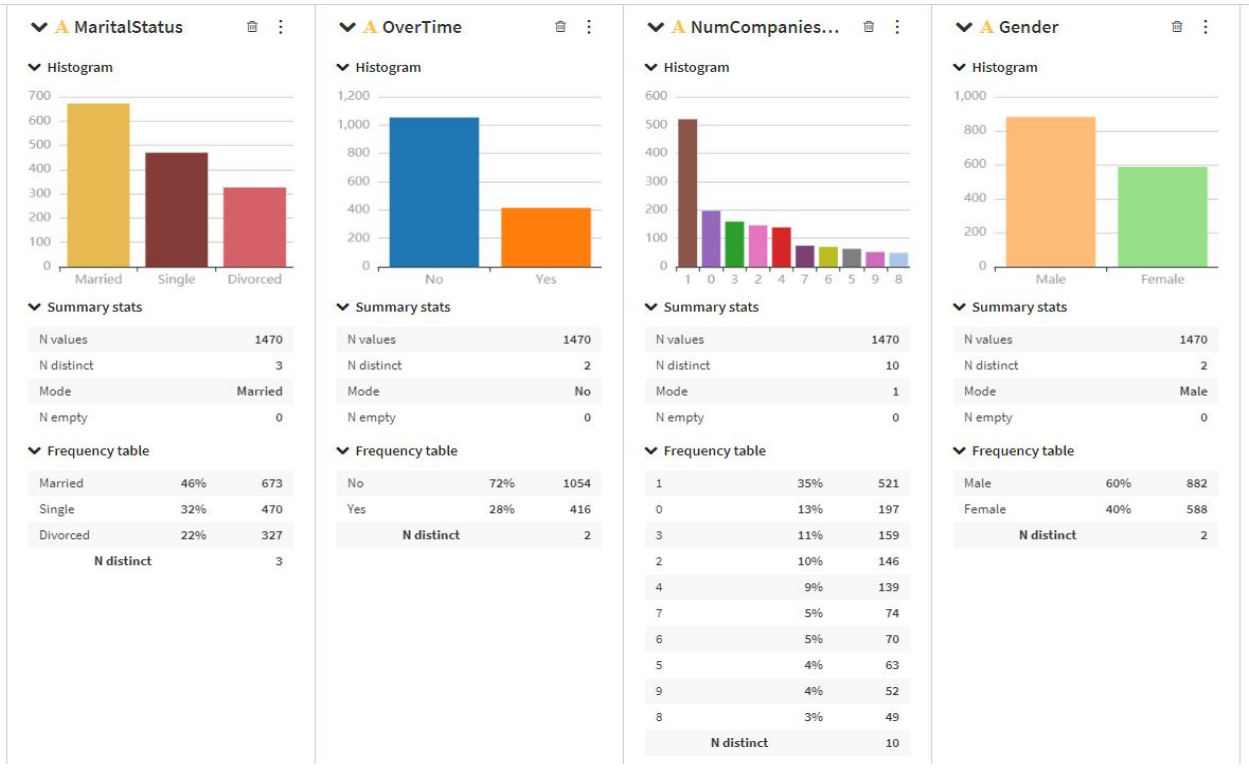
Frequency table

3	61%	893
2	23%	344
4	10%	153
1	5%	80
N distinct		4

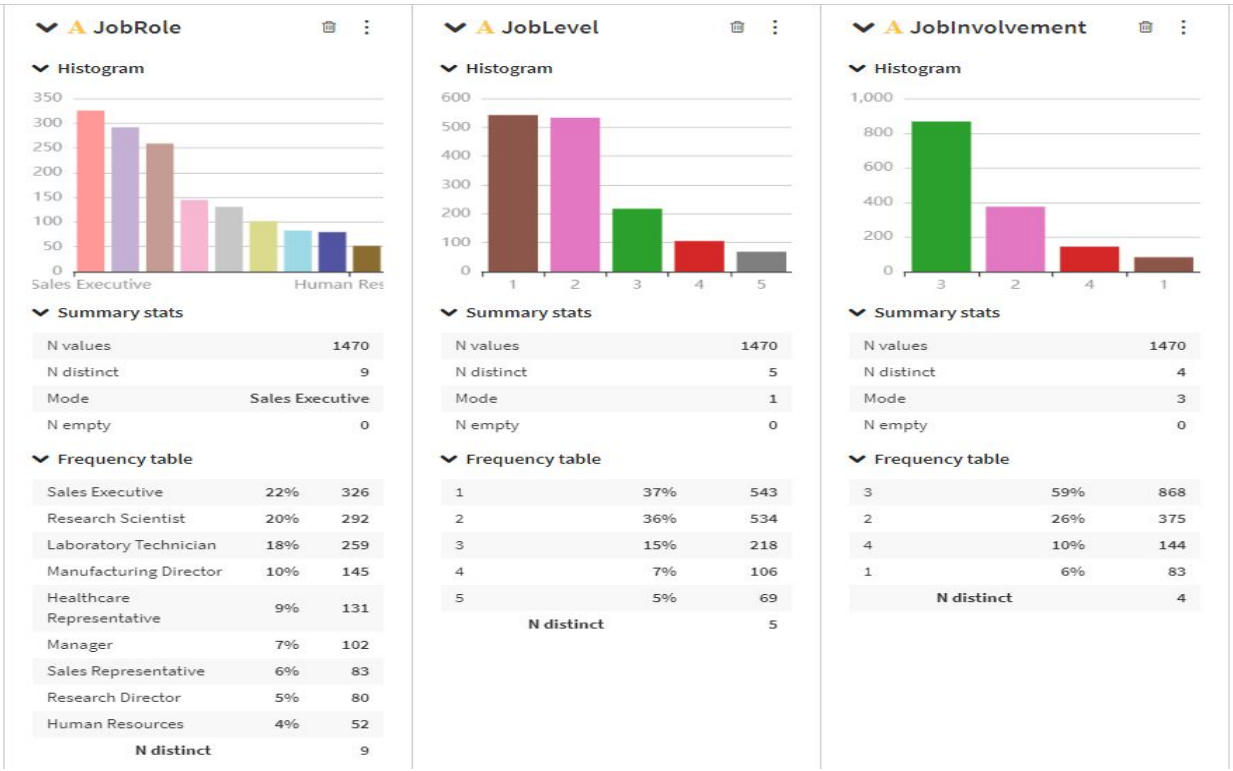
Variables: Relationship satisfaction, Business Travel, Department, Job Satisfaction



Variables: Marital Status, Overtime, Number of companies worked in, Gender



Variables: **Job Role, Job Level, Job Involvement**



JobRole

Histogram

Summary stats

N values	1470
N distinct	9
Mode	Sales Executive
N empty	0

Frequency table

Sales Executive	22%	326
Research Scientist	20%	292
Laboratory Technician	18%	259
Manufacturing Director	10%	145
Healthcare Representative	9%	131
Manager	7%	102
Sales Representative	6%	83
Research Director	5%	80
Human Resources	4%	52
N distinct		9

JobLevel

Histogram

Summary stats

N values	1470
N distinct	5
Mode	1
N empty	0

Frequency table

1	37%	543
2	36%	534
3	15%	218
4	7%	106
5	5%	69
N distinct		5

JobInvolvement

Histogram

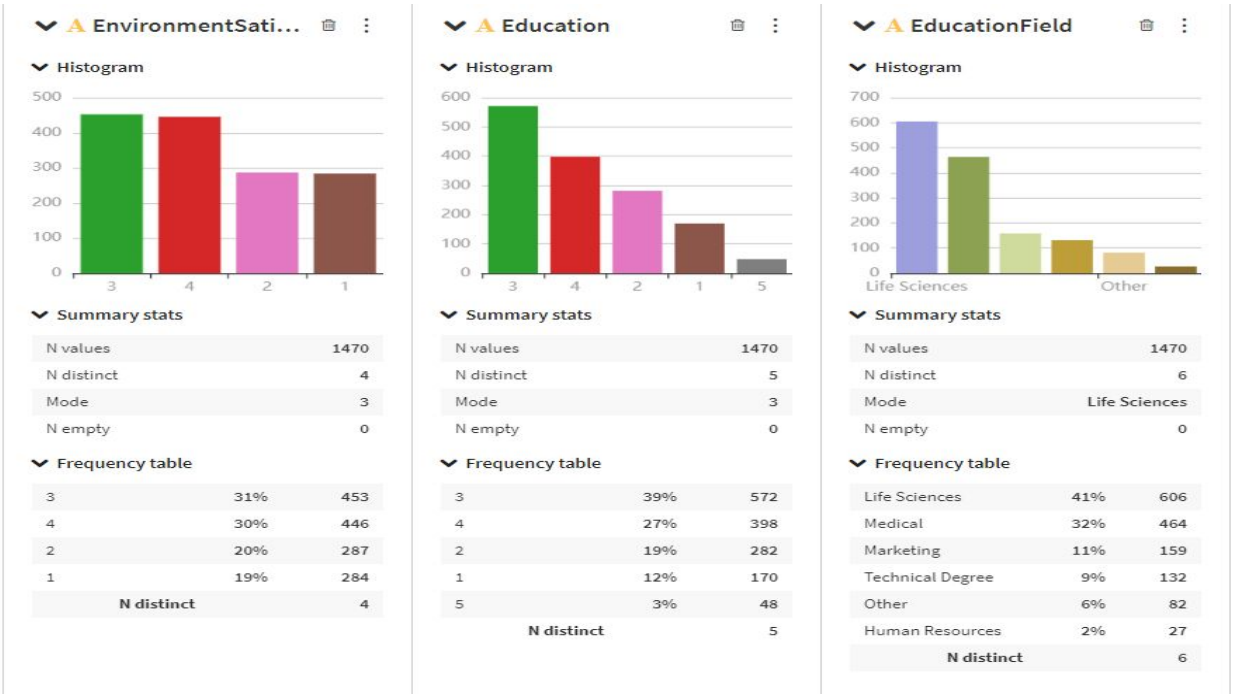
Summary stats

N values	1470
N distinct	4
Mode	3
N empty	0

Frequency table

3	59%	868
2	26%	375
4	10%	144
1	6%	83
N distinct		4

Variables: **Marital Status, Overtime, Number of companies worked in, Gender**



EnvironmentSati...

