HR Analytics Case Study

Group Name:

1. Swati Shejwalker
2. Vijaya Lakshmi Potturu
3. Benjamin Turner
4. Somasundaram Balasubramaniyam

Business Objective

Problem Statement: XYZ company has around 4000 employees with 15% employee attrition each year .

Business Challenges:

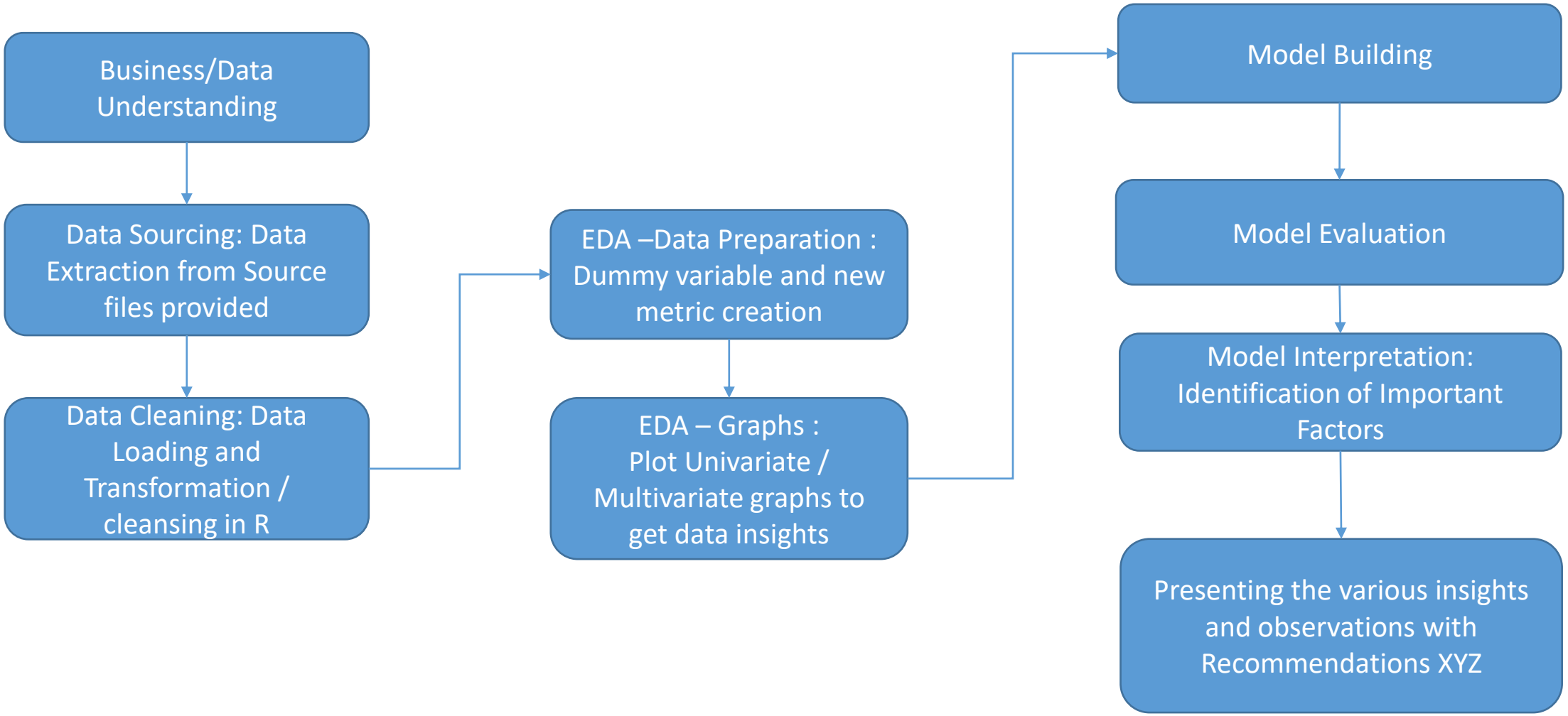
- Need to be find new resources from the talent pool available in the job market
- Projects get delayed, which makes it difficult to meet **timelines** resulting in a reputation loss among consumers and partners.
- Training new employees incurs **additional costs and time**.
- **Recruitment costs**.
- Time involved in acclimatise themselves to the company further delays project time lines.

Objective:

- They want to understand what factors they should focus on, in order to curb attrition.
- They also want to know which of these variables is most important and needs to be addressed right away.
- Overall objective is therefore to model the probability of attrition using a logistic regression.

Constraints : Only 1 year of Employee data has provided

Problem solving methodology



Data Understanding

- There are 5 datasets given Employee ID is Unique in all the files.
- General Employee data – This is master file with employee demographics and professional experience details

•Age
•Attrition
•Business Travel
•Department
•Distance From Home
•Education
•Education Field
•Employee Count
•EmployeeID
•Gender
•Job Level
•Job Role
•Marital Status
•Num Companies Worked
•Over18
•Percent S alary 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 with 24 variables.**

2. Employee Survey data – which provides employee survey data(3 features)

•Employee ID
•Environment Satisfaction
•Job Satisfaction
•Work Life Balance
• **4410 observations with 4 variables.**

3. Manager Survey Data – Performance of an employee under a manager(2 features)

•Employee ID
•Job Involvement
•Performance Rating
•**4410 observations with 3 variables.**

4. Intime Data – Employee check in time (1 year data)
Employee ID and in time details from 01/01/2015 to 31/12/2015

5. Outtime Data – Employee check out time (1 Year data)
Employee ID and outtime details from 01/01/2015to 31/12/2015

Data Sourcing : Assumptions

- Total Working Years contains 9 NA values. These employees were assumed to have started working at age 24 years (Based on median value).
- Num Companies Worked contains 19 NA values. These rows were removed because of lack of clarity as to how values should be calculated and the small number of rows affected.
- 12 columns from in_time.csv and out_time.csv contained only NAs. These were assumed to be public holidays and were removed from the dataset, Other assumption NAs in these files could be absences or Employee Leaves (e.g. due to sickness or annual leave). These columns removed from the dataset.
- 585 records found to be having discrepancy as Num Companies Worked = 0 and difference of Total Working Years and Years At Company is more than 0 (e.g. employee id 2) so that means the person would have worked in at least one company before XYZ company or there may be possibility that Years At Company only counts the years completed by employee and not the months hence the difference, So leave these records as is considering the volume of such records.
- Variable X in in time and out ime represents Employee ID and hence changed the name of column so it can be merged with other datasets
- Variables converted to factors according to the data dictionary descriptions

