

# An HR Analytics Case study

based on

## Logistic regression

predicting

## Employee Attrition



*A company's employees are its greatest asset and your people are your product.*

-Richard Branson

## Problem Statement

Every year ~15% of employees leave XYZ company and need to be replaced with the talent pool available in the job market. This percentage of attrition(employees leaving, either on their own or because they got fired) impacts the company negatively, because of the following reasons:

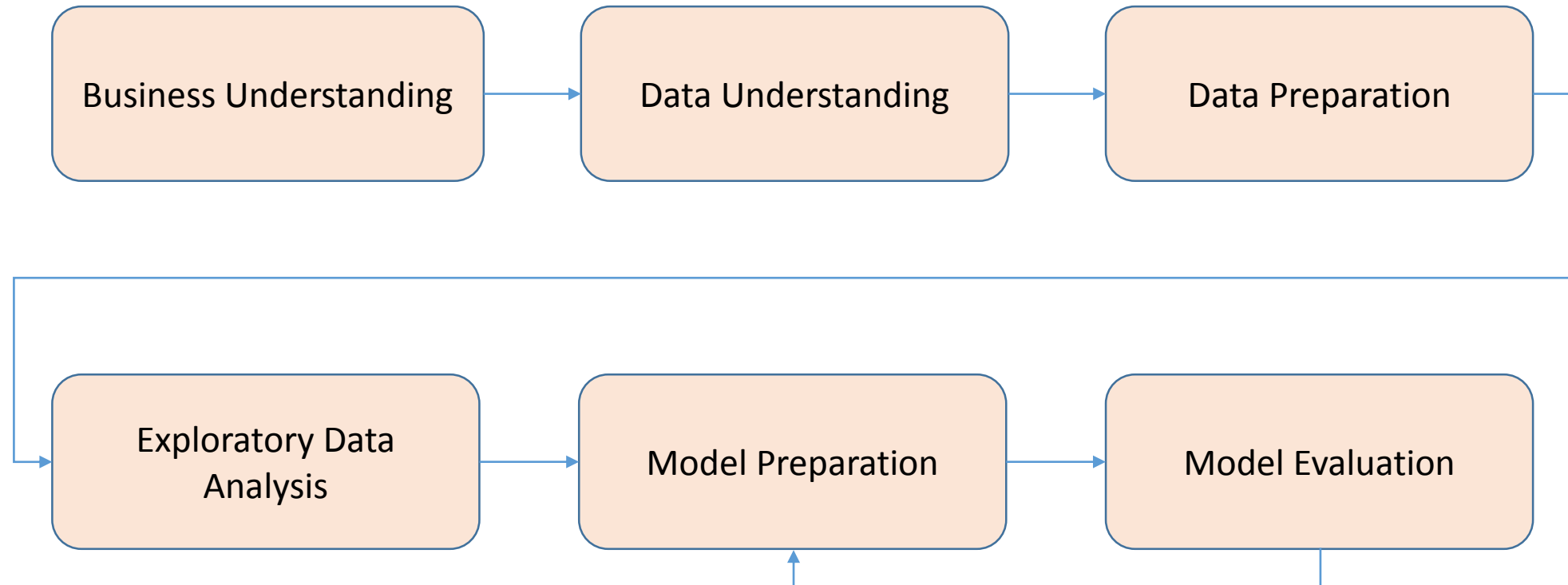
- The former employees' projects get delayed, which makes it difficult to meet **timelines**, resulting in a **reputation loss** among consumers and partners
- A sizeable department has to be maintained, for the purposes of **recruiting** new talent
- More often than not, the new employees have to be **trained** for the job and/or given time to acclimatize themselves to the company

## Objective

Model the probability of attrition using logistic regression :

- To understand what factors XYZ company should focus on to curb attrition, in other words, what changes should be made in the workplace, in order to get most of their employees to stay

# Problem Solving Methodology



## About the Data in hand

### General data

Contains demographic data and other behavioral details of the employees

- Age
- Attrition
- Business Travel
- Department
- Distance From Home
- Education
- Education Field
- Employee Count
- Employee ID
- Gender
- Job Level
- Job Role
- Marital Status
- Monthly Income
- Num. Companies Worked
- Over18
- Percent Salary Hike
- Standard Hours
- Stock Option Level
- Total Working Years
- Training Times Last Year
- Years At Company
- Years Since Last Promotion
- Years With Curr. Manager