Histogram

Summary stats

N values	1470
N distinct	4
Mode	3
N empty	0

Frequency table

3	31%	453
4	30%	446
2	20%	287
1	19%	284
N distinct		4

Education

Histogram

Summary stats

N values	1470
N distinct	5
Mode	3
N empty	0

Frequency table

3	39%	572
4	27%	398
2	19%	282
1	12%	170
5	3%	48
N distinct		5

EducationField

Histogram

Summary stats

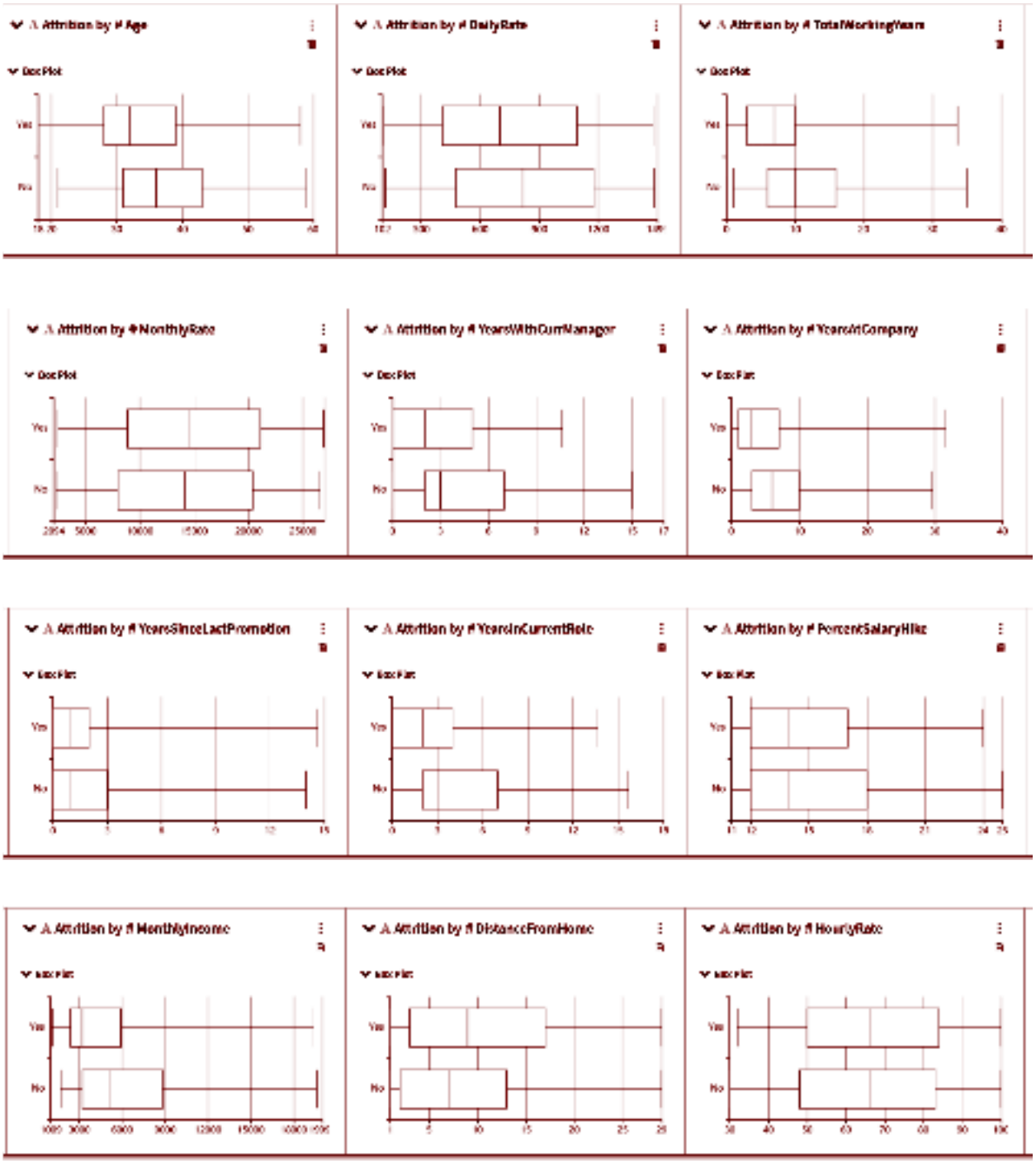
N values	1470
N distinct	6
Mode	Life Sciences
N empty	0

Frequency table

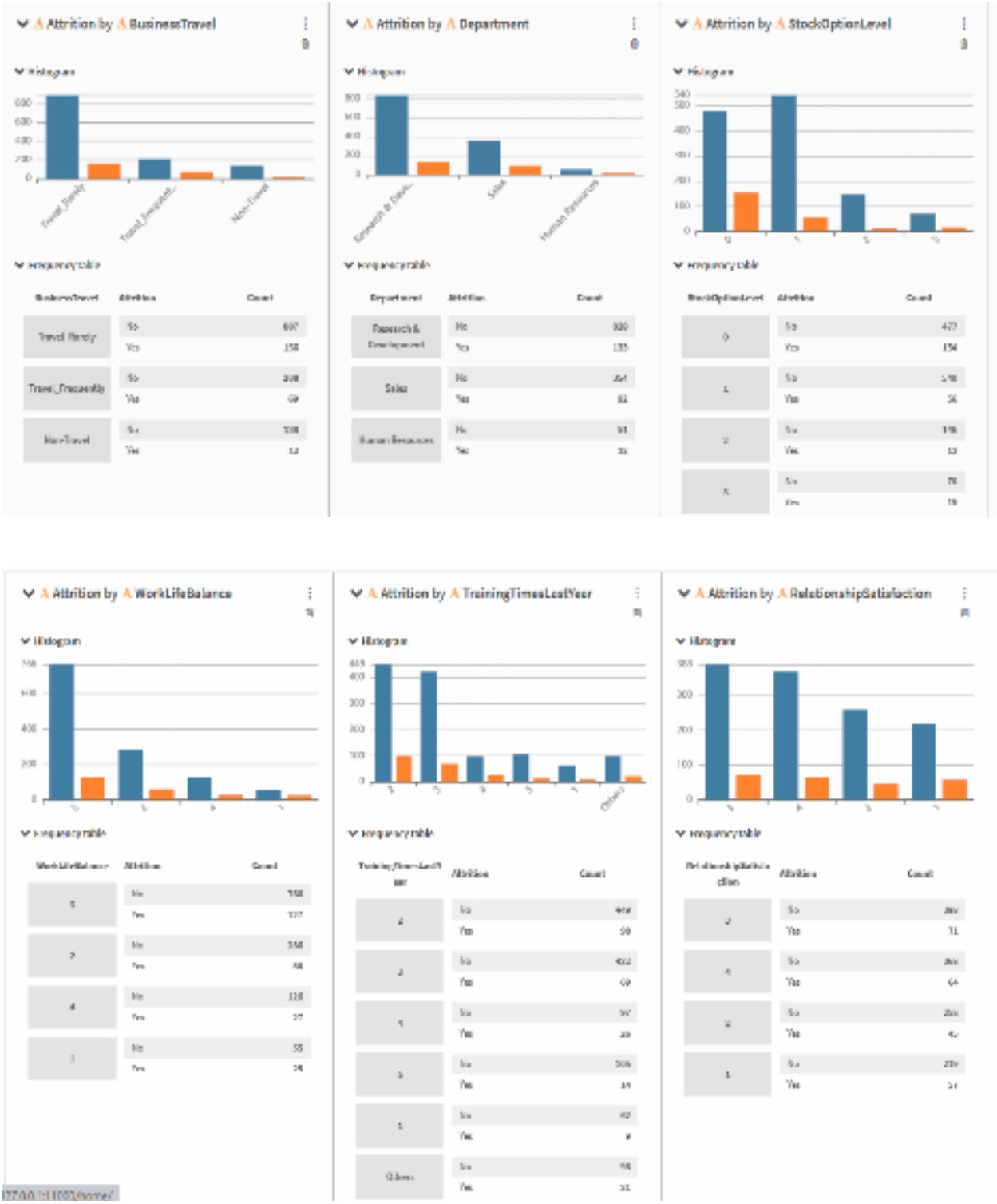
Life Sciences	41%	606
Medical	32%	464
Marketing	11%	159
Technical Degree	9%	132
Other	6%	82
Human Resources	2%	27
N distinct		6

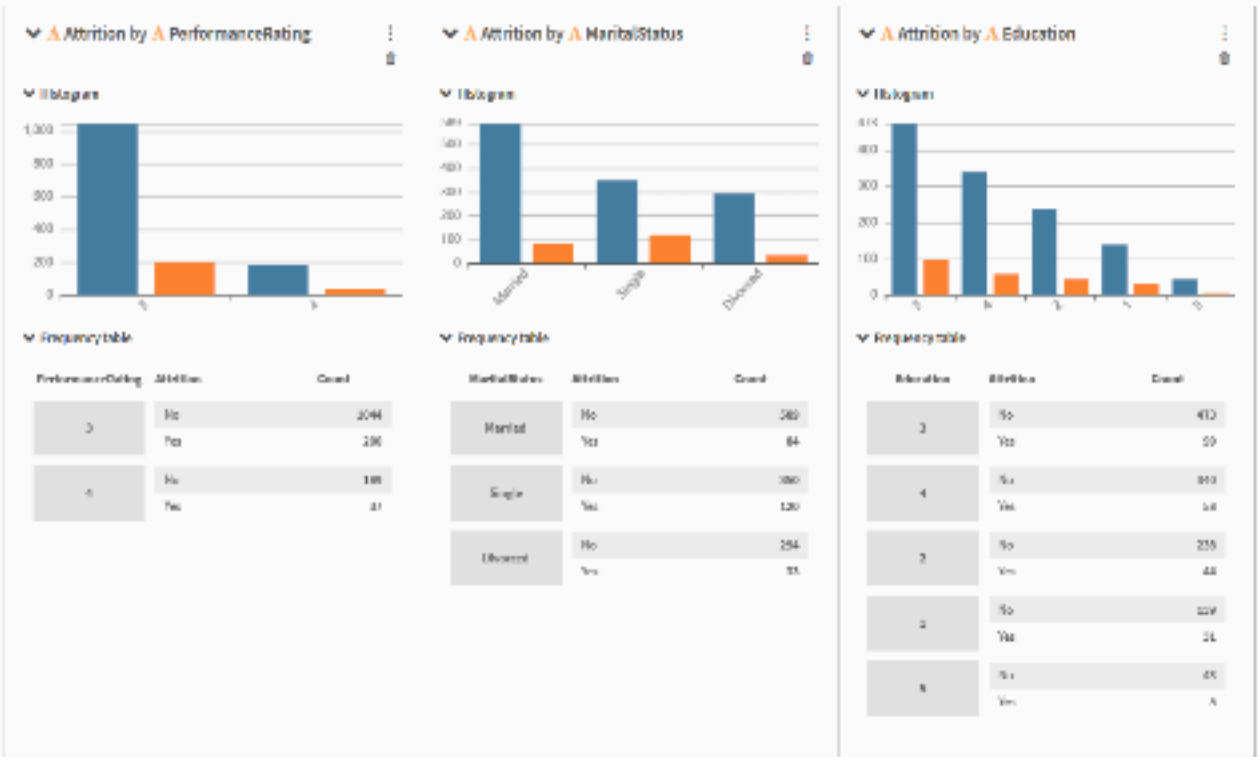
Bivariate Analysis (Target [Categorical] vs Continuous predictors):