Data cleaning : Data Quality Checks

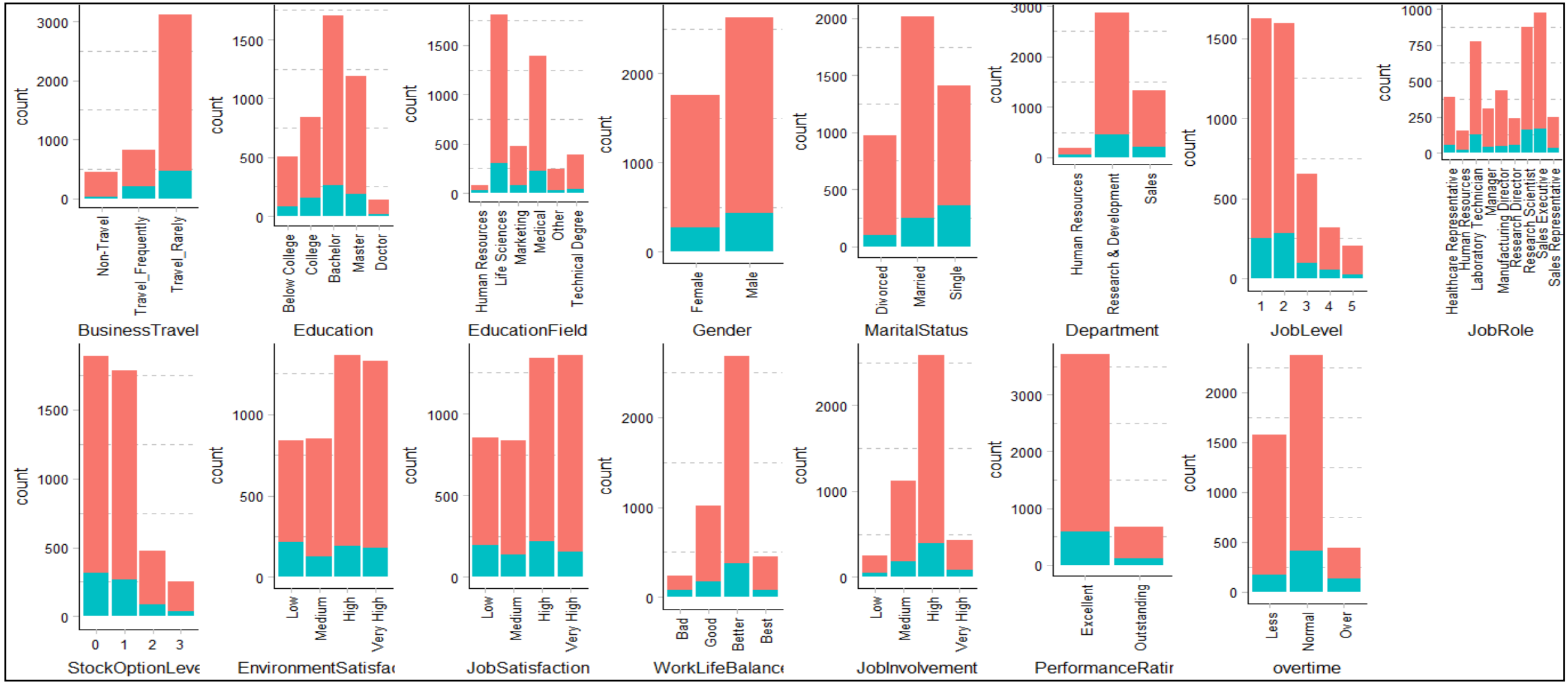
- NAs for Environment Satisfaction, Job Satisfaction, Work Life Balance were replaced with median values.
- Over18, Standard Hours and Employee Count each contained a single value and were not included in the model.
- **New Metrics:** average daily hours ,number of absences and overtime flag calculated for each employee.
- **Outliers:** capped at the level of the 5th and 95th centiles.
- Numeric variables scaled to standardised ranges.
- Dummy variables created for categorical variables.

EDA Summary

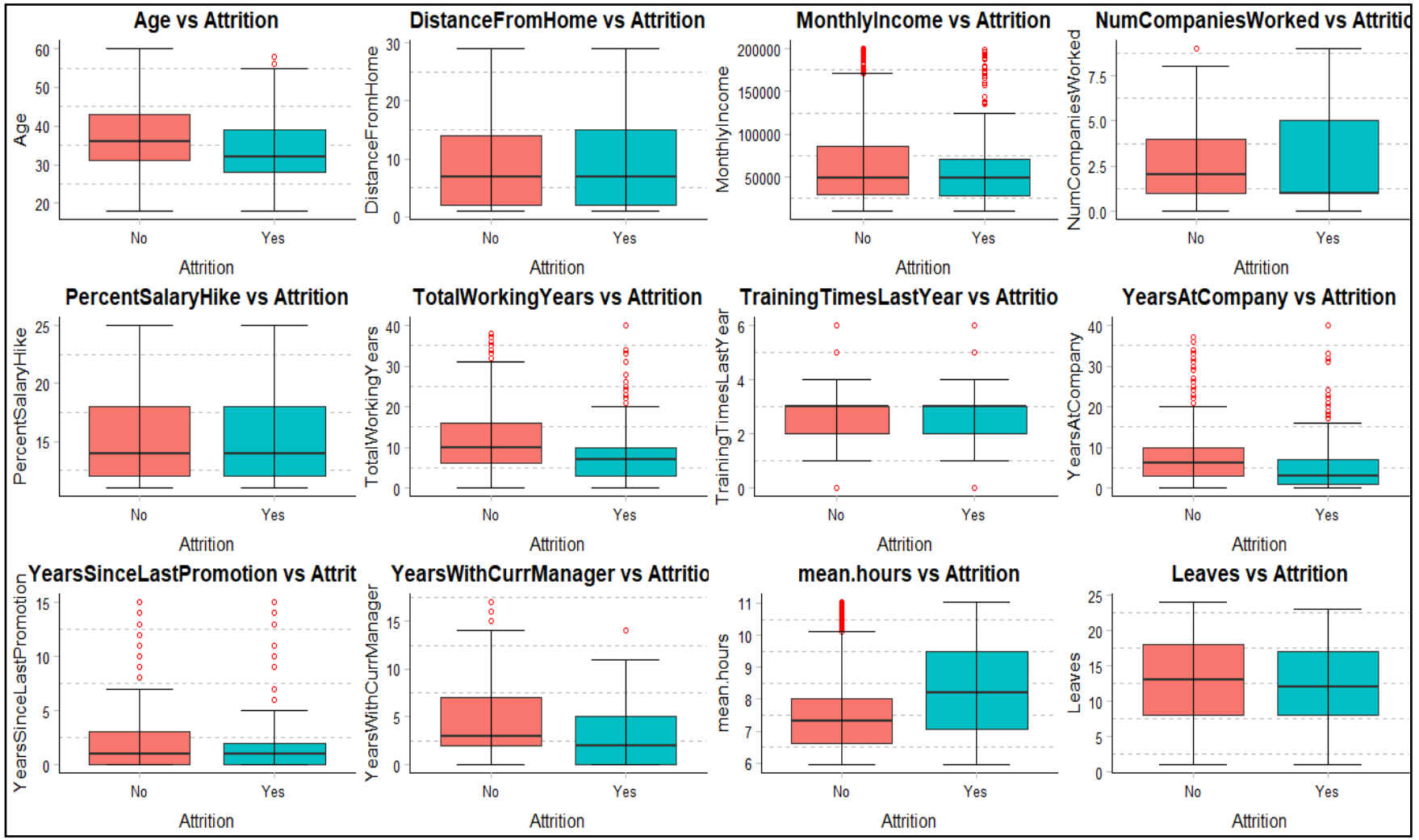
- **Univariate Analysis :**
  - Youngest employees (in their 20s and 30s) are most likely to leave. There is another peak in the 60s, likely due to retirement.
  - Higher attrition amongst employees who have been with their manager <10 years.
  - Attrition rate is higher among employees travelling frequently for business.
  - Attrition rate is higher in Human resources department .
  - Attrition rate is higher in college educated employees.
  - Attrition rate is higher in HR-educated employees in education field.
  - Research directors, research scientist and sales exec have the highest attrition.
  - Single employees more likely to leave, divorced are least likely.
  - Attrition highest for those with low environmental satisfaction.
  - Attrition highest for those with low job satisfaction, low for those with very high satisfaction
  - Attrition highest for those with bad work life balance.
  - Attrition highest for those with Outstanding performers.
  - Attrition highest for those with < 3 times Training Times Last Year
  - Attrition highest for those who worked over time more.

