HR ANALYTICS CASE STUDY

SUBMISSION

Group Name:

1. Swati Kamat
2. Saurabh Kumar Singh
3. Rounak Gurjar
4. Suraj Kumar Talreja

Objective/Goals

Background

- A large company named XYZ, employs around 4000 employees.
- However, around 15% of the employees leave the company and that is a big concern for the company as it has to be replaced with the talent pool available in the job market.
- The company has contracted an HR analytics firm and want to understand what factors they should focus on, in order to curb attrition.

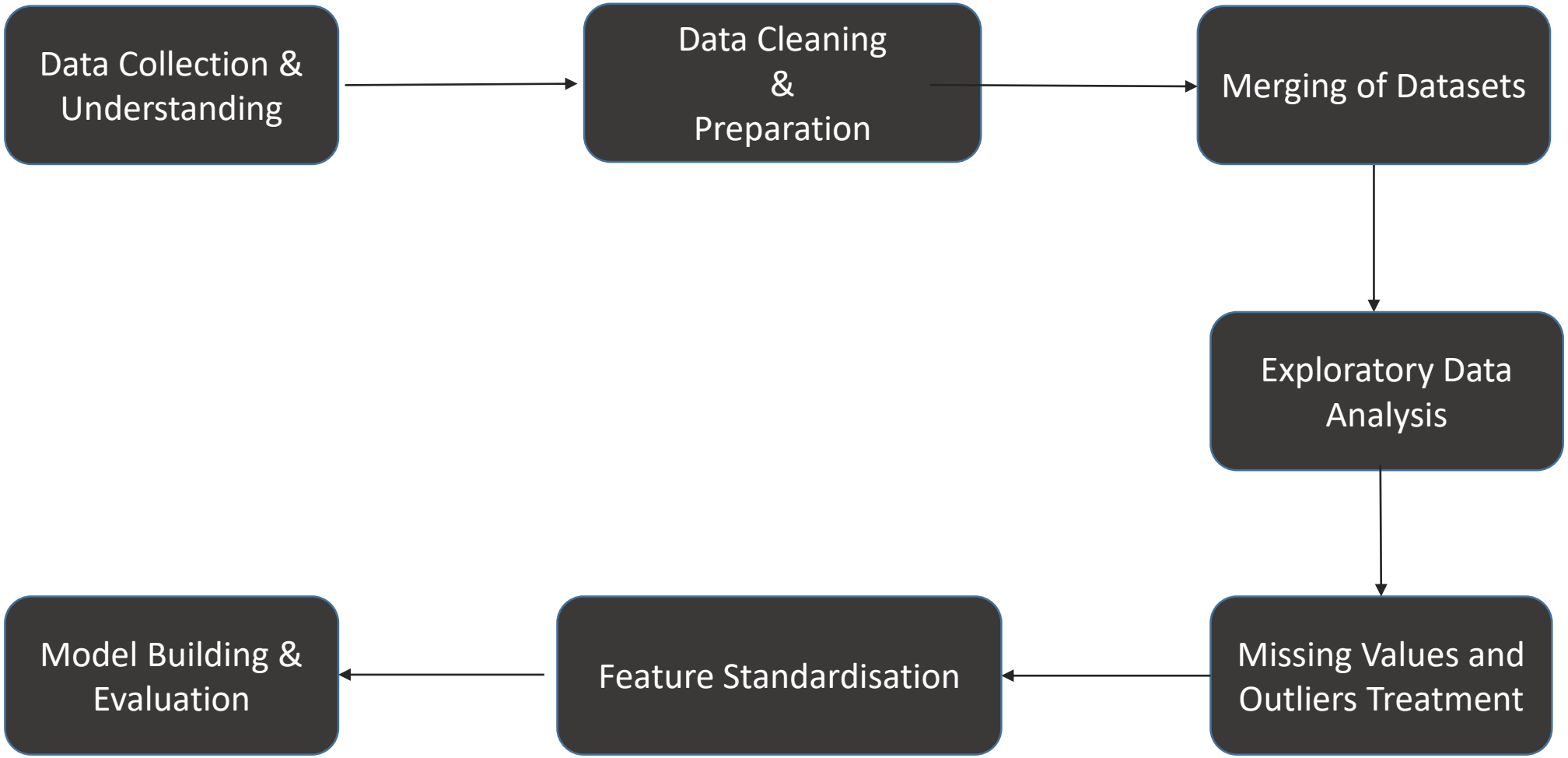
Objective

- To understand factors on which company should focus and what changes they should make to their workplace, to curb attrition.
- To model the probability of attrition using a logistic regression.

Strategy

- Perform Data cleaning & Preparation
- Perform Exploratory Data Analysis
- Model building & evaluation, identify the factors company should focus on.