**4410 observations & 24 variables**

#### Legend

- Categorical
- Continuous
- Key value used to merge

### Employee survey

Contains survey data from employees

- Employee ID
- Environment Satisfaction
- Job Satisfaction
- Work Life Balance

**4410 obs. & 4 var.**

### In time

Contains a year worth data of the employees' office in time in date time format

- Employee ID
- In-time details for days 01/01/2015 through 12/31/2015

**4410 obs. & 262 var.**

### Manager survey

Contains manager survey data for the employees

- Employee ID
- Job Involvement
- Performance Rating

**4410 obs. & 3 var.**

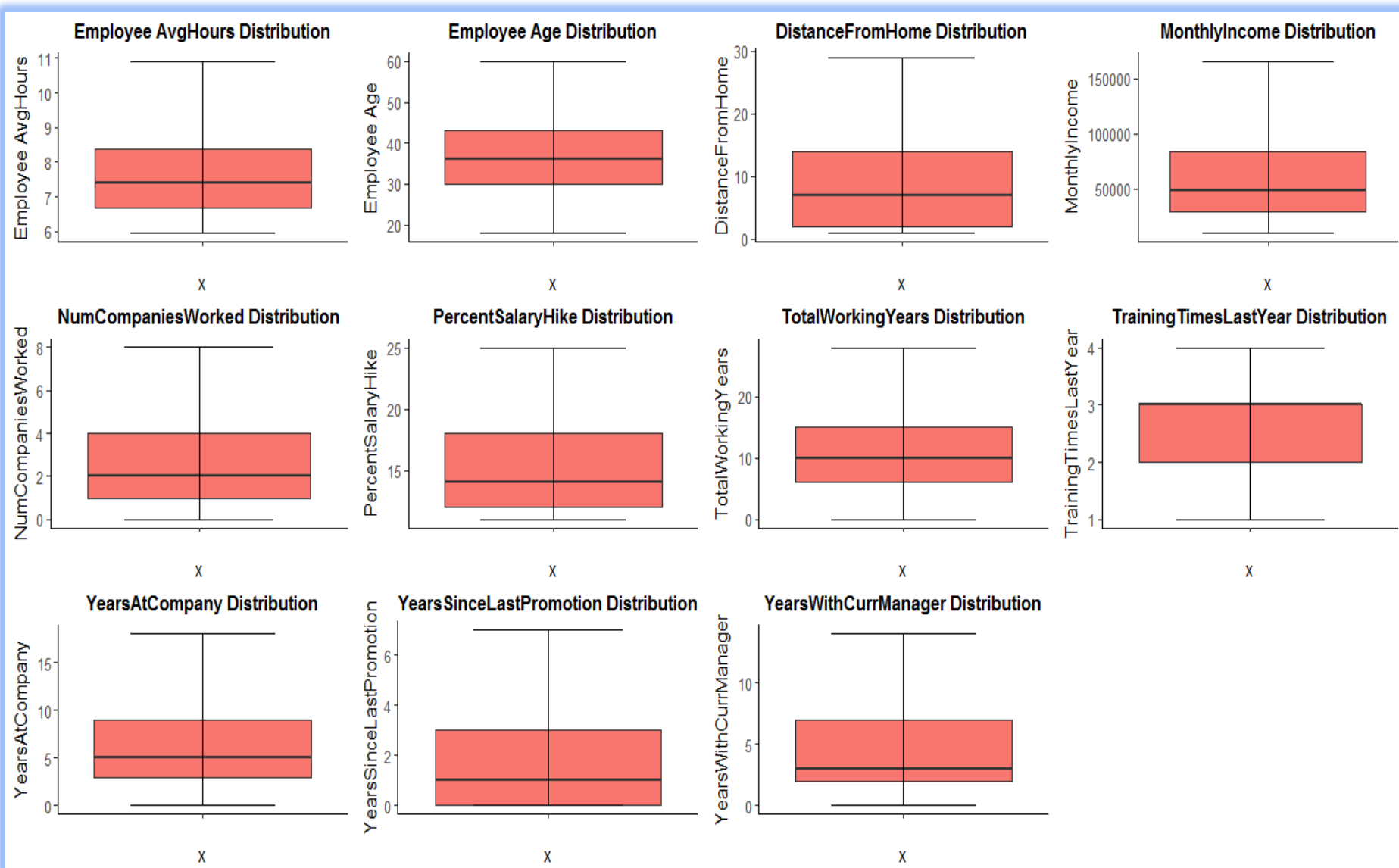
### Out time

Contains a year worth data of the employees' office out time in date time format

- Employee ID
- Out-time details for days 01/01/2015 through 12/31/2015

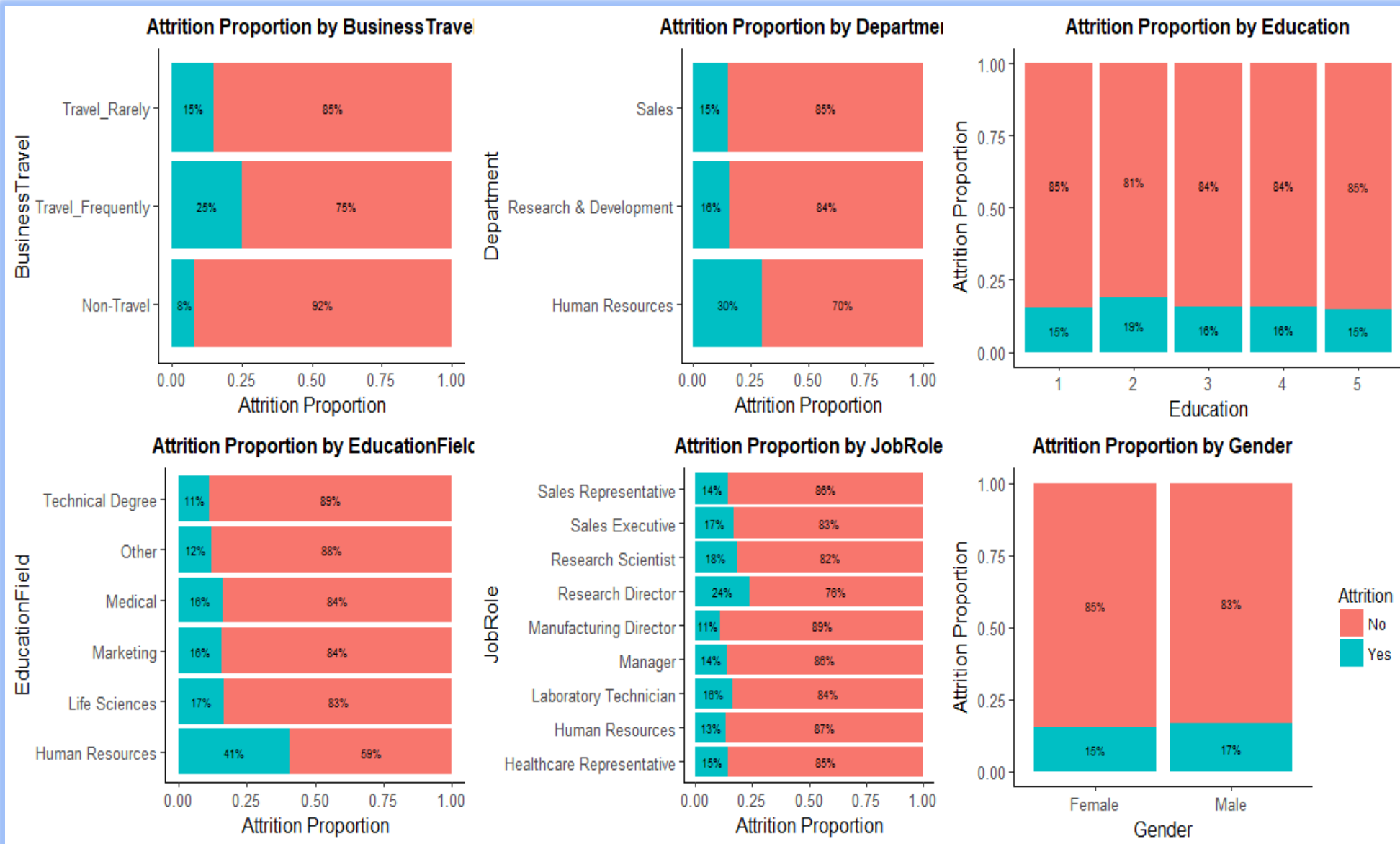
**4410 obs. & 262 var.**

- **Convert In\_time and Out\_time data** from wide format to long format in-order to calculate the average working hours and then eventually **merge** with the master employee data
- **New Variable Time In Office** derived based on the average working hours.
  - If the employee works > 10 Avg. Hours we will classify him as working overtime