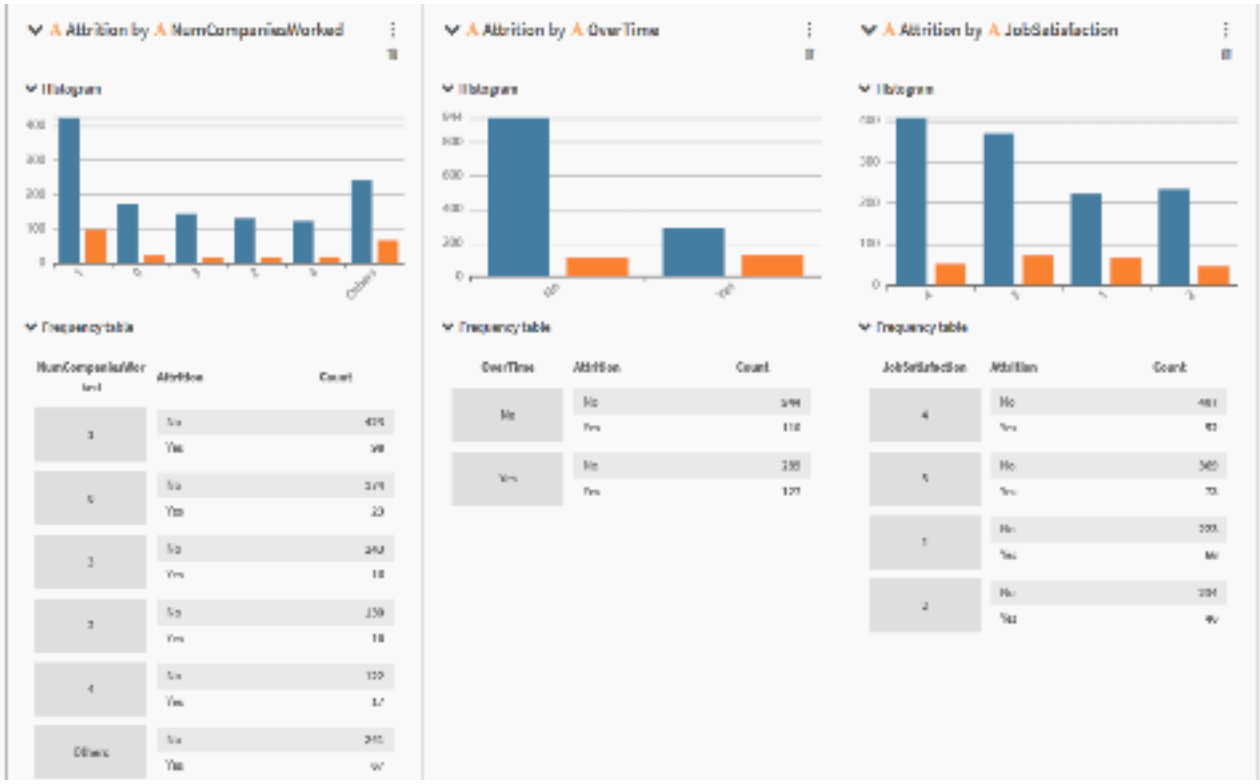
Charts and Metrics: Box Plots



Bivariate Analysis (Target [Categorical] vs Categorical predictors):




Attrition by PerformanceRating
Histogram
Frequency table
Attrition by MaritalStatus
Histogram
Frequency table
Attrition by Education
Histogram
Frequency table


Attrition by NumCompaniesWorked
Histogram
Frequency table
Attrition by OverTime
Histogram
Frequency table
Attrition by JobSatisfaction
Histogram
Frequency table


Attrition by JobRole
Histogram
Frequency table
Attrition by EducationField
Histogram
Frequency table
Attrition by JobLevel
Histogram
Frequency table

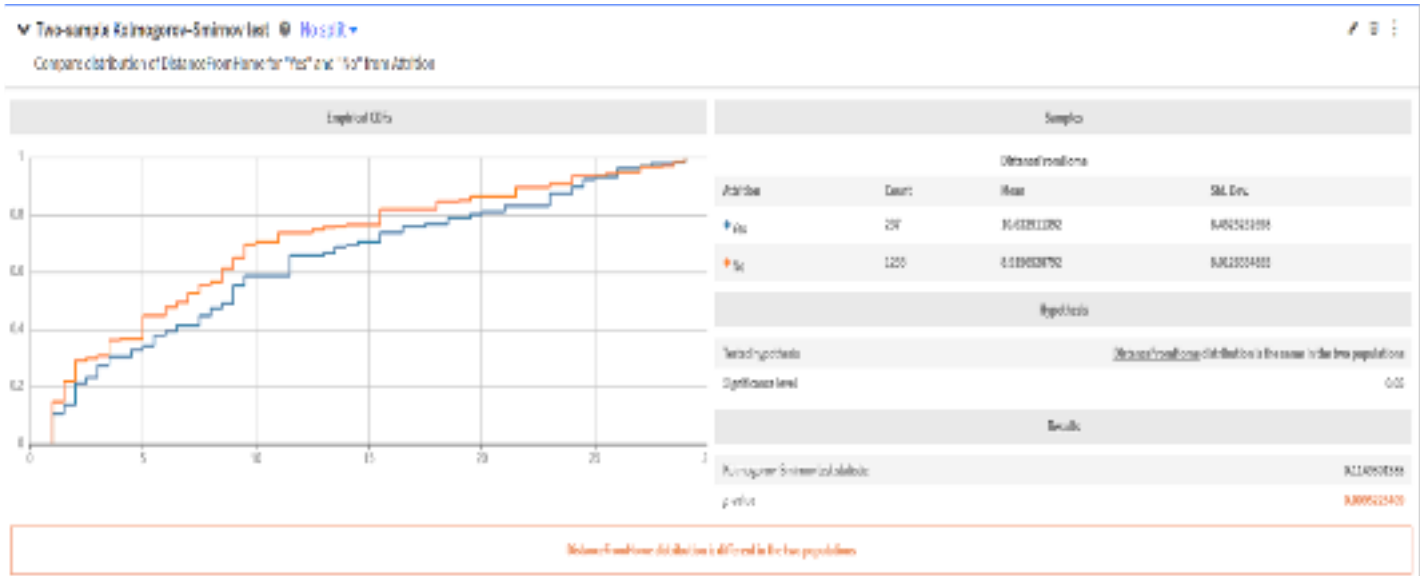

Attrition by Gender
Histogram
Frequency table
Attrition by EnvironmentSatisfaction
Histogram
Frequency table
Attrition by JobInvolvement
Histogram
Frequency table

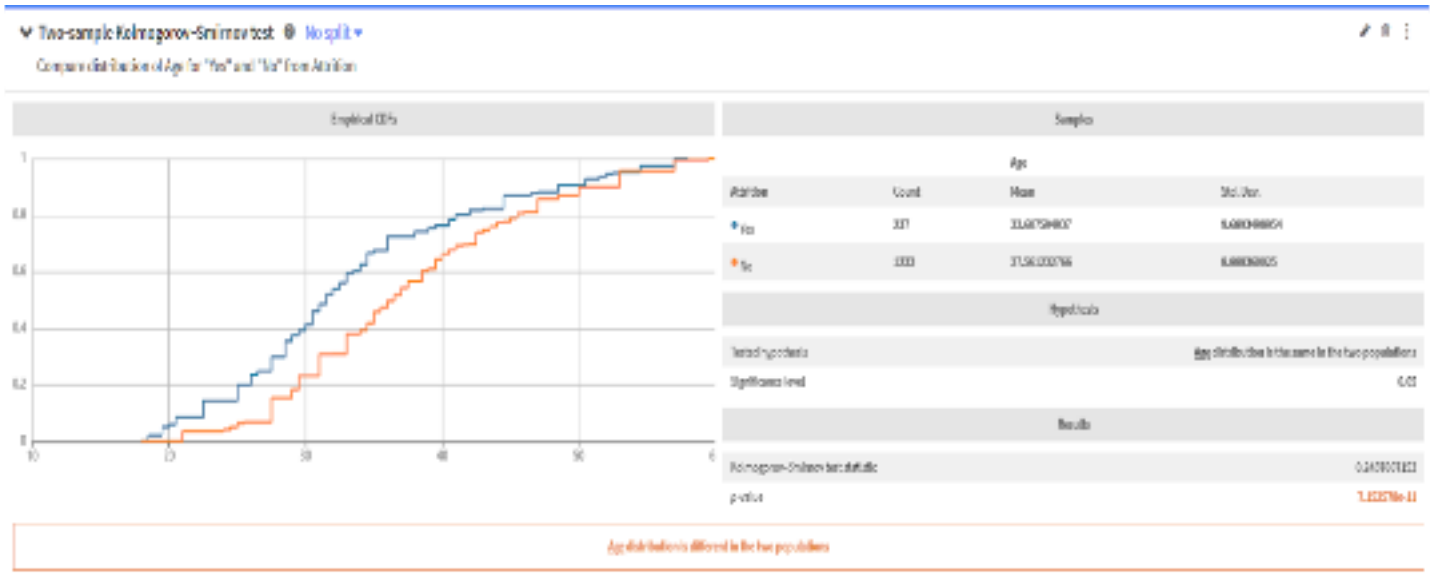
Statistical Test (For Categorical predictors): Chi-Square statistic