Problem Solving Methodology



Data Collection & Understanding

Employee Survey Data
EmployeeID
Employee Satisfaction
JobSatisfaction
WorkLifeBalance

GENERAL DATA	
AGE	Attrition
BusinessTravel	Department
DistanceFromHome	Education
EducationField	EmployeeCount
EmployeeID	Gender
JobLevel	JobRole
MaritalStatus	MonthlyIncome
NumCompaniesWorked	Over18
PercentSalaryHike	StandardHours
StockOptionlevel	TotalWorkingYears
TrainingTimesLastYear	YearsAtCompany
YearsSinceLastPromotion	YearsWithCurrManager

Manager Survey Data
EmployeeID
JobInvolvement
PerformanceRating

In Time of Employee
Day wise Intime of Employees for the year 2015

Out Time of Employee
Day wise out time of Employees for the year 2015

Data is available in 5 Data sets as described above

Data Cleaning & Preparation

General Data Set

- Remove Employee Count, Over18 & Standard Hours from data frame as have same values for all rows.
- Created categorical variables of following numerical variables as per data dictionary and business understanding.

Education		DistanceFromHome		Age	
1	Below College	Within 2 Kms	Proximity	Below 30	Youth
2	College	2 to 7 Kms	Near	Between 30 & 42	Mid-Age
3	Bachelor	7 to 15 Kms	Far	Above 40	Seniors
4	Master	More than 15 Kms	VeryFar		
5	Doctor				

- Replace NA values of NumCompaniesWorked & TotalWorkingYears with mode of column.

Employee Survey Data Set

- Replace NA values of EnvironmentSatisfaction, JobSatisfaction & WorkLifeBalance with mode of column.
- Created categorical variables of following numerical variables as per data dictionary.

EnvironmentSatisfaction		JobSatisfaction		WorkLifeBalance	
1	Low	1	Low	1	Bad
2	Medium	2	Medium	2	Good
3	High	3	High	3	Better
4	Very High	4	Very High	4	Best

Data Cleaning & Preparation

Manager Survey Dataset

- Created categorical variables of following numerical variables as per data dictionary.

JobInvolvement		PerformanceRating	
1	Low	1	Low
2	Medium	2	Good
3	High	3	Excellent
4	Very High	4	Outstanding

In and Out Time dataset

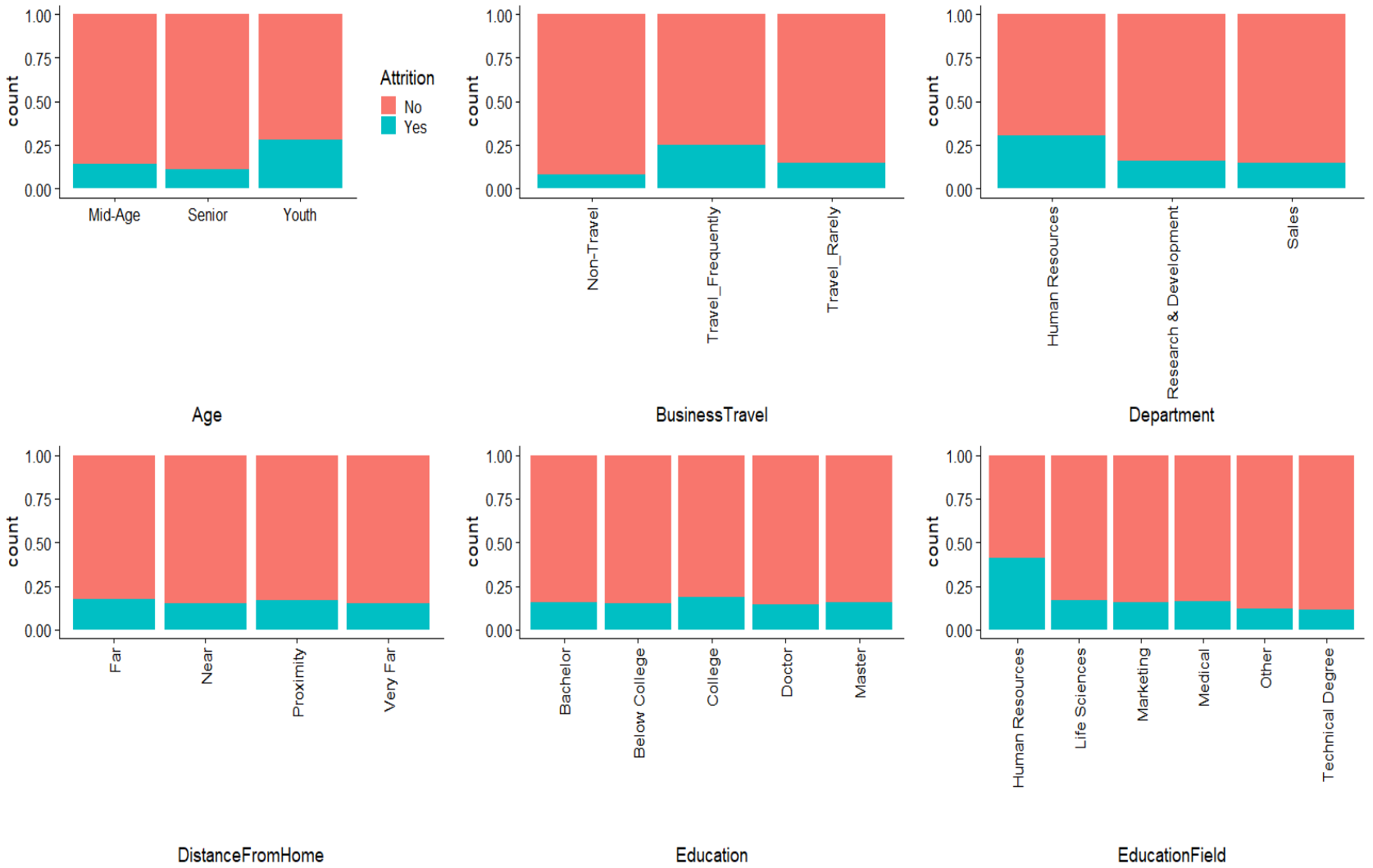
- Calculate the difference between in and out time of employees using POSIXct function.
- Perform rollup operation and calculate Average working hours, Extra Hours & Leaves taken by employees and make data frame of these three columns with employee ID.