  - If the employee works < 7 Avg. Hours we will classify him as working less time
  - And if the employee works somewhere between 7 & 10 Avg. Hours classify as normal
- **NA Values treated:**
  - NA values (Less than 1%) in columns Num. Companies Worked , Total Working Years, Environment Satisfaction, Job Satisfaction, Work Life Balance are replaced with Mode.
- **Single constant value columns removed:**
  - The columns Employee Count, Over18 and Standard Hours contains a single constant value which doesn't help in analysis. Hence these were removed.
- **Employee ID column removed**
- **Categorical variables:** All Categorical variables converted to Factors and then to dummy variables before model building.
- **Continuous variables:** All Continuous variables scaled.



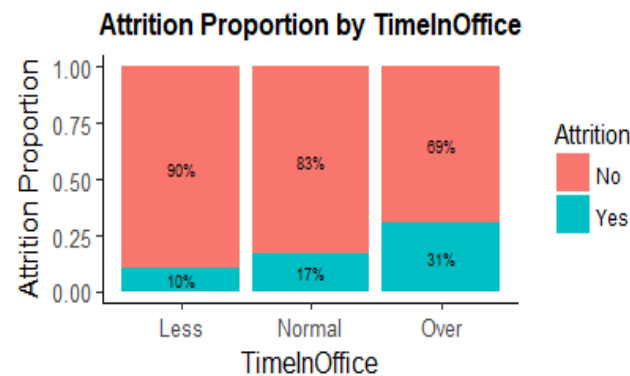
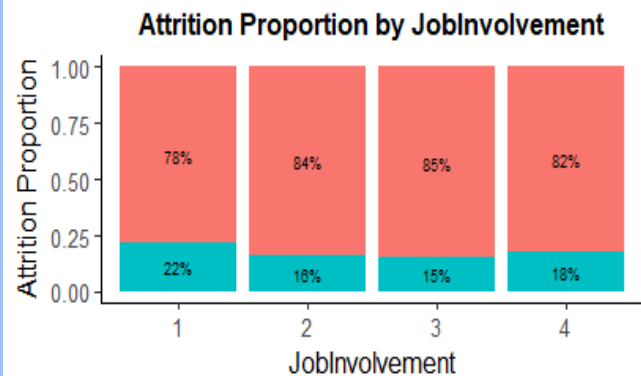
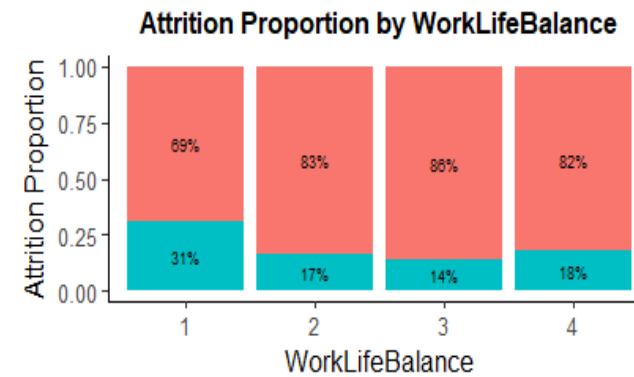
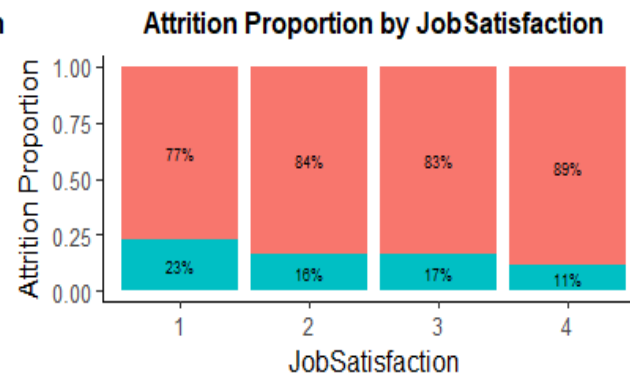
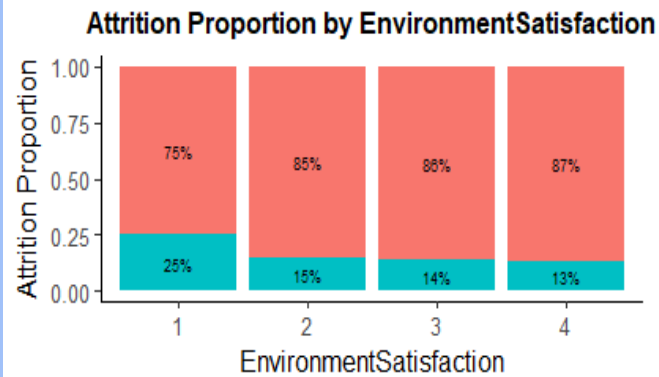
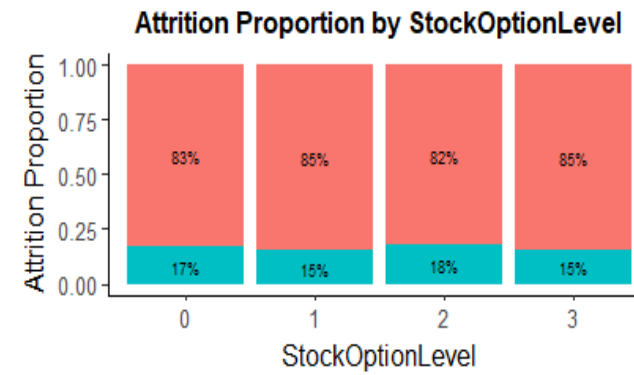
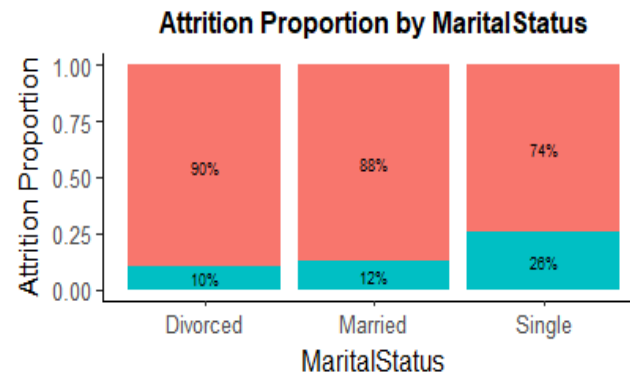
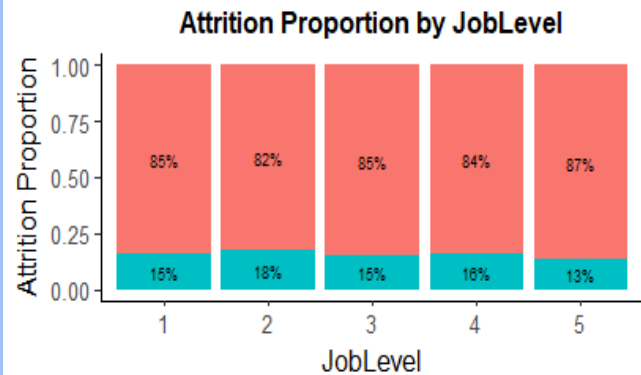
Handled Outliers in the following variables by capping:

- **Avg. Hours**
- **Monthly Income,**
- **Num. Companies Worked,**
- **Total Working Years**
- **Training Times Last Year,**
- **Years At Company,**
- **Years Since Last Promotion**
- **Years With Curr. Manager**



Attrition is higher for the following variables

- **Frequent BusinessTravel**
- **Department - Human Resources department**
- **Education field - Human Resources**
- **Job Role - Research Director**



Attrition  
■ No  
■ Yes

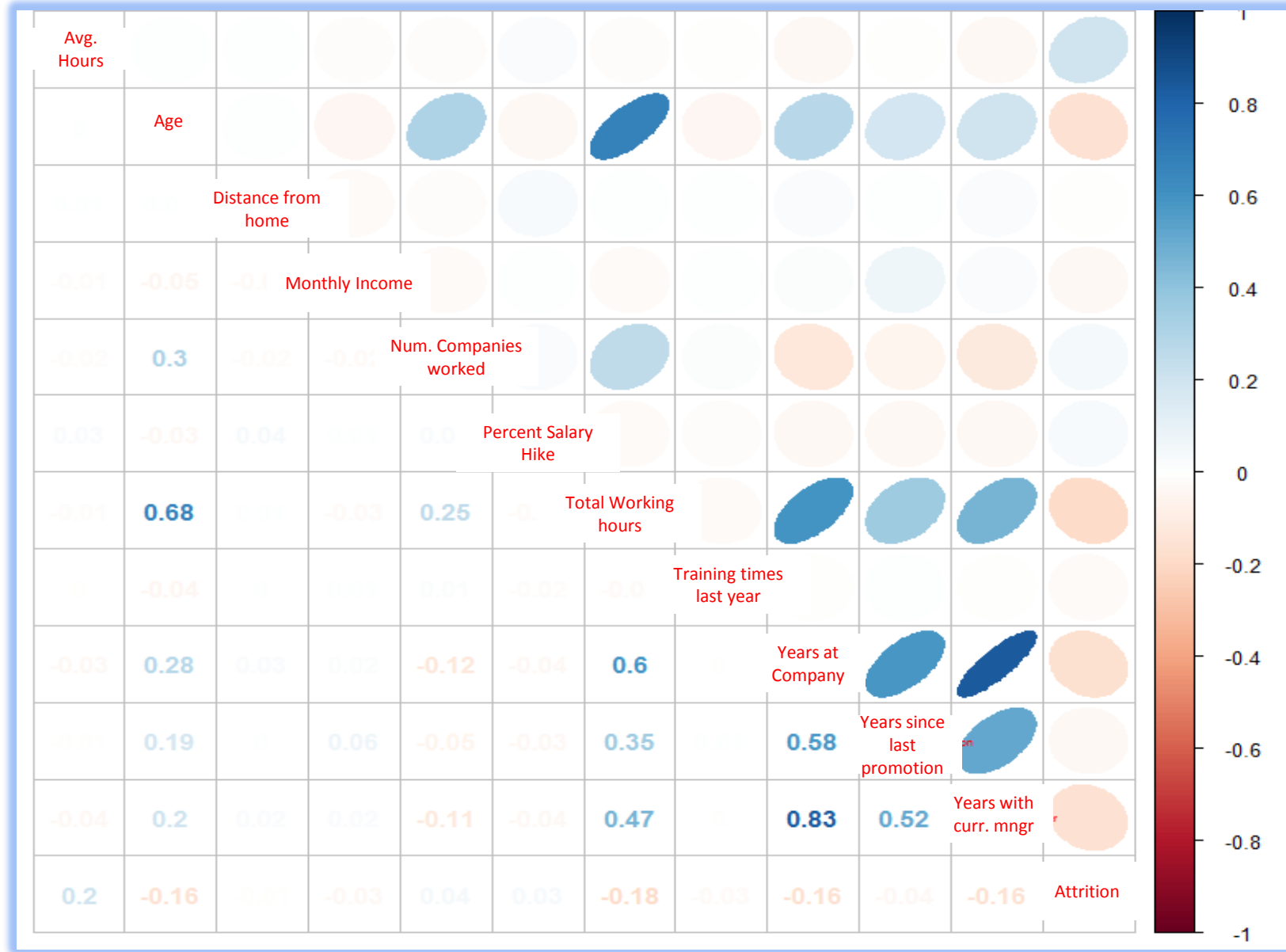
Attrition is higher for the following variables  
 Frequent BusinessTravel