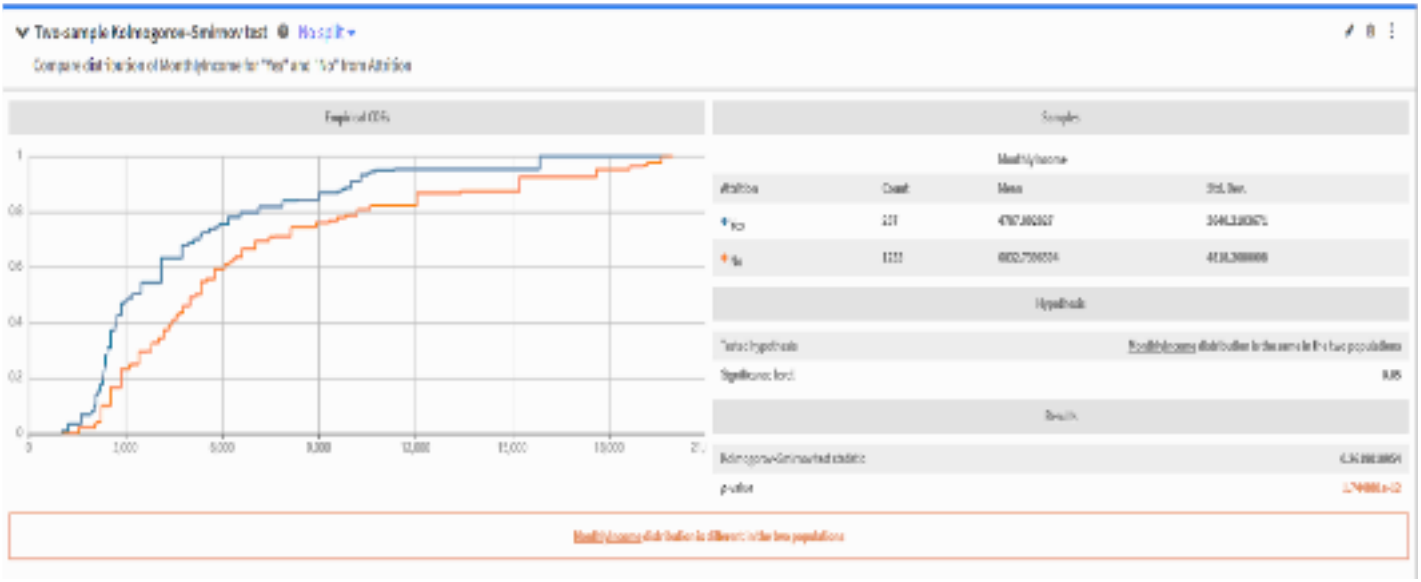
Hypotheses: **The categorical predictor and the Target variables are independent**

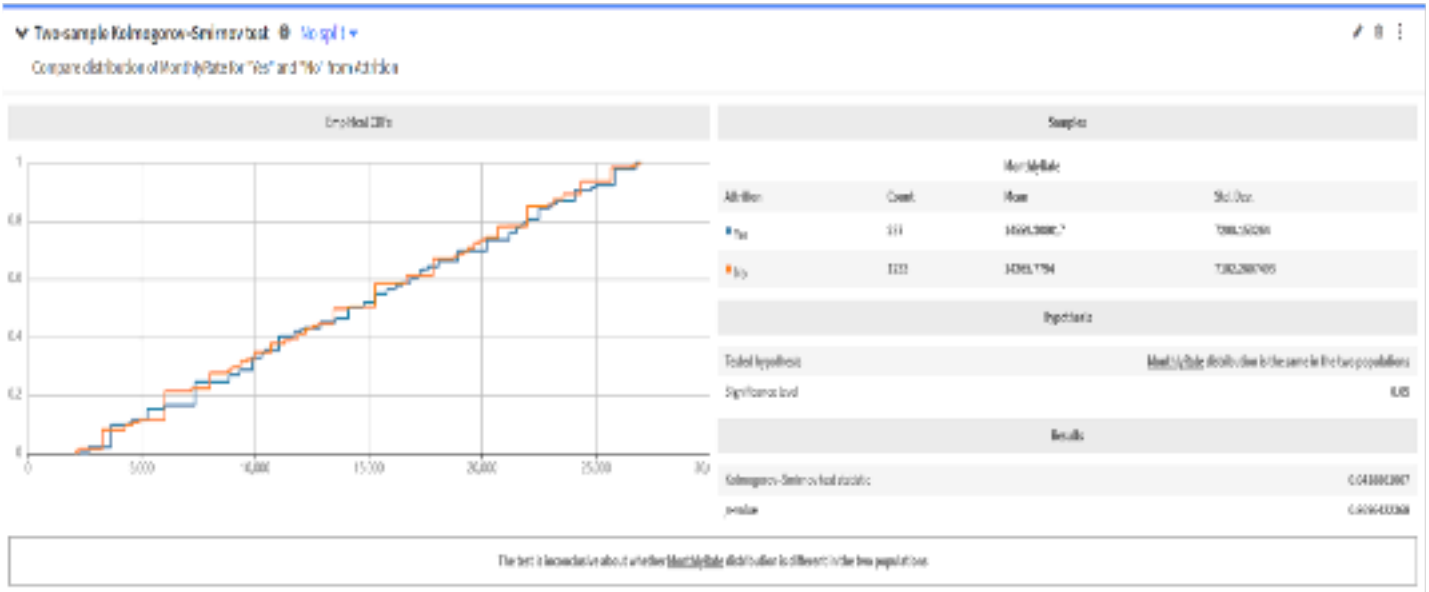
Categorical Variable (Predictor)	Test Inference
Gender	Inconclusive
Business Travel	Not Independent
Education Field	Not Independent
Department	Not Independent
Marital Status	Not Independent
Job Role	Not Independent
Over Time	Not Independent
Education	Inconclusive
Relationship Status	Inconclusive
Number of Companies Worked	Not Independent
Job Satisfaction	Not Independent
Performance Rating	Inconclusive
Job level	Not Independent
Job Involvement	Not Independent
Stock option level	Not Independent
Training time last year	Not Independent
Work-life Balance	Not Independent

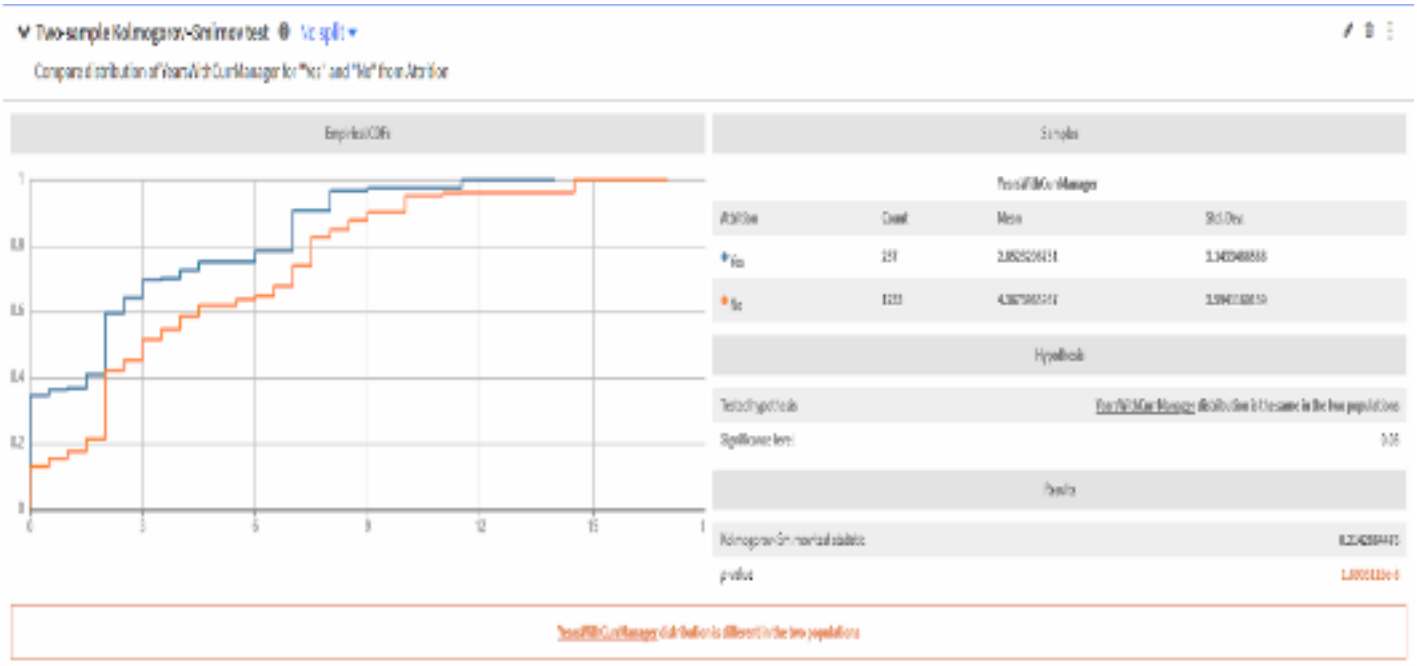
Statistical Test (For Continuous predictors): 2-sample Kolmogrov-Smirnov statistic