Most Important driving factors : Age, Total Working Years, Work life Balance, Job Satisfaction, Environment Satisfaction, Marital status, Job Role, Education Field, Department, Training Times Last Y ear ,and business Travel

Univariate Analysis - Categorical Variables Vs Attrition



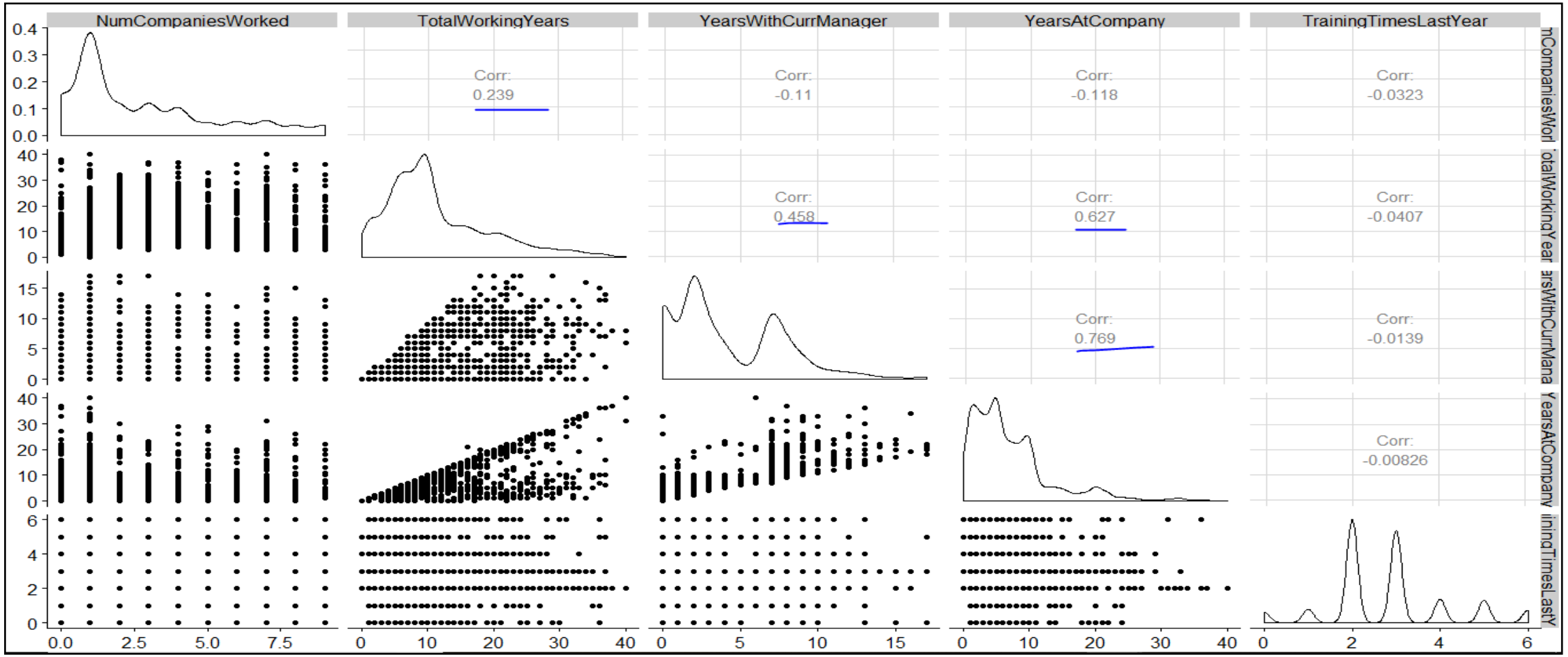
Multivariate Analysis
Continuous Variables Vs Attrition