- **Single (Marital Status)**
- **Low Environment Satisfaction**
- **Low Job Satisfaction**
- **Low Work Life balance**
- **Low Job Involvement**

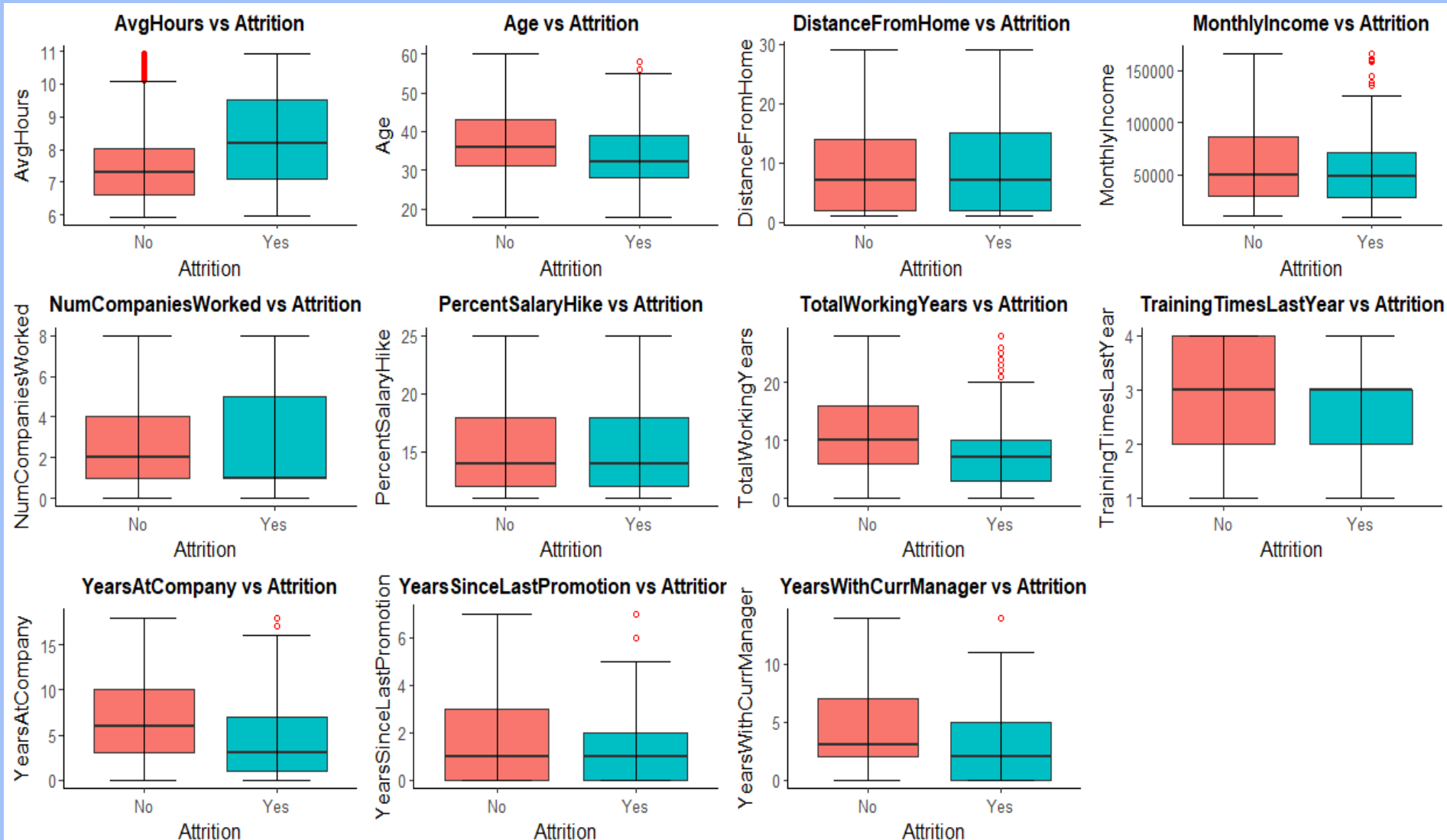


# Multivariate Analysis

## Continuous variables

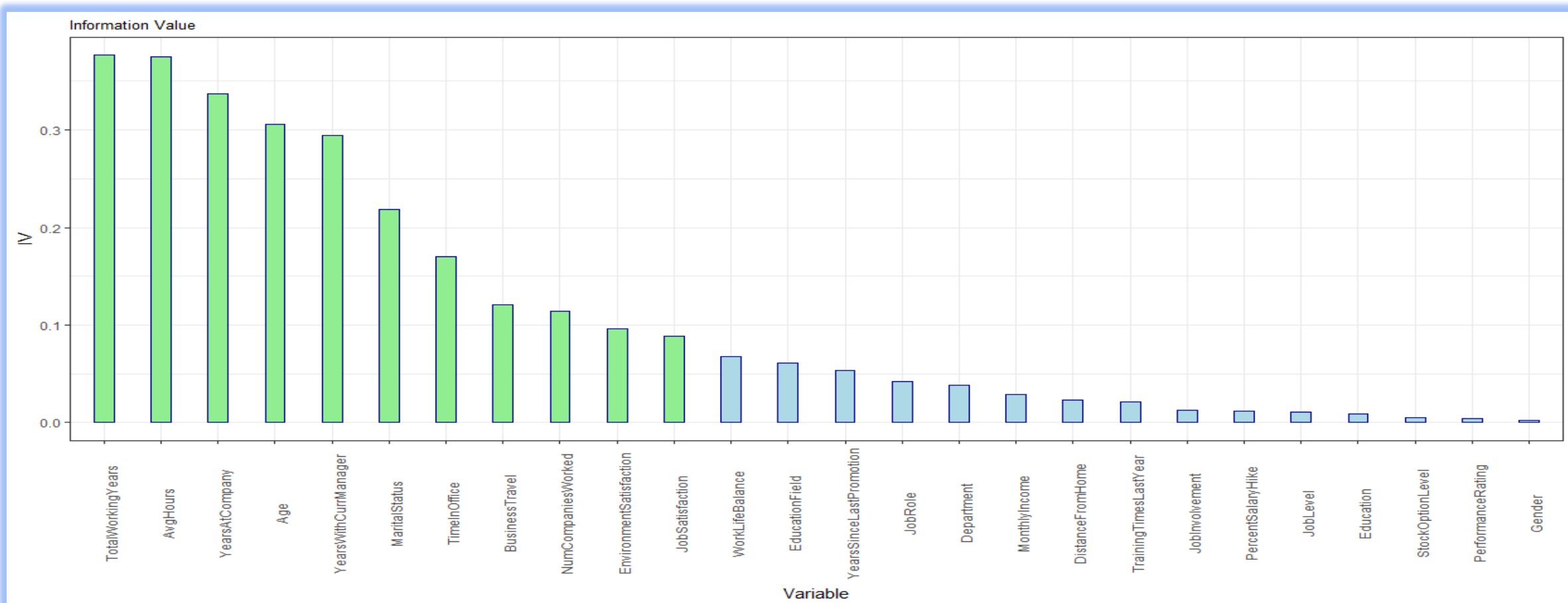


- As expected **Age & total working years** are strongly correlated
- Total working years** shows strong correlation with **YearsAtCompany**
- Attrition shows **+ve correlation** with
  - **Average working hours**
- Attrition shows **-ve correlation** with
  - **Age**
  - **Total Working years**
  - **Years at Company**
  - **Years with current Manager**