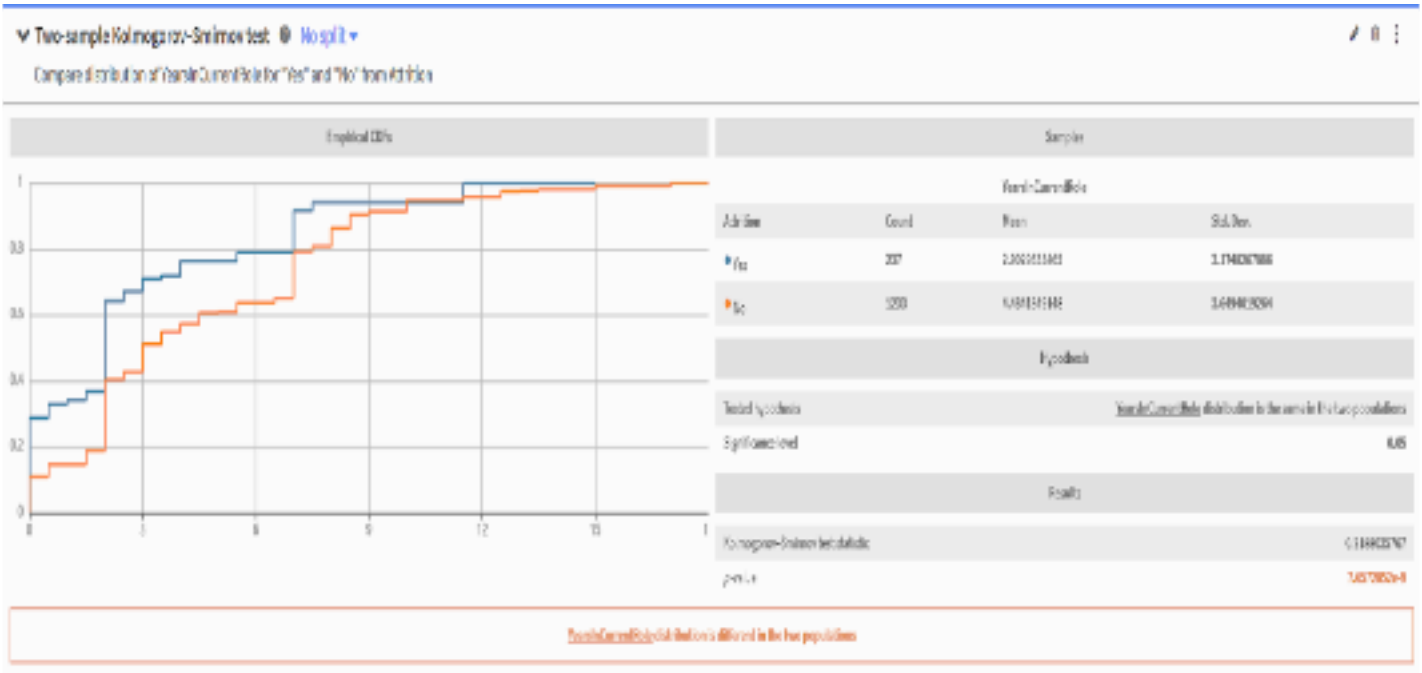


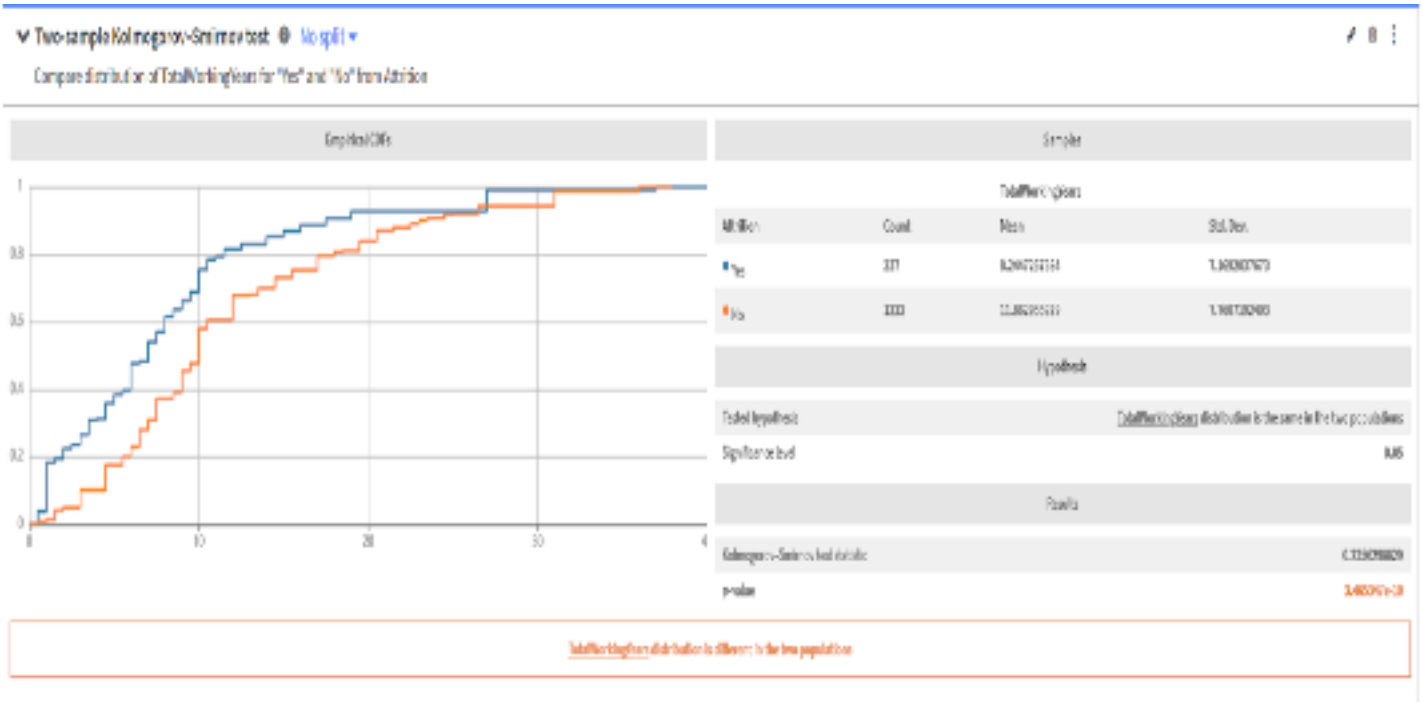


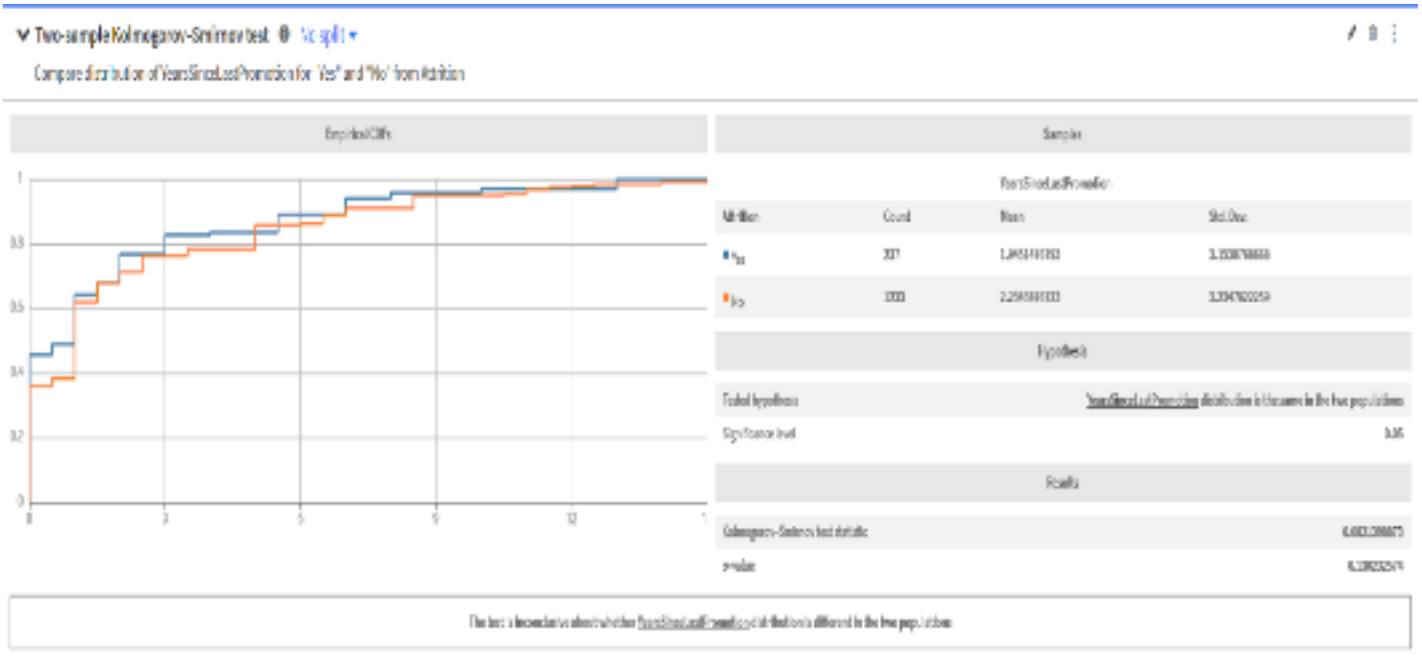

Two-sample Kolmogorov-Smirnov test
No split
Compare distribution of MonthlyRate for "Yes" and "No" from Attrition
Empirical CDFs
Samples
Hypothesis
Results
The test is inconclusive about whether MonthlyRate distribution is different in the two populations


Two-sample Kolmogorov-Smirnov test
No split
Compare distribution of YearsWithCurrManager for "Yes" and "No" from Attrition
Empirical CDFs
Samples
Hypothesis
Results
YearsWithCurrManager distribution is different in the two populations


Two-sample Kolmogorov-Smirnov test
Empirical CDFs
Samples
Hypothesis
Results


Two-sample Kolmogorov-Smirnov test
Empirical CDFs
Samples
Hypothesis
Results


Two-sample Kolmogorov-Smirnov test
No split
Compare distribution of TotalWorkingYears for "Yes" and "No" from Attrition
Empirical CDFs
Samples
Hypothesis
Results
p-value


Two-sample Kolmogorov-Smirnov test
No split
Compare distribution of YearsSinceLastPromotion for "Yes" and "No" from Attrition
Empirical CDFs
Samples
Hypothesis
Results
p-value

```
library(caret)
```

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

```
##
## Attaching package: 'caret'
```

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

```
library(randomForest)
```

```
## randomForest 4.6-14
```

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

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

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

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

```
library(xgboost)
```

```
##
## Attaching package: 'xgboost'
```

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

```
library(psych)
```

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

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

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

```
library(ROCR)
library(ggplot2)
```

Step-1: Train-Test Dataset splitting:

Train-Test split (We use Test Dataset as the validation set also)

```
set.seed(1234)
ind<- sample(2, nrow(dfn), replace=TRUE, prob=c(0.8,0.2))
training <- dfn[ind==1,]
testing <- dfn[ind==2,]
y_train <- training[,2]
x_train <- training[,-2]
x_test <- testing[,-2]
y_test<- testing[,2]
```

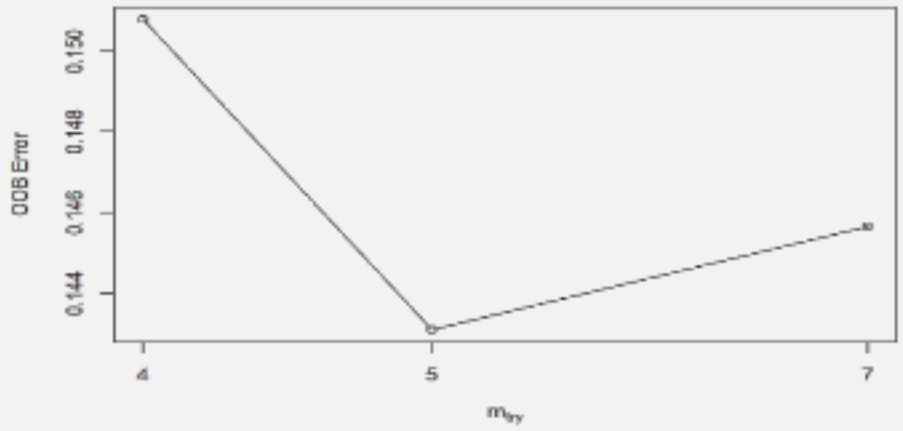
Step-2: Chosen Algorithm-1: Random Forest:

We tune the Random Forest with its tuning package tuneRF. It gives us an optimal value of mtry (mtry: Number of variables randomly sampled as candidates at each split).