Attrition is higher for the below variables

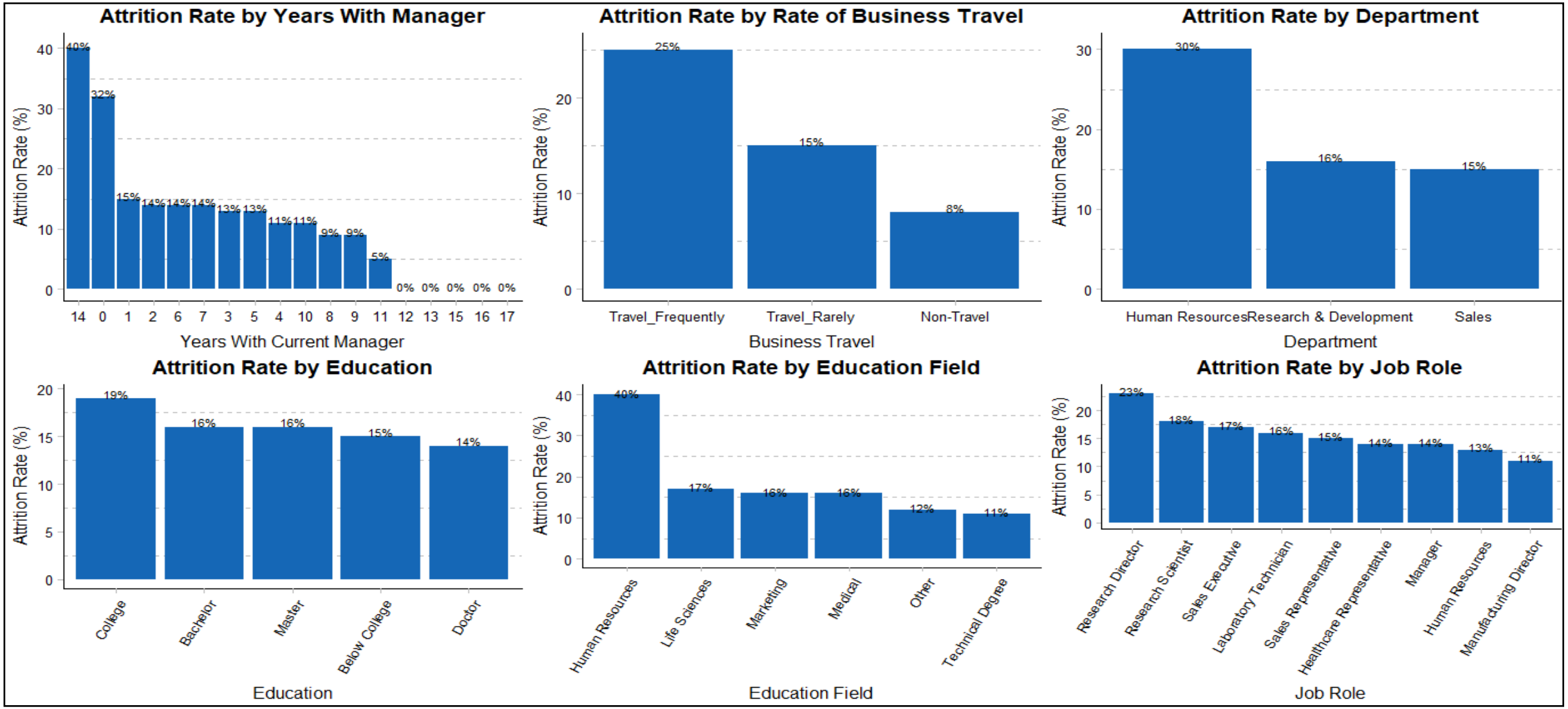
- **Less number of total working years**
- **Less Age**
- **Less number of years at company**
- **Less number of years with same current manager**
- **More with Avg working are high**
- **Less number of companies worked**

Multicollinearity

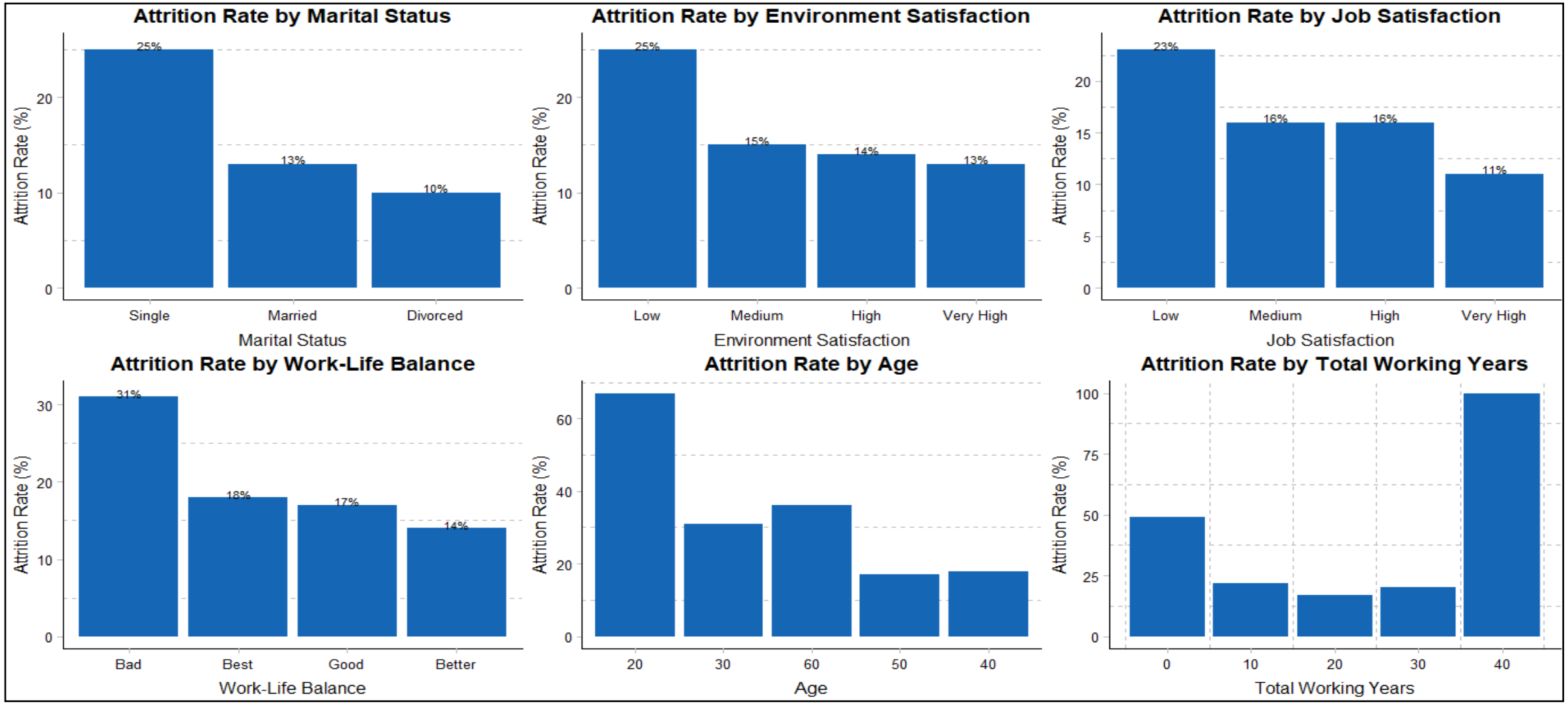


There is high correlation between Total working years, Years With Current Manager, Years At Company,