Merging of Datasets

- Merge General, Employee Survey, Manager Survey and data frame created from In and Out data sets.

Exploratory Data Analysis

Categorical Variables

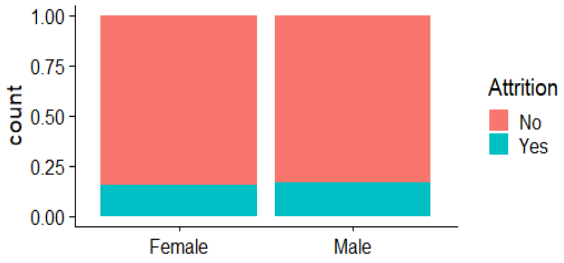


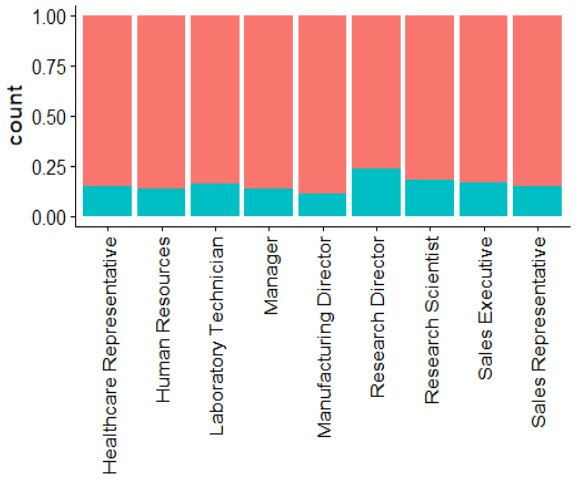
Following category of employees have high attrition rate :

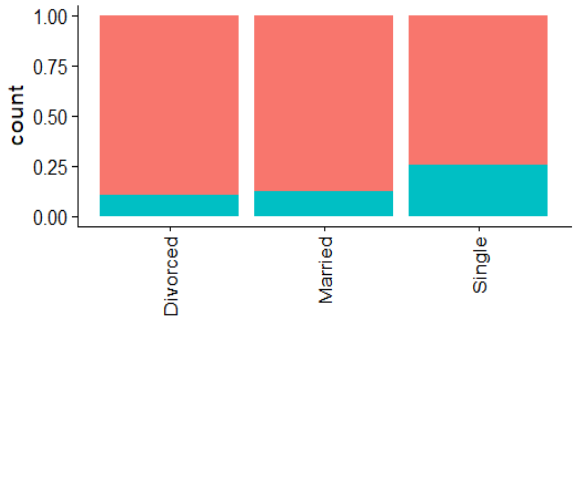
- Youth (Age < 30)
- Employees who travel frequently for business purpose
- Employees from HR Department
- Employees from HR education field