  - **Number of companies worked**




Attrition is higher for the following variables  
Frequent BusinessTravel

- **More Average working hours (over time)**
- **Lesser Age**
- **Less number of total working years**
- **Less number of years at company**
- **Less number of years with current manager**
- **Less number of companies worked**

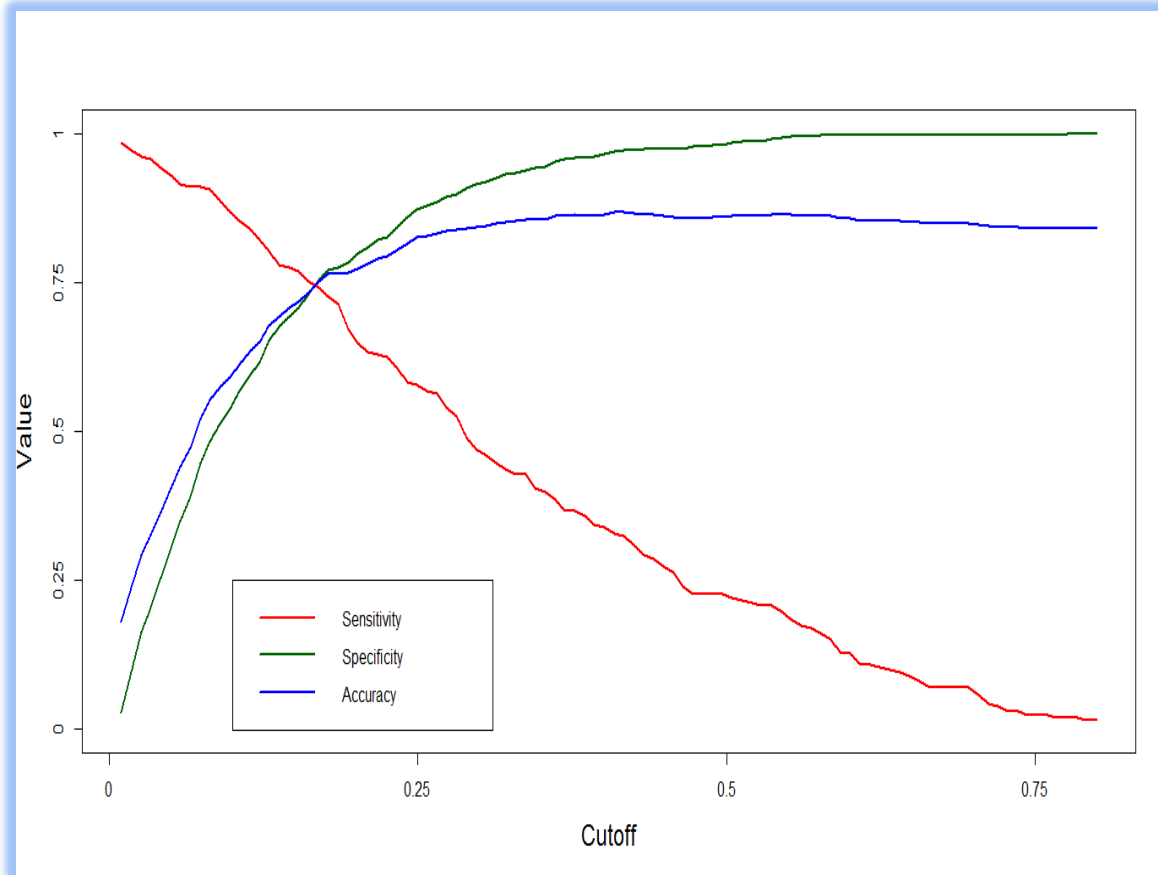


- All those with green bars are variables with higher information value and hence are most significant variables
- These are the same variables that have been identified as significant in the earlier multivariate analysis