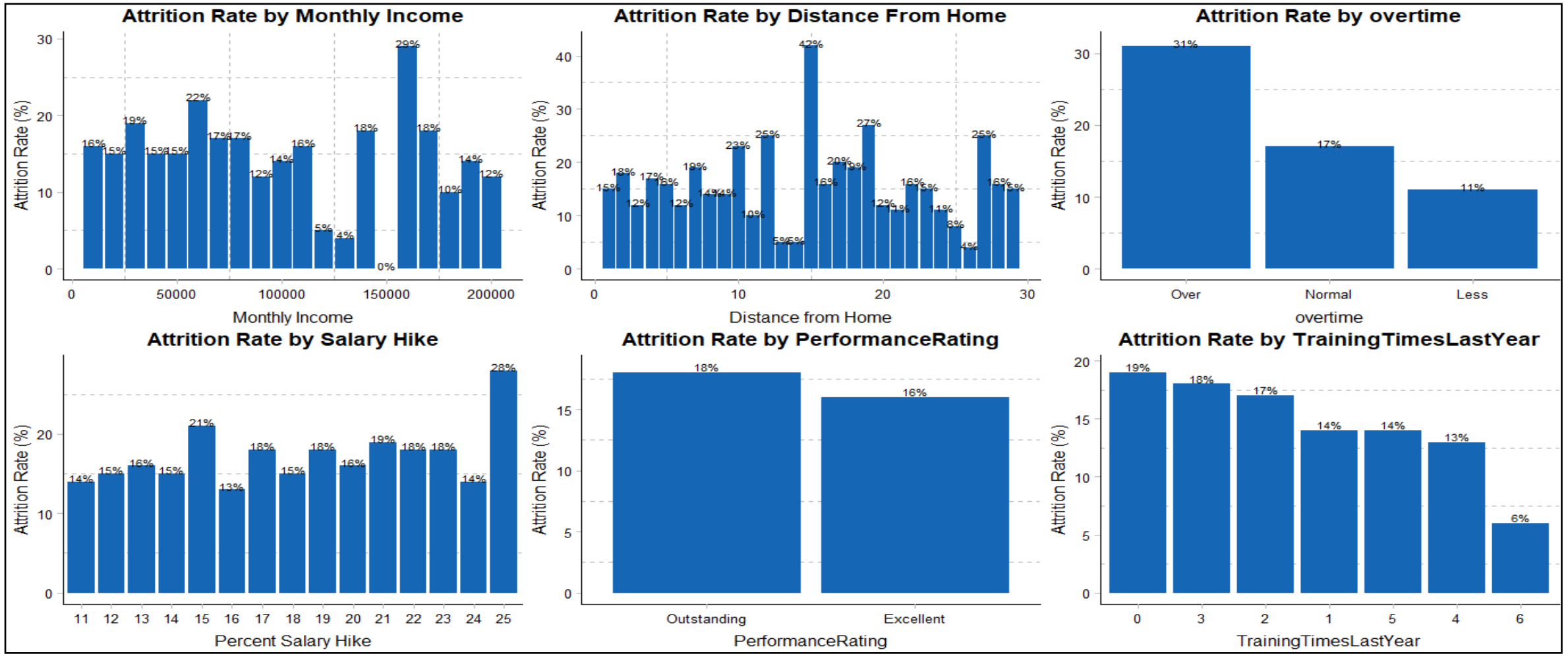
Exploratory Data Analysis -1



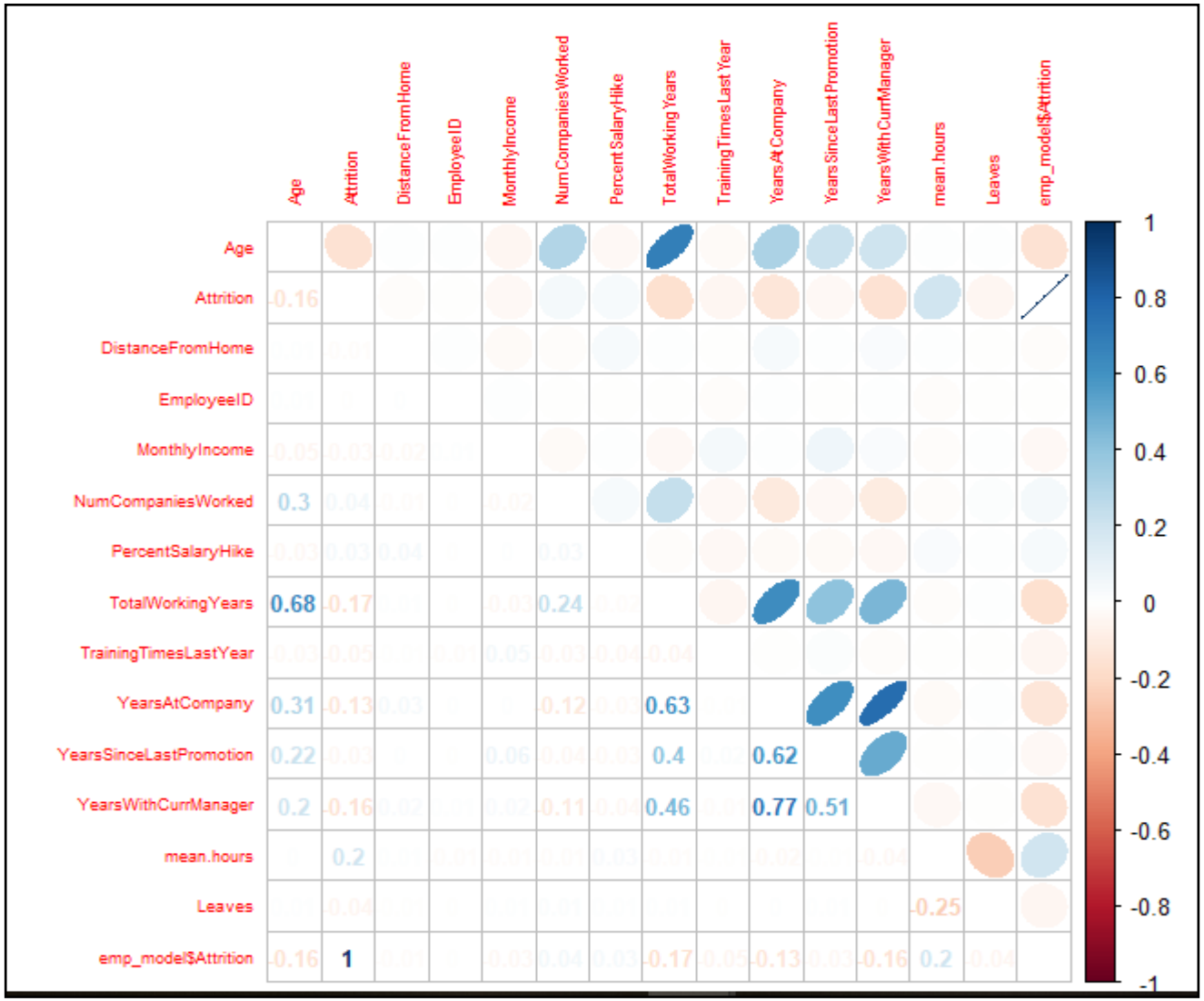
Exploratory Data Analysis -2



Exploratory Data Analysis - 3



Multivariate Analysis



- **Year At Company** and **Years with Current Manager** are highly correlated.
- **Years with Current Manager , total working years, Years since last promotion** are highly correlated.
- As expected **Age** and **Total working years** are strongly correlated.
- **Total working years** shows a strong positive correlation with **Years at company**.
- **Attrition** shows a positive correlation with **mean working hours**.
- **Attrition** shows –ve correlation with :
  - **Age**
  - **Total Working Years**
  - **Years at Company**
  - **Years with current manager**

Model Building.