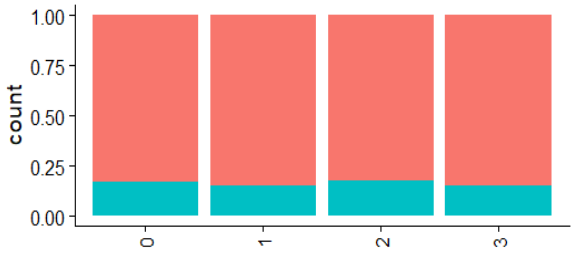
Exploratory Data Analysis

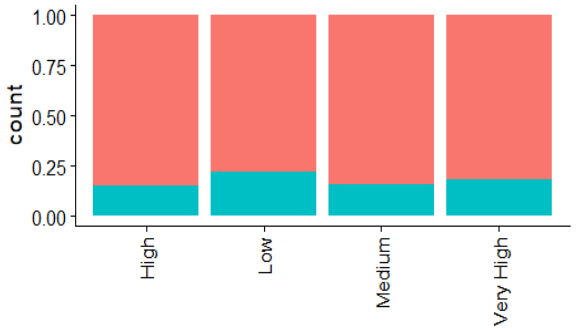
Categorical Variables

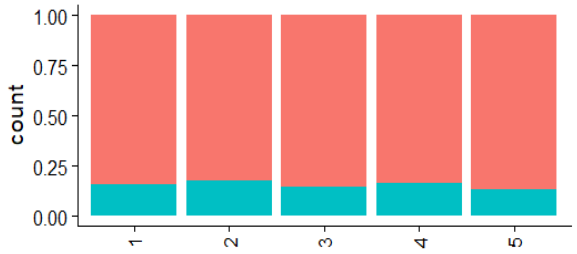










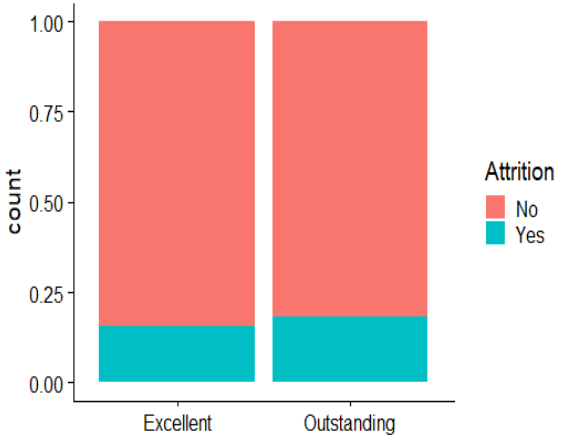


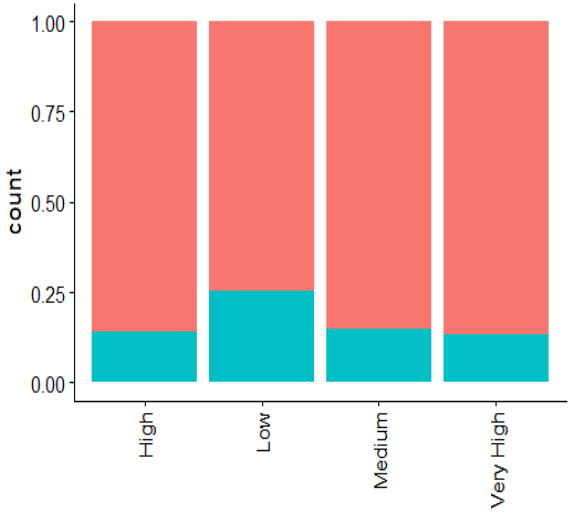
Following category of employees have high attrition rate :

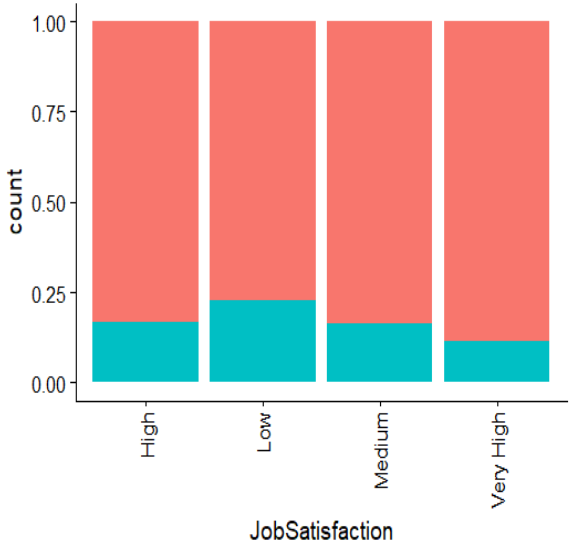
- Employees who have job role as Research Director
- Employees who are unmarried.
- Employees who have low job involvement