 Most Significant  
 Least Significant

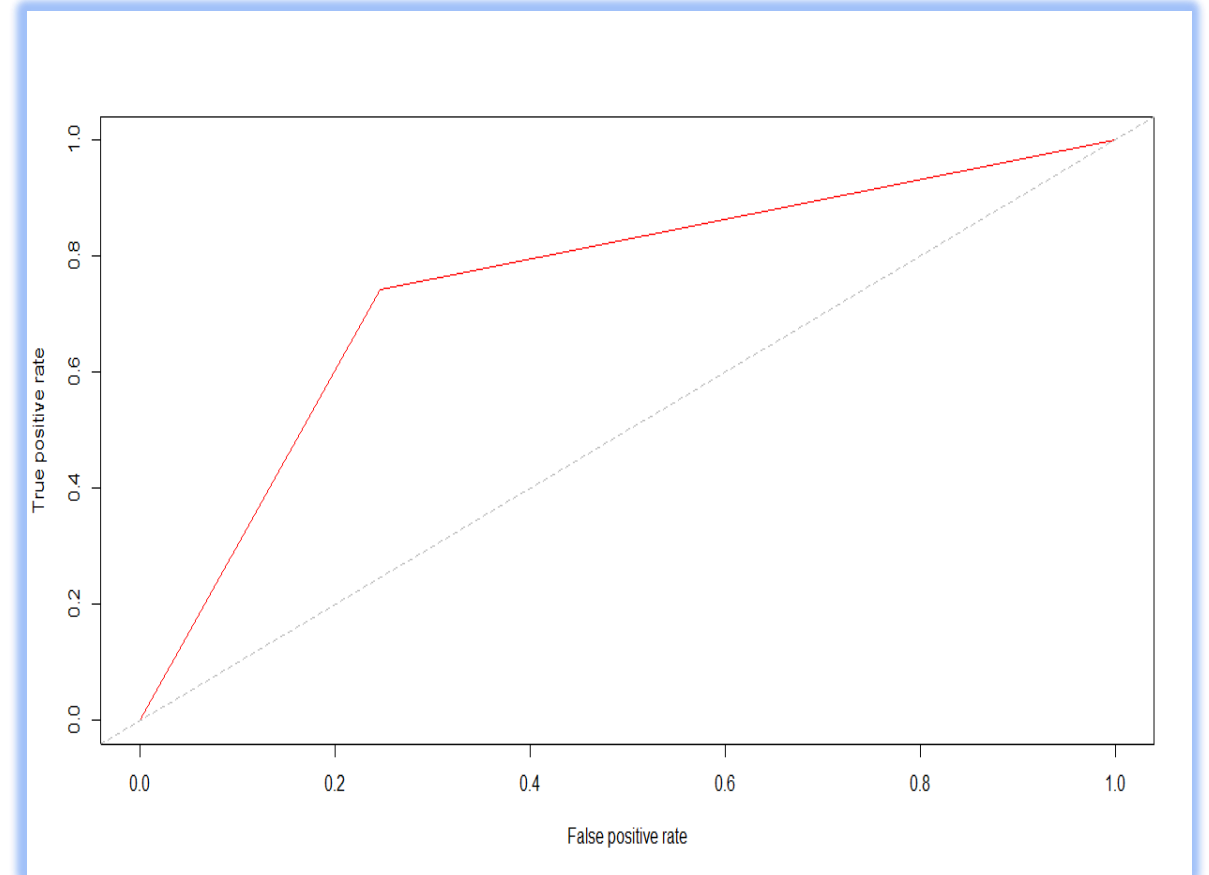
## Steps involved in model preparation:

- The master employee data was divided into **test** & **train** in the ratio of **70%** & **30%** respectively.
- The **Generalized linear model** (glm) algorithm was used to run against the train data.
- The initial model m1 was created and then was run into step AIC procedure to select the best model based on *Akaike Information Criteria* ignoring the least significant variables.
- The model out of step AIC procedure m2 was then examined for variables with less significance (high P value > 0.05) and with High multi collinearity (VIF > 2). Variables with low P value and high multi collinearity were removed one by one building a new model in each iteration.
- A total of 29 model iterations were executed to identify the variables with utmost significance to the model.
- All the final variables selected for the model are Significant i.e P value almost = 0 and negligible & have VIF < 3 indicating no major multi collinearity
- The final best model was identified as **m28** which was then evaluated based on the test dataset



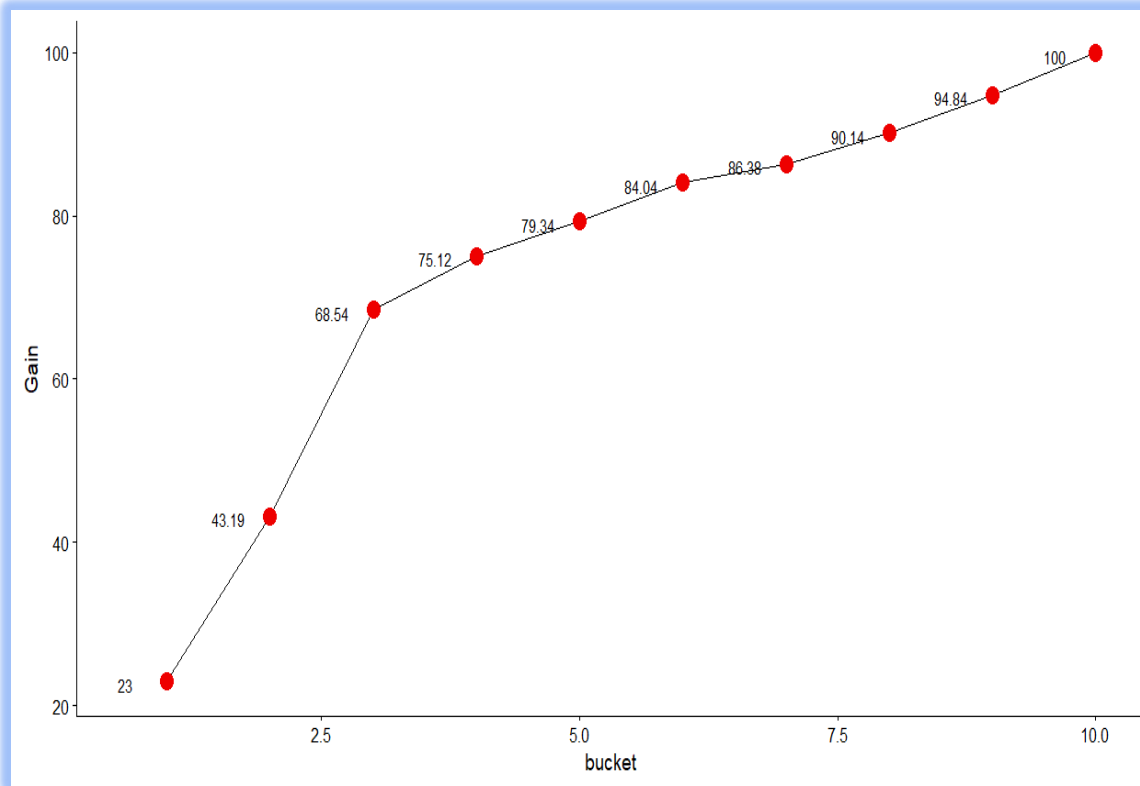
**Accuracy : 75% | Sensitivity : 74% | Specificity : 75%**