- Normalised all the continuous variables with scaling.
- Converted all the categorical variables to dummy variable to convert to numerical variables.

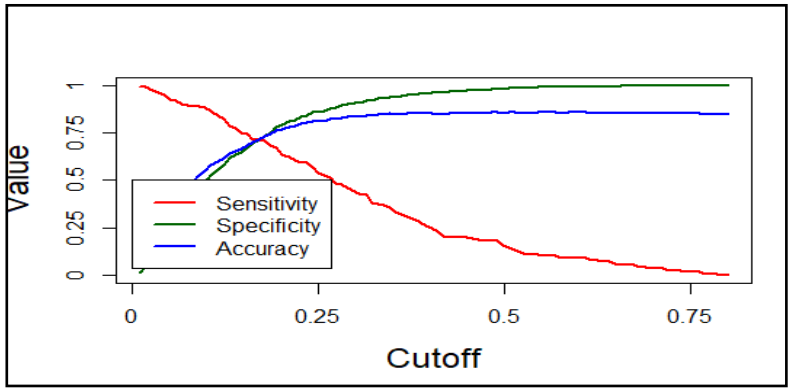
The master employee file is divided in to 70% of data used for training, 30% reserved for testing.

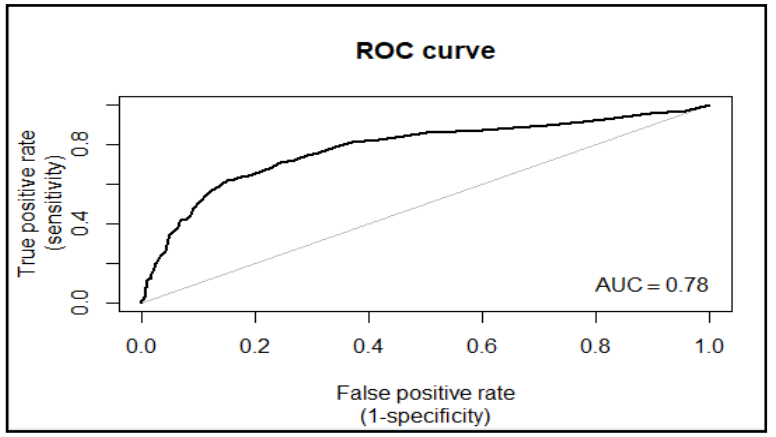
- Initial model created using all variables on the training data with **generalised linear model(glm).**
- stepAIC() function used (stepwise selection) to select the best model based on Akaike information criteria to eliminate some of the variables
- Backward selection used to remove further variables based on significance(P value > 0.05) and VIF >2 .
- We optimised the model to reduce number of variables with ANOVA the final model 10 significant variables.

Model Evaluation

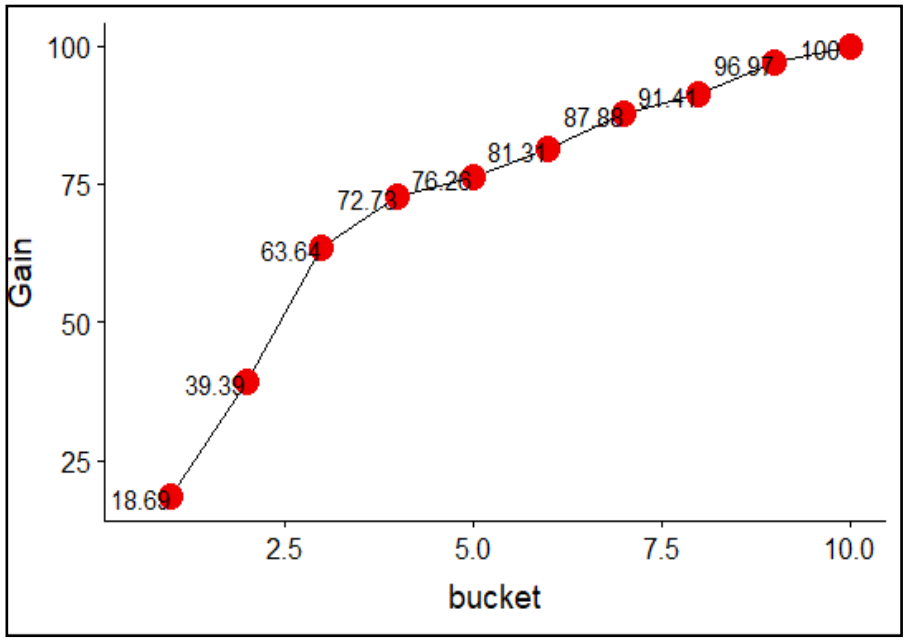
- Generated confusion matrix with different cut-off levels.
- Lift & Gain chart plotted for final model.
- To find the optimum cut-off ,we have verified with Accuracy, Sensitivity, Specificity for 1% to 80% probability values on final model.
- Sensitivity and specificity calculated for each model at the optimum cut-off.
- The optimum Cut off is chosen at 0.169596
- KS statistics calculated.
- ROC Curve plotted.
- Plotted Residual & Q-Q Plot’s , Coefficient Plot ,Residual plots to visualize the model accuracy
-