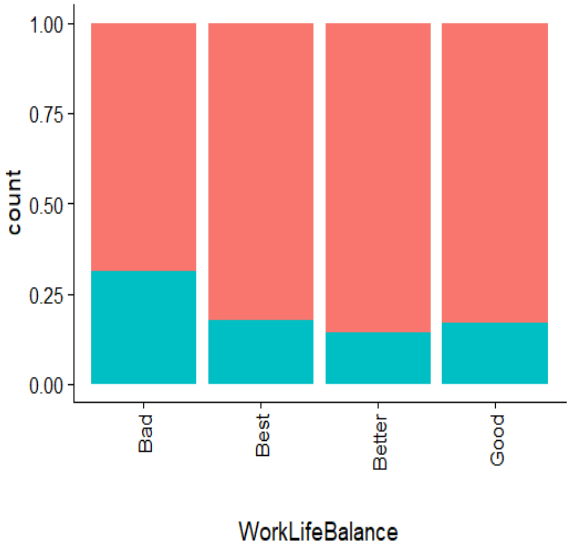
Exploratory Data Analysis

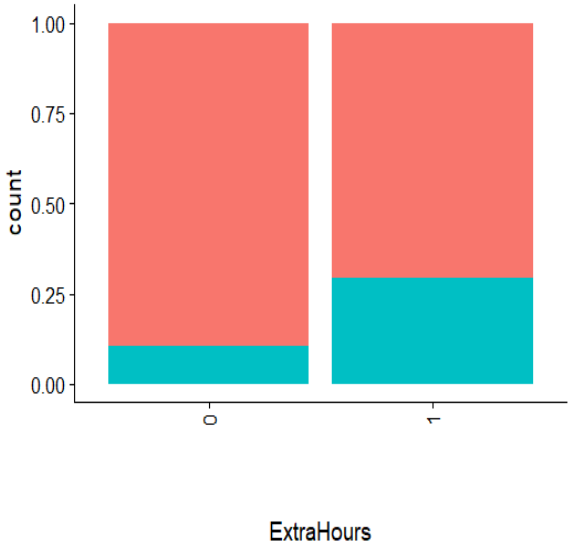
Categorical Variables











Following category of employees have high attrition rate :

- Employees who have Low Environment Satisfaction
- Employees who have Low Job Satisfaction
- Employees who have Bad Work life balance
- Employees who are working extra hours

Missing Values and Outliers Treatment

Missing Values treated in **General Data Set**

- NumCompaniesWorked , TotalWorkingYears, EnvironmentSatisfaction, JobSatisfaction & WorkLifeBalance
- Replacing NA's with the mode of columns.

Outliers treatment (for numerical variables) :

- Identify the Outliers (using boxplot & histogram)
- Capping the Outliers that lie outside the 1.5 * *IQR* limits, such that
  - replacing those observations outside the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile.
- Columns MonthlyIncome, NumCompaniesWorked, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsSinceLastPromotion, Average Hours and YearsWithCurrManager contains outliers and treated.

Feature Standardisation

- Changed Attrition column from categorical to binary.
- Normalized following continuous columns using Z-Score standardization :
  - MonthlyIncome
  - NumCompaniesWorked
  - PercentSalaryHike
  - TotalWorkingYears
  - TrainingTimesLastYear
  - YearsAtCompany
  - YearsSinceLastPromotion
  - YearsWithCurrManager
  - AverageHours
  - Leaves
- Created dummy variables of categorical variables.

Model Building

Following steps taken to build Logistic Regression Model :

- Split the Data set into train(70%) and test(30%).
- Start model building with all the variables, use stepAIC to eliminate all the insignificant variables having high multicollinearity.
- Start eliminating variables one by one having high VIF value (VIF > 2) and low significance.
- Build the final model having all significant variables and low multicollinearity.

Final model have 10 significant variables as given below :

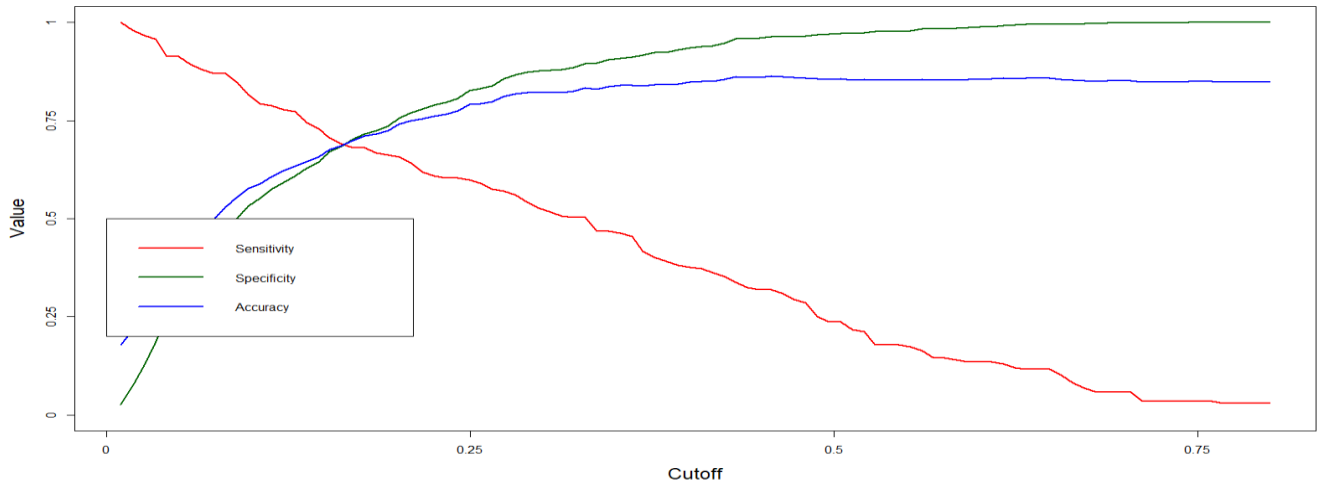
```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.41465    0.11293 -21.383  < 2e-16 ***
NumCompaniesWorked               0.38333    0.05707   6.716 1.86e-11 ***
TotalWorkingYears               -0.59033    0.08512  -6.935 4.06e-12 ***
YearsSinceLastPromotion          0.48853    0.07118   6.863 6.75e-12 ***
YearsWithCurrManager            -0.41775    0.08184  -5.104 3.32e-07 ***
Age.xYouth                       0.71707    0.13424   5.342 9.21e-08 ***
JobRole.xManufacturing.Director -0.82987    0.21704  -3.824 0.000132 ***
MaritalStatus.xMarried          -0.39279    0.11066  -3.550 0.000386 ***
EnvironmentSatisfaction.xLow     0.95720    0.12848   7.450 9.31e-14 ***
JobSatisfaction.xVery.High      -0.79082    0.12829  -6.164 7.07e-10 ***
ExtraHours                       1.64734    0.11318  14.556  < 2e-16 ***
```