Random Forest: Tuning with TuneRF

```
beatmtry <- tuneRF(x_train, y_train,stepFactor = 1.5,improve = 1e-5,ntree=500)
```

```
## mtry = 8  OOB error = 14.51%
## Searching left ...
## mtry = 4      OOB error = 14.08%
## -0.05357143 1e-05
## Searching right ...
## mtry = 7      OOB error = 14.57%
## -0.01785714 1e-05
```



Random Forest: Using the best

Now we implement our Random Forest model with the best value obtained from the previous tuning.

```
modelrf <- randomForest(x_train,y_train,ntree = 500, mtry = 5, importance = TRUE)
modelrf
```

```
##
## Call:
##  randomForest(x = x_train, y = y_train, ntree = 500, mtry = 5,      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 14.65%
## Confusion matrix:
##      No Yes class.error
## No  969   7 0.007172131
## Yes 165  33 0.833333333
```

Prediction and Model evaluation:

```
pred= predict(modelrf,training)
pred1= predict(modelrf,testing)
confusionMatrix(testing$Attrition,pred1)
```

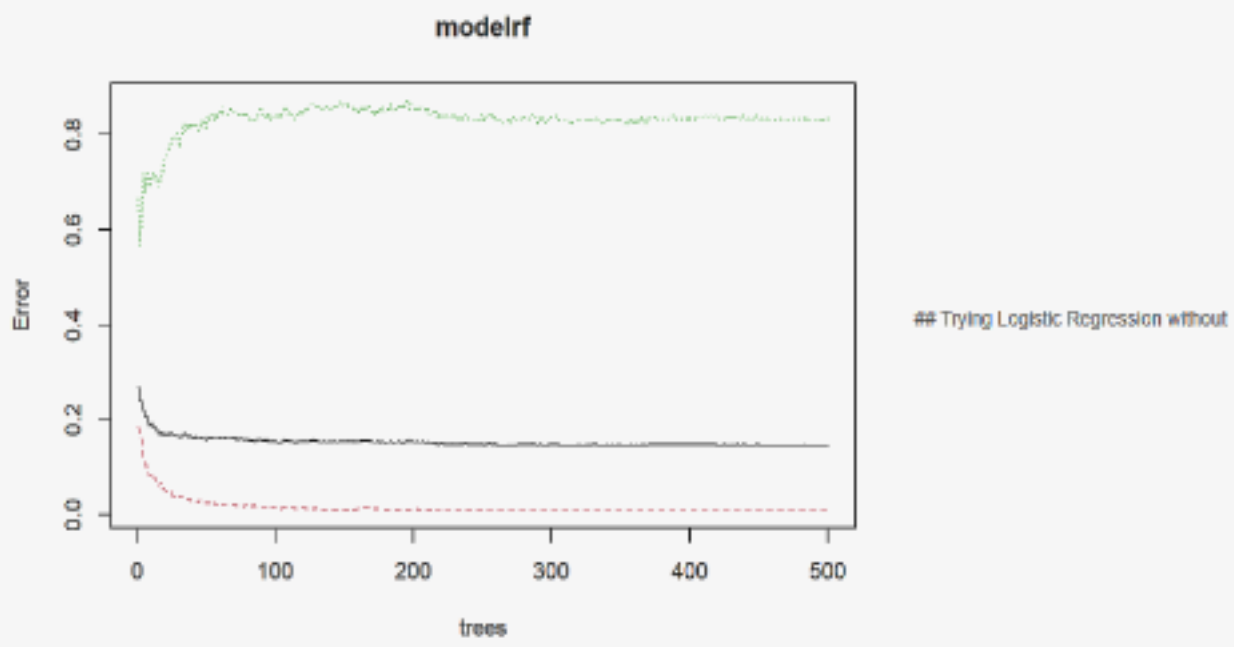
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  255   2
##        Yes  35   4
##
##               Accuracy : 0.875
##                 95% CI : (0.8318, 0.9104)
##    No Information Rate : 0.9797
##    P-Value [Acc > NIR] : 1
##
##                  Kappa : 0.1478
##
##  Mcnemar's Test P-Value : 1.435e-07
##
##             Sensitivity : 0.8793
##             Specificity : 0.6667
##          Pos Pred Value : 0.9922
##          Neg Pred Value : 0.1026
##              Prevalence : 0.9797
##          Detection Rate : 0.8615
##    Detection Prevalence : 0.8682
##       Balanced Accuracy : 0.7730
##
##        'Positive' Class : No
##
```

Evaluation metrics: Accuracy and Cohen's Kappa score

```
postResample(pred1,testing$Attrition)
```

```
##  Accuracy     Kappa
## 0.8750000 0.1475369
```

```
plot(modelrf)
```



Trying Logistic Regression without

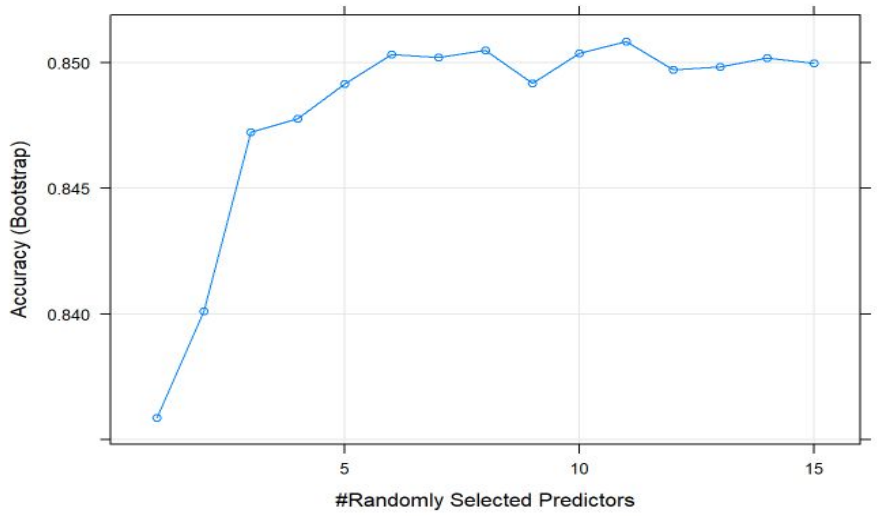
Though we obtained quite high accuracy (more than 87 percent) if we look at the confusion matrix and the kappa statistic, we know that the model has inherited a tendency of assigning most instances into 1 single class. This is not good as by this, the model is not performing much better than a random predictor when it comes to imbalanced classification problem.

Improving model performance with Grid Search:

```
control <- trainControl(method='repeatedcv',number=10, repeats=3,search='grid')
tunegrid <- expand.grid(.mtry=(1:15))
rf_grid <- train(Attrition ~.,data = training,method='rf',metric='Accuracy',tuneGrid=tunegrid)
print(rf_grid)
```

```
## Random Forest
##
## 1174 samples
##   30 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1174, 1174, 1174, 1174, 1174, 1174, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    1    0.8358643  0.00000000
##    2    0.8401150  0.04413782
##    3    0.8472237  0.12180184
##    4    0.8477639  0.14036962
##    5    0.8491479  0.15956118
##    6    0.8503300  0.17977988
##    7    0.8502000  0.18472463
##    8    0.8504905  0.19738680
##    9    0.8491863  0.19464988
##   10    0.8503695  0.20411793
##   11    0.8508467  0.21653101
##   12    0.8497251  0.21299959
##   13    0.8498307  0.21401074
##   14    0.8501891  0.21963302
##   15    0.8499790  0.22173824
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 11.
```

```
plot(rf_grid)
```



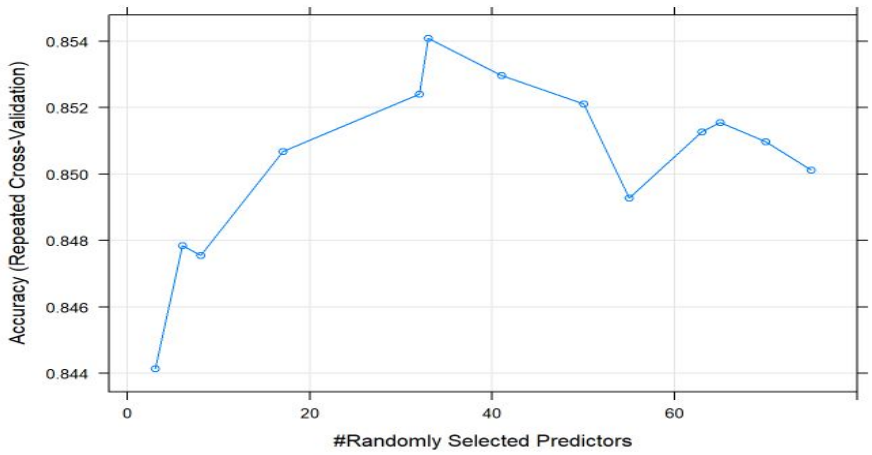
Random Forest:Tuning further

Improving model performance with Random Search:

```
controlrand <- trainControl(method='repeatedcv',number=10, repeats=3,search='random')
set.seed(2)
mtry <- sqrt(ncol(x_train))
rf_random <- train(Attrition ~.,data = training,method='rf',metric='Accuracy',tuneLength=15,trControl=controlrand)
print(rf_random)
```

```
## Random Forest
##
## 1174 samples
##   30 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 1056, 1057, 1056, 1056, 1056, 1056, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    3    0.8441374  0.1172869
##    6    0.8478461  0.1773340
##    8    0.8475587  0.1869937
##   17    0.8506854  0.2330907
##   32    0.8524022  0.2803375
##   33    0.8540922  0.2924259
##   41    0.8529599  0.2984886
##   50    0.8521099  0.3062337
##   55    0.8492780  0.3008756
##   63    0.8512626  0.3135094
##   65    0.8515548  0.3116346
##   70    0.8509728  0.3216045
##   75    0.8501253  0.3165339
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 33.
```