**Accuracy** is how accurate the model in predicating the Attrition  
**Sensitivity** of a model is the proportion of yeses (or positives) correctly predicted by the model as yeses (or positives).  
**Specificity** is equal to the proportion of nos (or negatives) correctly predicted by the model as nos (or negatives)



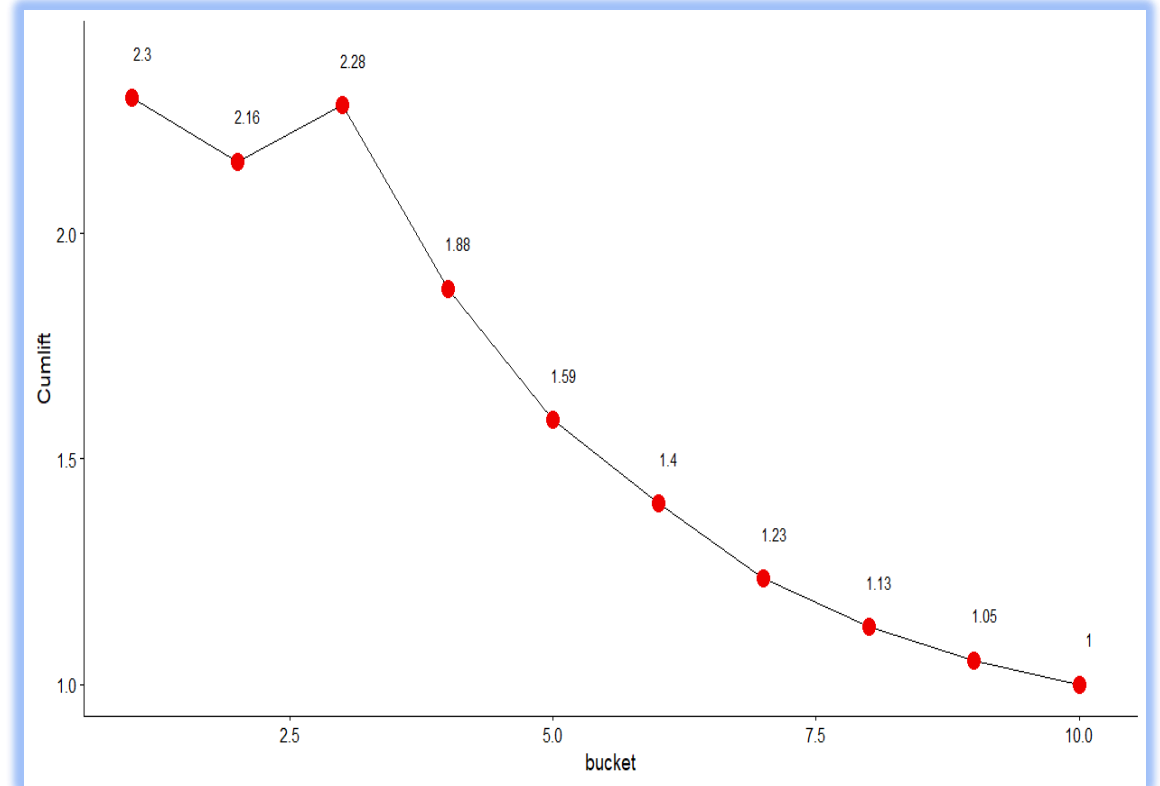
**AOC : 0.75**

**AUC Area under curve** Plotting true positive rate against False positive rate. This is a measure of how well a parameter can distinguish between two diagnostic groups



**Gain : 75% at 4<sup>th</sup> decile**

This means that if we sort all employees according to probability, then among the top 40% customers of this sorted list, we would find 75% of all employees that were likely to leave the company



**Lift : 2.28**

This means that the model's gain by the end of the 3rd decile is 2.28 times that of a random model's gain at the end of 3 deciles. In other words, the model catches 2.28 times more attrition than a random model would have caught.

## Model's Discriminative power metrics

Accuracy	75%
Sensitivity	74%
Specificity	75%
AOC	0.75
KS Statistic	50%
Gain	75% @ 4 <sup>th</sup> Decile
Lift	2.28%

## Summary of factors

The analysis and model results show that the company should focus on employees with the following attributes to curb attrition :

- Frequent Business Travel
- Single (Marital Status)
- Low Environment Satisfaction
- Low Job Satisfaction
- Low Work Life balance
- Higher average working hours
- Lesser Age
- Less Total Working years
- Higher Years since last Promotion
- Less Years with current Manager
- Less Number of companies worked