Model Evaluation

Final Model (**Cutoff probability – 0.16**)

Confusion Matrix

	REFERENCE	
PREDICTION	No	Yes
No	764	64
Yes	352	143

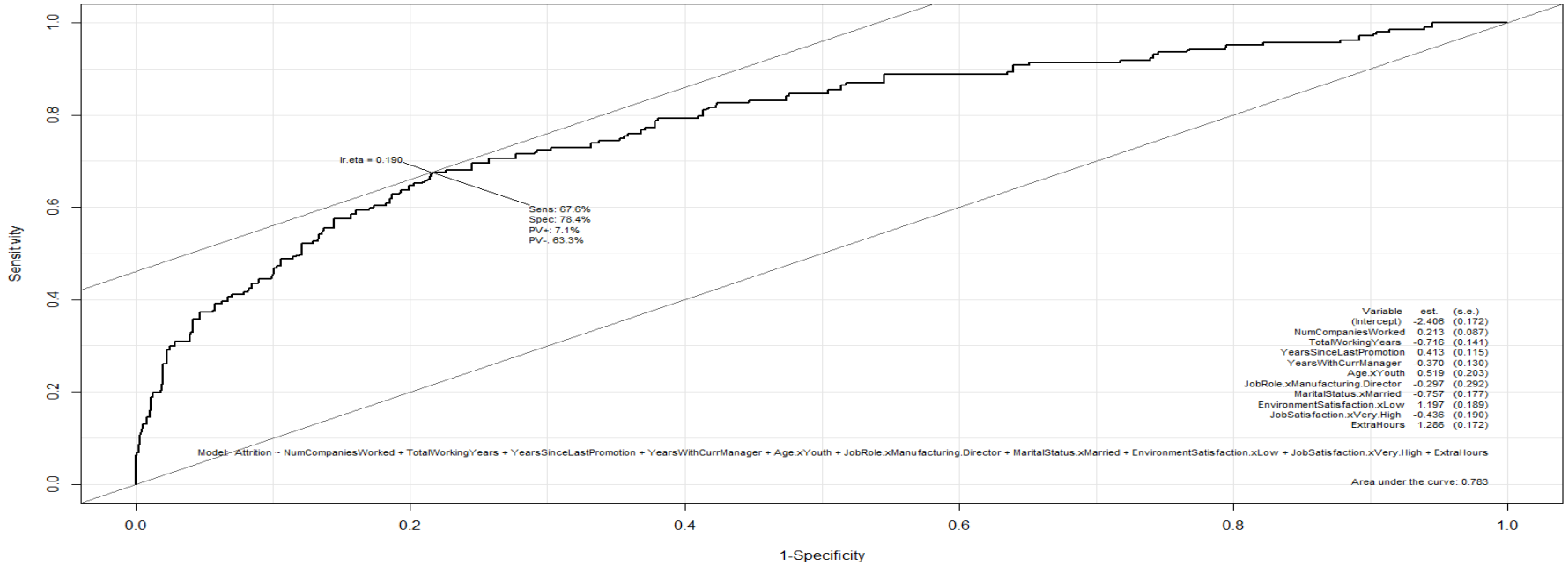


Accuracy -> 0.6855631 , Sensitivity -> 0.6908213 ,

Specificity -> 0.6845878

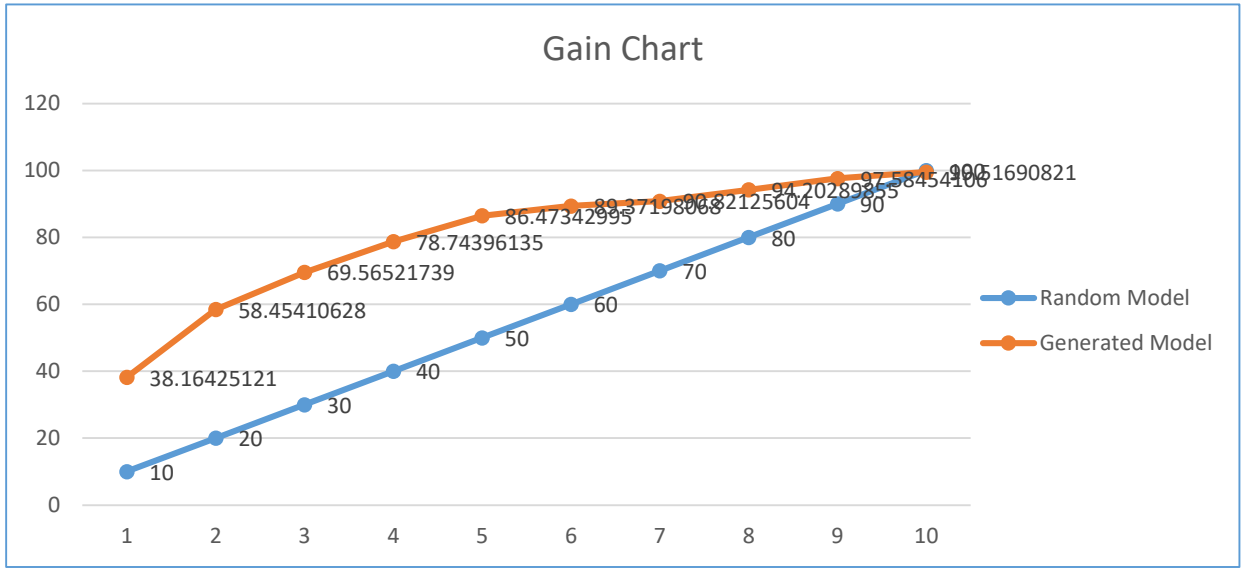
Area Under ROC curve –> 0.783

This is a good Model



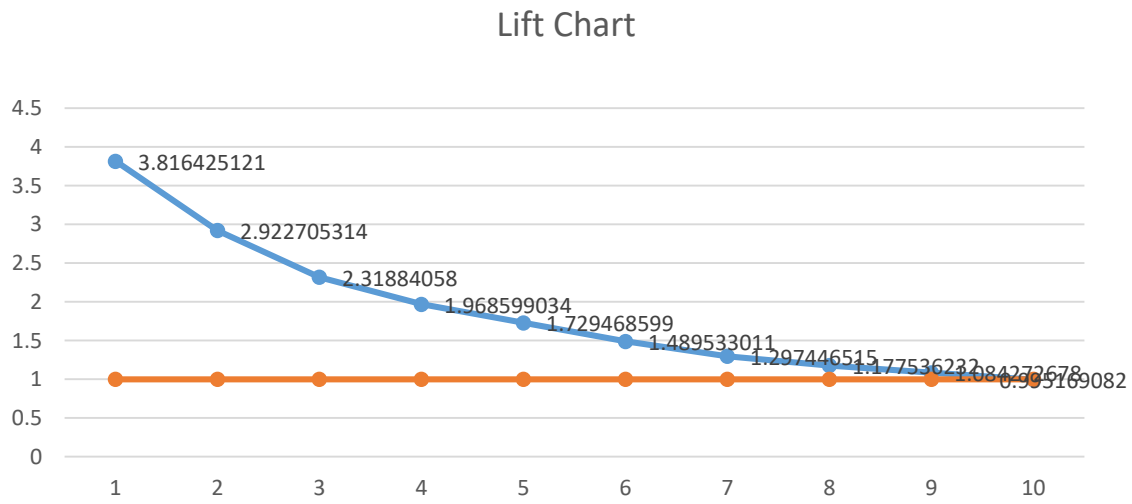
Gain and Lift charts for Generated Model

bucket	total	totalresp	Cumresp	Gain	Cumlift
1	132	79	79	38.16425	3.816425
2	131	42	121	58.45411	2.922705
3	132	23	144	69.56522	2.318841
4	131	19	163	78.74396	1.968599
5	132	16	179	86.47343	1.729469
6	131	6	185	89.37198	1.489533
7	132	3	188	90.82126	1.297447
8	131	7	195	94.2029	1.177536
9	132	7	202	97.58454	1.084273
10	131	4	206	99.51691	0.995169