```
plot(rf_random)
```



Variable Importance: Random Forest Model

```
importancerf <-round(importance(modelrf),2)
newimp <-data.frame(importancerf)
print(newimp)
```

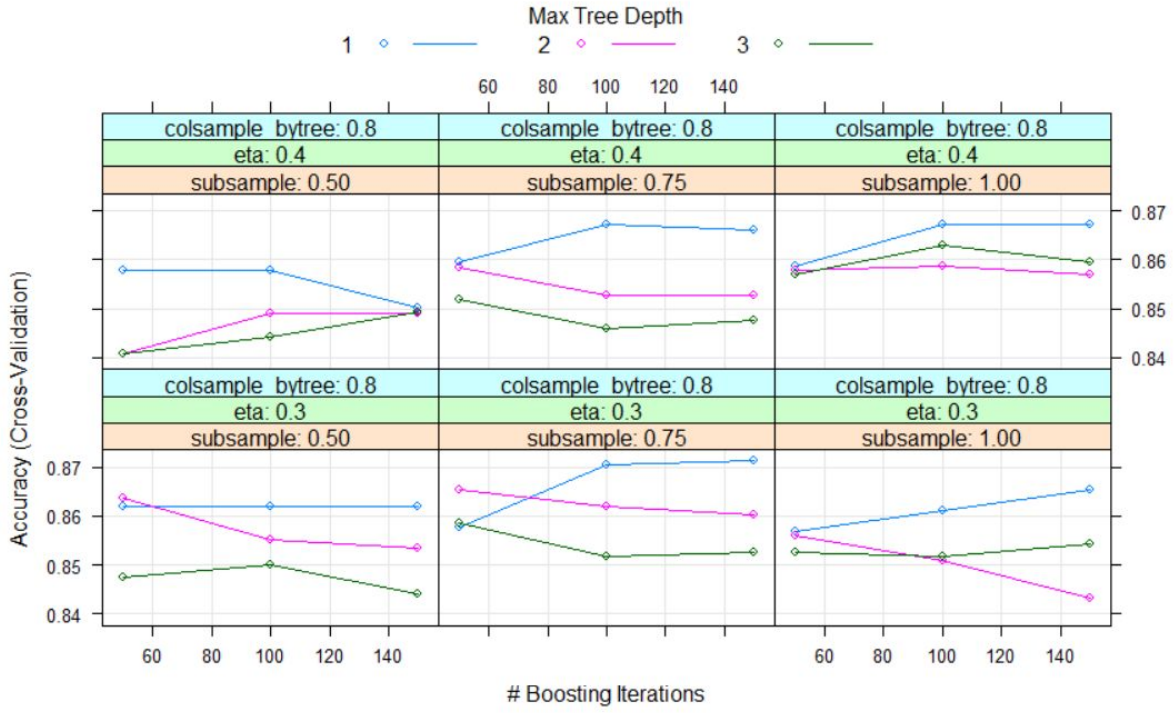
```
##                             No   Yes MeanDecreaseAccuracy MeanDecreaseGini
## i..Age                    8.86  7.94                11.89            19.63
## BusinessTravel            0.88 -0.53                 0.59             3.94
## DailyRate                -0.17 -0.41                -0.36            16.92
## Department                1.74  4.77                 3.73             3.90
## DistanceFromHome          1.38 -0.09                 1.21            13.92
## Education                 2.78 -1.03                 1.94             7.90
## EducationField            1.29 -0.15                 1.12             6.69
## EnvironmentSatisfaction   1.72  6.76                 4.94            11.51
## Gender                    0.56 -0.56                 0.31             2.81
## HourlyRate               -1.25 -1.43                -1.73            14.23
## JobInvolvement            2.34  3.66                 3.96             8.94
## JobLevel                  6.03  8.00                 9.34             7.42
## JobRole                   4.32  5.78                 6.60            10.39
## JobSatisfaction           3.63  1.29                 3.74             9.59
## MaritalStatus             4.91  8.45                 8.11             7.50
## MonthlyIncome             9.49  8.47                13.12            23.15
## MonthlyRate              -0.02 -1.77                -0.77            15.42
## NumCompaniesWorked        3.36  0.86                 3.49            21.11
## OverTime                 15.57 21.07                23.20            17.99
## PercentSalaryHike         2.64  0.29                 2.35            11.54
## PerformanceRating         0.14 -0.79                -0.23             1.23
## RelationshipSatisfaction  0.70  0.33                 0.82             7.91
## StockOptionLevel          8.39  9.48                11.47            10.23
## TotalWorkingYears         7.82  4.04                 9.36            15.11
## TrainingTimesLastYear     0.20 -1.29                -0.46            12.49
## WorkLifeBalance           2.86  3.36                 3.92             8.70
## YearsAtCompany            8.26  4.00                 9.80            13.64
## YearsInCurrentRole        4.65  3.36                 6.05             8.33
## YearsSinceLastPromotion   3.94 -0.56                 3.39             7.69
## YearsWithCurrManager      6.64  3.34                 8.04             9.49
```

Chosen Algorithm 2: Gradient boosted Tree (XGBOOST):

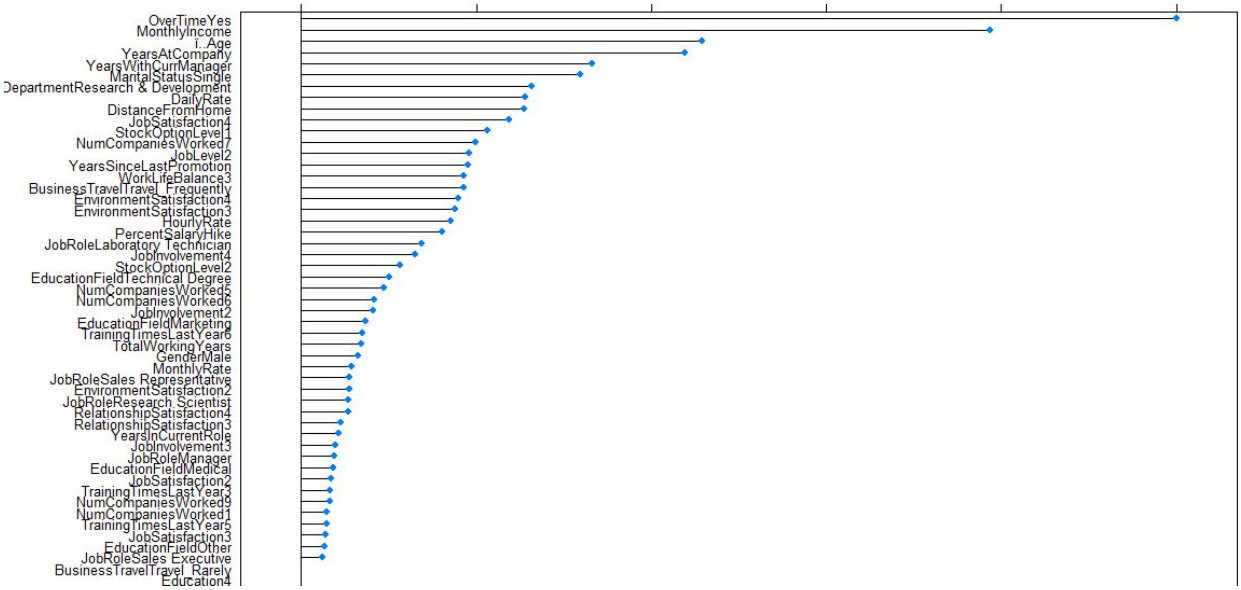
```
> #Model-3: Gradient Boosted Tree (XG Boost)
> controlxg <- trainControl(method='repeatedcv',number=10, repeats=3)
> modelxgcv <- train(Attrition ~., data=training, method='xgbTree',trControl=controlxg)

> #Model-3: Gradient Boosted Tree (XG Boost)
> controlxg <- trainControl(method='cv',number=10)
> modelxgcv <- train(Attrition ~., data=training, method='xgbTree',trControl=controlxg)
> plot(modelxgcv)
```

Plotting the model:



Variable Importance plotting:



Performance Evaluation:

```
> mean(predxg==testing$Attrition)
[1] 0.8885135
> postResample(predxg,testing$Attrition)
 Accuracy     Kappa
0.8885135 0.3858149
```

The Kappa value has improved from Random Forest but still it has not crossed the boundary of 0.40. Accuracy is extremely good. We will try a simpler model with a little trade off of accuracy and will try to improve the kappa value.

Chosen Algorithm-3: Logistic Regression without cross validation:

```
modellogr= glm(formula=Attrition ~.,data = training,family = binomial)
summary(modellogr)
```

```
##
## Call:
## glm(formula = Attrition ~ ., family = binomial, data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9178  -0.4359  -0.1893  -0.0500   3.5867
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -9.081e+00  7.163e+02  -0.013 0.989885
## ï..Age                          -3.531e-02  1.673e-02  -2.110 0.034871 *
## BusinessTravelTravel_Frequently  1.708e+00  4.888e-01   3.495 0.000474 ***
## BusinessTravelTravel_Rarely      8.776e-01  4.457e-01   1.969 0.048967 *
## DailyRate                       -4.929e-04  2.731e-04  -1.805 0.071120 .
## DepartmentResearch & Development 1.508e+01  7.163e+02   0.021 0.983198
## DepartmentSales                  1.415e+01  7.163e+02   0.020 0.984237
## DistanceFromHome                 5.688e-02  1.351e-02   4.210 2.55e-05 ***
## Education2                      -8.201e-02  3.882e-01  -0.211 0.832690
## Education3                       1.072e-01  3.392e-01   0.316 0.752036
## Education4                       8.258e-02  3.740e-01   0.221 0.825240
## Education5                      -4.717e-01  9.462e-01  -0.499 0.618109
## EducationFieldLife Sciences     -1.269e+00  1.063e+00  -1.194 0.232592
## EducationFieldMarketing         -6.571e-01  1.120e+00  -0.587 0.557468
## EducationFieldMedical           -1.310e+00  1.064e+00  -1.231 0.218426
## EducationFieldOther             -1.110e+00  1.148e+00  -0.967 0.333746
## EducationFieldTechnical Degree  -1.508e-01  1.065e+00  -0.142 0.887447
## EnvironmentSatisfaction2        -1.167e+00  3.319e-01  -3.515 0.000440 ***
## EnvironmentSatisfaction3        -1.498e+00  3.185e-01  -4.705 2.54e-06 ***
## EnvironmentSatisfaction4        -1.485e+00  3.127e-01  -4.749 2.04e-06 ***
## GenderMale                       3.766e-01  2.267e-01   1.662 0.096604 .
## HourlyRate                       1.552e-03  5.413e-03   0.287 0.774396
## JobInvolvement2                 -1.448e+00  4.271e-01  -3.390 0.000698 ***
## JobInvolvement3                 -1.626e+00  3.998e-01  -4.067 4.76e-05 ***
## JobInvolvement4                 -2.645e+00  5.836e-01  -4.532 5.84e-06 ***
## JobLevel2                       -1.252e+00  5.186e-01  -2.415 0.015754 *
## JobLevel3                        1.503e-01  8.224e-01   0.183 0.855011
## JobLevel4                       -1.135e+00  1.403e+00  -0.809 0.418501
## JobLevel5                        2.156e+00  1.827e+00   1.180 0.238136
## JobRoleHuman Resources           1.554e+01  7.163e+02   0.022 0.982686
## JobRoleLaboratory Technician     9.875e-01  6.755e-01   1.462 0.143790
```

```
## TrainingTimesLastYear3           -1.746e+00  5.379e-01  -3.247 0.001167 **
## TrainingTimesLastYear4           -1.201e+00  6.106e-01  -1.966 0.049243 *
## TrainingTimesLastYear5           -1.945e+00  6.517e-01  -2.984 0.002845 **
## TrainingTimesLastYear6           -2.261e+00  7.799e-01  -2.899 0.003742 **
## WorkLifeBalance2                 -1.390e+00  4.823e-01  -2.883 0.003943 **
## WorkLifeBalance3                 -1.864e+00  4.612e-01  -4.041 5.32e-05 ***
## WorkLifeBalance4                 -1.342e+00  5.364e-01  -2.501 0.012380 *
## YearsAtCompany                    1.456e-01  4.712e-02   3.091 0.001996 **
## YearsInCurrentRole               -1.867e-01  5.885e-02  -3.173 0.001507 **
## YearsSinceLastPromotion           1.575e-01  5.047e-02   3.121 0.001801 **
## YearsWithCurrManager             -1.804e-01  5.658e-02  -3.188 0.001434 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1065.40  on 1173  degrees of freedom
## Residual deviance:  617.79  on 1098  degrees of freedom
## AIC: 769.79
##
## Number of Fisher Scoring iterations: 15
```

Model Evaluation: Anova and Chi square :