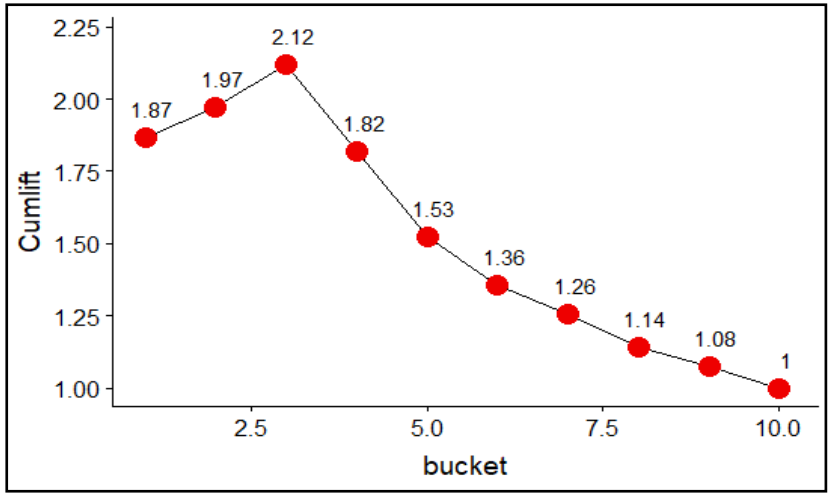
Model Evaluation: Gain & Lift charts





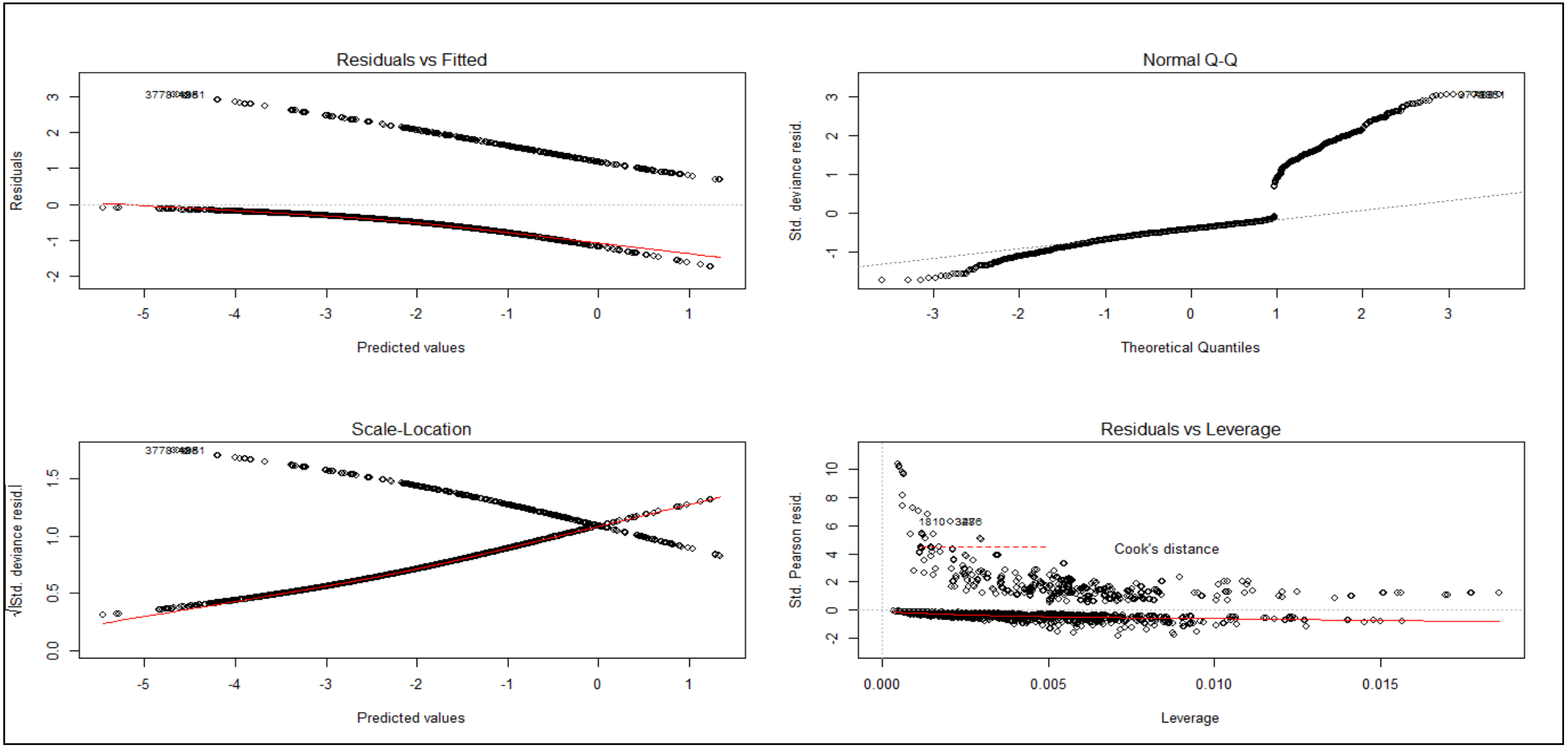
Confusion Matrix			
		test_actual	
		No	Yes
test_pred	No	802	57
	Yes	312	141