Based on the Gain values of this model,

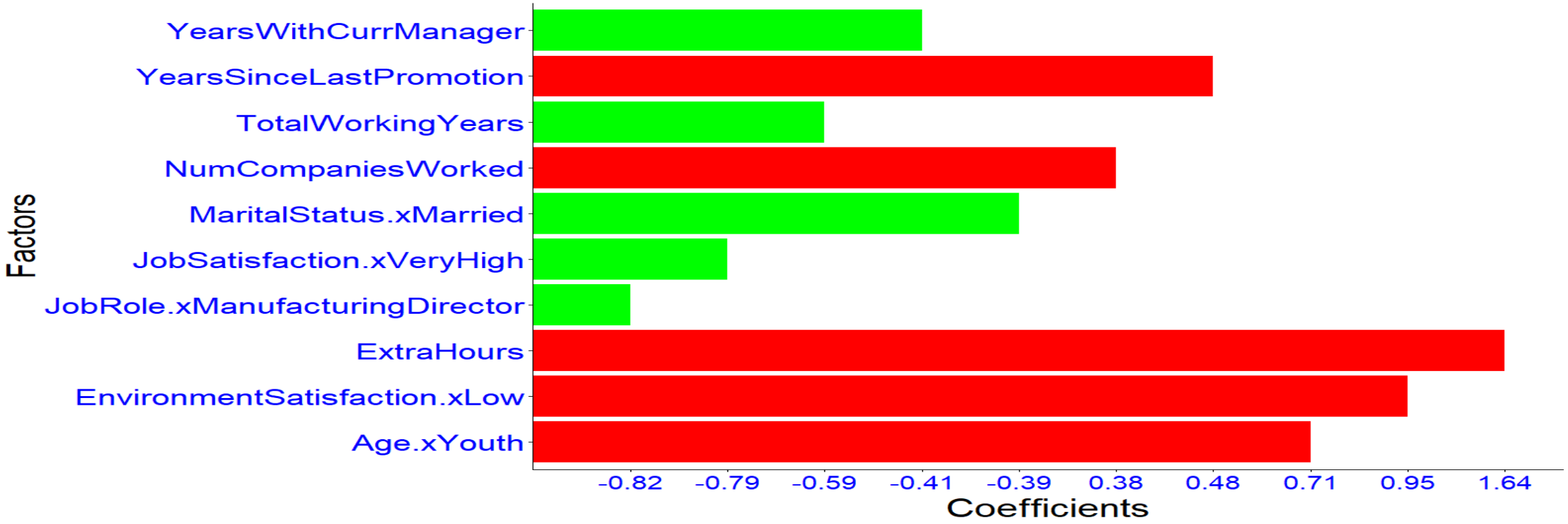
If we choose this model, sort all the Employees according to Probability and contact top 30% in this sorted list, we will be able to catch 69.56% of the Employees that are going to leave the organization.



Considering the 3rd Decile,

Lift indicates the generated model is able to address 2.3 times of the Employees as compared to random model.

SIGNIFICANT FACTORS CONTRIBUTING TO EMPLOYEE ATTRITION



FACTORS (LEADING TO EMPLOYEE ATTRITION) : ExtraHours, EnvironmentSatisfaction.xLow, Age.xYouth, YearsSinceLastPromotion, NumCompaniesWorked

SAFE FACTORS (LEADING TO EMPLOYEE RETENTION) : MaritalStatus.xMarried, YearsWithCurrManager, TotalWorkingYears, JobSatisfaction.xVeryHigh, JobRole.xManufacturingDirector

RECOMMENDATION

FACTORS	INFERENCES	SUGGESTION
Extra Hours	More an Employee works extra hours on average, more are the chances of him/her leaving the company	Company should ensure that project estimations are given correctly, so that there is no pressure of delivery at last moment. Proper Knowledge Transfer within the team to load balance
Low Environment Satisfaction	Lesser Environment Satisfaction is the main feature for Employee to leave	Conducive, clean and safe work environment. Hygiene in Canteen. Motivation in terms of coupons/vouchers and recognition
Age < 30 (Youth)	Employees with age less than 30 are more tend to leave the organization. With the increase in age, they tend to adapt to environment	Provide challenging opportunities to the youth. Reward the senior members.
YearsSinceLastPromotion	The more is the gap between promotions, more is the tendency of Employee to leave	Even if promotions can not be given, there must be the policies to motivate employees so that they feel connected and can relate their work to organizations growth.

RECOMMENDATION – CONTD.

FACTORS	INFERENCES	SUGGESTION
NumCompaniesWorked	More the Number of companies Employee has worked, more is the tendency of him quitting	Company should understand Employees aspirations. For such employees, at the hiring time, Company should check proper reasons for switching the organizations
MaritalStatus.xMarried	Those who are married, less likely to quit	Married employees get settled. For singles, company should locate them near to their native place
YearsWithCurrManager	If Employee works with same manager for longer time, he/she is likely to quit	Company should have behavioral trainings, so that managers are approachable and a healthy relationship with Employees
TotalWorkingYears	People with more experience are less likely to quit	Company should utilize the Experience
JobSatisfaction.xVeryHigh	People with Very high Job satisfaction are least likely to quit	Mechanism to share the feedback. Promote the positive ones.
JobRole.xManufacturingDirector	Employees with Job role as Manufacturing director are least likely to quit	Company should have authorities and responsibilities defined at every job role. This gives clarity and vision.