```
anova(modellogr,test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Attrition
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                      1173    1065.40
## ï..Age                    1   31.304      1172    1034.09 2.206e-08 ***
## BusinessTravel            2   13.020      1170    1021.07 0.0014882 **
## DailyRate                 1    2.394      1169    1018.68 0.1217685
## Department                2   10.679      1167    1008.00 0.0047989 **
## DistanceFromHome          1    4.494      1166    1003.50 0.0340162 *
## Education                 4    3.313      1162    1000.19 0.5068056
## EducationField            5    7.167      1157     993.02 0.2085009
## EnvironmentSatisfaction   3   23.610      1154     969.41 3.013e-05 ***
## Gender                    1    0.901      1153     968.51 0.3426024
## HourlyRate                1    0.111      1152     968.40 0.7393123
## JobInvolvement            3   27.752      1149     940.65 4.095e-06 ***
## JobLevel                  4   50.926      1145     889.72 2.314e-10 ***
## JobRole                   8   15.056      1137     874.67 0.0580707 .
## JobSatisfaction           3   15.611      1134     859.06 0.0013624 **
## MaritalStatus             2   28.387      1132     830.67 6.851e-07 ***
## MonthlyIncome             1    0.097      1131     830.57 0.7554789
## MonthlyRate               1    1.255      1130     829.32 0.2625076
## NumCompaniesWorked        9   26.103      1121     803.21 0.0019654 **
## OverTime                  1  100.632      1120     702.58 < 2.2e-16 ***
## PercentSalaryHike         1    0.669      1119     701.91 0.4134010
## PerformanceRating         1    0.360      1118     701.55 0.5486294
## RelationshipSatisfaction  3    9.912      1115     691.64 0.0193257 *
## StockOptionLevel          3   11.209      1112     680.43 0.0106457 *
## TotalWorkingYears         1    0.475      1111     679.96 0.4907803
## TrainingTimesLastYear     6   11.935      1105     668.02 0.0634238 .
## WorkLifeBalance           3   18.831      1102     649.19 0.0002963 ***
## YearsAtCompany            1    0.984      1101     648.21 0.3211244
## YearsInCurrentRole        1   11.516      1100     636.69 0.0006899 ***
## YearsSinceLastPromotion   1    8.765      1099     627.93 0.0030707 **
## YearsWithCurrManager      1   10.138      1098     617.79 0.0014525 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Performance evaluation: Logistic Regression

```
modellogr$aic
```

```
## [1] 769.7875
```

```
predlr= predict(modellogr, testing, type = "response")
binpred= ifelse(predlr>0.5,"Yes","No")
mean(binpred == testing$Attrition)
```

```
## [1] 0.8716216
```

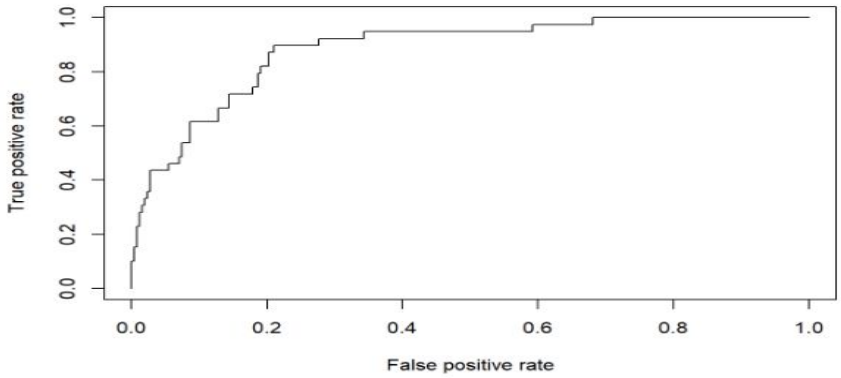
```
postResample(binpred,testing$Attrition)
```

```
##  Accuracy     Kappa
## 0.8716216 0.4133723
```

Thus we have obtained a good AIC score (That implies our model has learned quite good from the training Data), Accuracy of 87% and a substantially good Kappa score for imbalance Datasets: 0.41.

Plotting the ROC for Logistic regression and calculating AUC:

```
pr<- prediction(predlr, testing$Attrition)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```

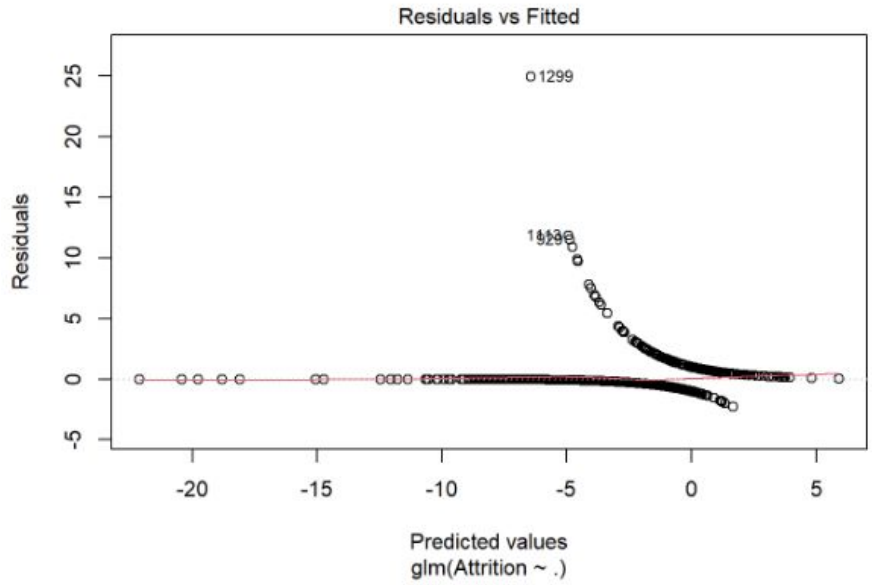


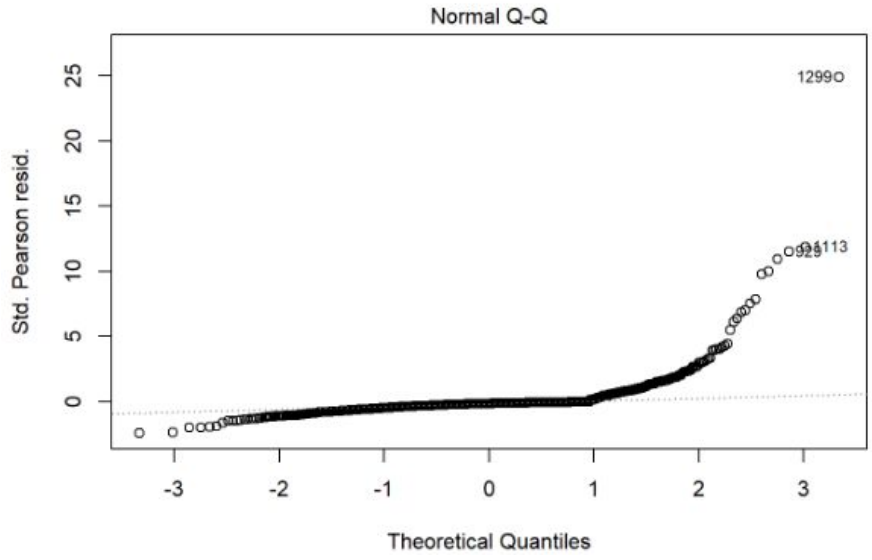

```
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.8841664
```

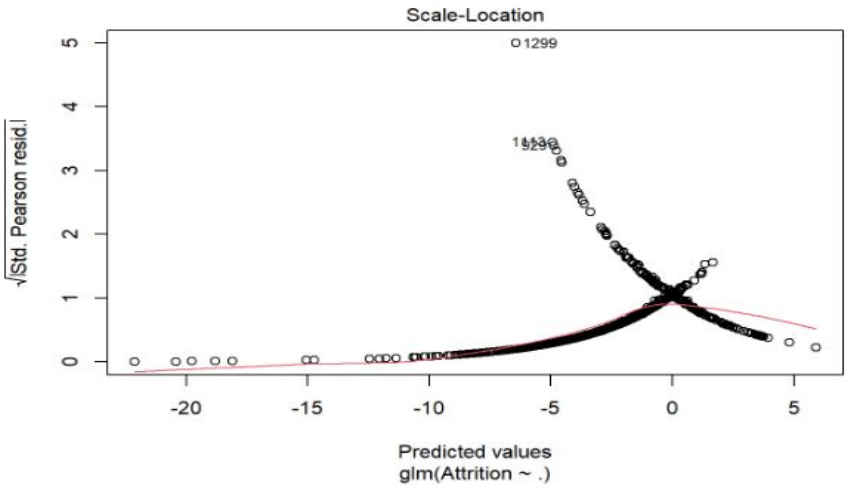
Plotting the Model: Logistic Regression

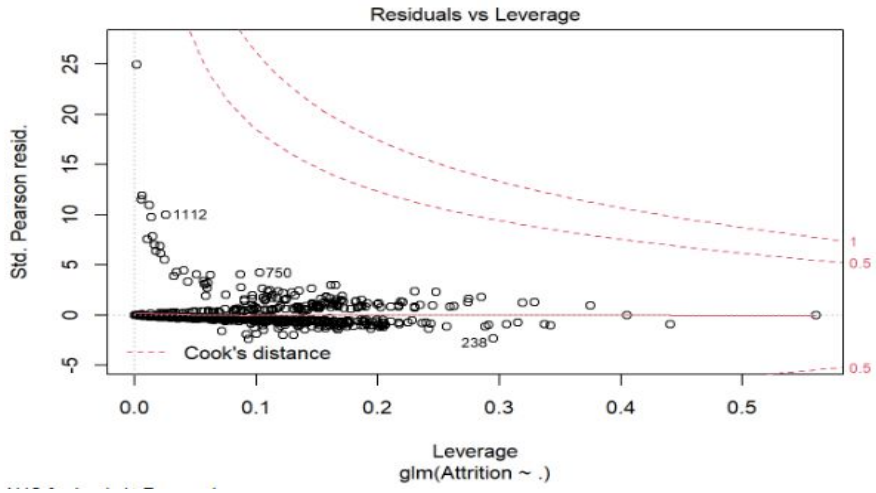
```
plot(modellogr)
```





Plotting the Model: Logistic Regression (continued)





##Plotting the ROC and calculating

Conclusion and Future Vision: **In this project we performed descriptive statistics, exploratory analysis and predictive modeling. Our analysis claims that we have achieved an 88% accuracy. By all means of evaluation we can choose Logistic Regression or XGBoost for prediction of Attrition based on the specific project requirement. Further tuning of the models and more intensive preprocessing will result in improved performance.**