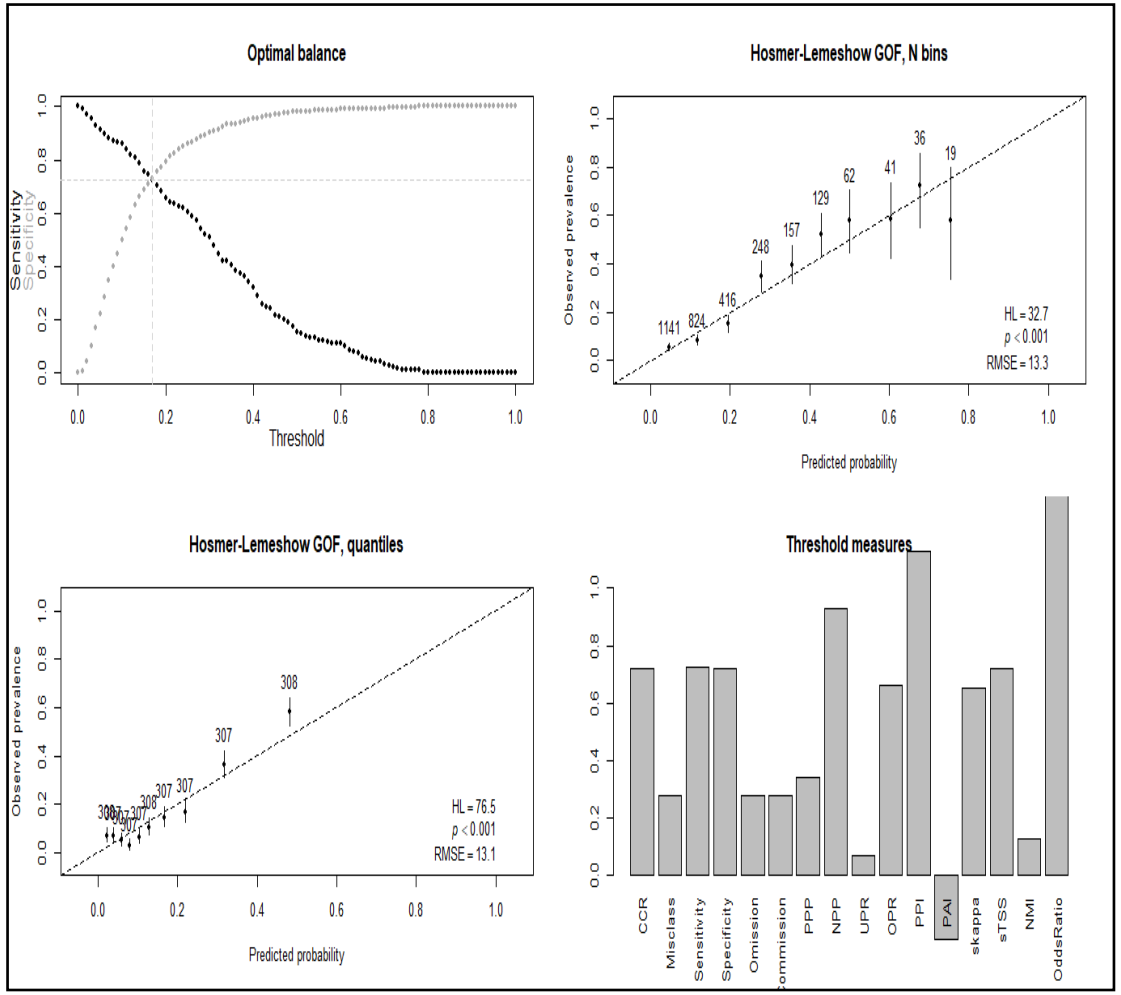


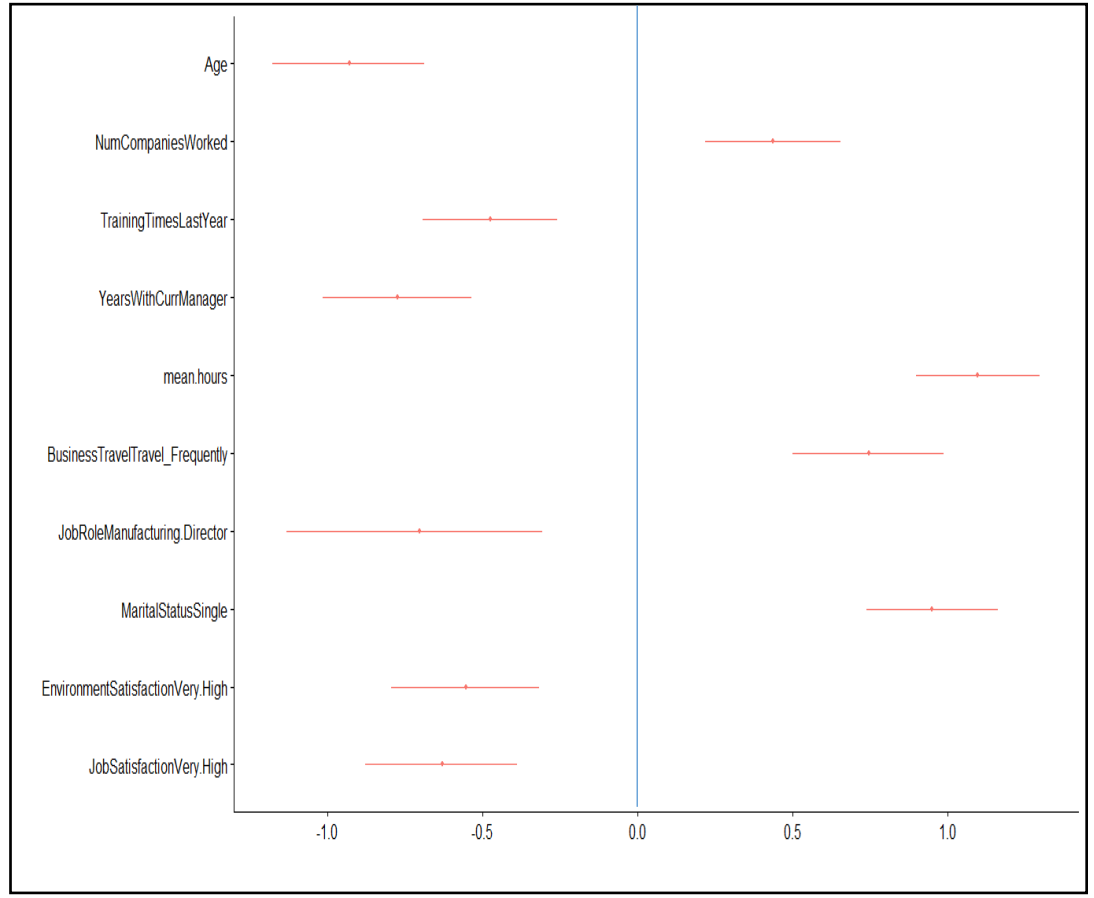
Model Statistics	
Accuracy	0.7140963
Sensitivity	0.7121212
Specificity	0.7160714
Cut-off	0.169596
K statistic	0.4281926
Gain	72.2 at 4th deciles
Lift	2.12 %
AUC	0.7527552

Final Model Residual & Q-Q Plot's



Model Evaluation (coefficients & Thresholds)





Conclusion - Model Interpretation Model 32

Variable	Coefficient	Interpretation
Age	-0.46173	younger employees are more likely to leave.
NumCompaniesWorked	0.21871	Employees with many previous companies are more likely to leave.
TrainingTimesLastYear	-0.23732	Employees taking more training are less likely to leave.
YearsWithCurrentManager	-0.38914	Employees that have been under the same manager for a long period are less likely to leave.
Mean.hours	0.54652	Employees working more hours are more likely to leave.
BusinessTravelTravel_Frequently	0.74532	Employees who travel frequently are more likely to leave.
JobRoleManufacturing.Director	-0.70223	Employees with the job role Manufacturing.Director are less likely to leave.
MaritalStatusSingle	0.95056	Single employees more likely to leave than married or divorced.
EnvironmentSatisfactionVery.High	-0.55117	Employees with very high environment satisfaction are less likely to leave.
JobSatisfactionVery.High	-0.62901	Employees with very.high job satisfaction are less likely to leave.

Variables in **bold** are those with the highest magnitude coefficients and therefore largest impact on attrition.

Recommendation

- Business Travel is an important factors to influence attrition so company needs to plan more virtualization and remote handling to avoid frequent travel.
- One of the highly influential factor is over time hours which employee is putting in office ,company should plan proper project plans which is required by management .This will improve employee motivation factors with work life balance to curb the attrition.
- Promotions has also come out as an important factor to influence the attrition. Promotions and rewarding should be priority for the company to improve employee career level motivation to reduce attrition.
- Changing of managers very frequently will impact performance ratings and promotion of an employee so need to focus on this area.
- Job satisfaction, Environmental satisfaction & work life balance are obvious reasons to influence attrition.HR team should enhance employee engagement related organisation events, flexible working hours, team outings etc..
- Employees who are taking more training are less likely to leave. So management should encourage employees to attend more trainings and up skill which will improve their satisfaction.