“HR DATA ANALYTICS CASE STUDY”

Submission

Aoushnik Aich(**Ooch0DbAN)**

Problem statement:

A large company named **XYZ**, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of **attrition** (employees leaving, either on their own or because they got fired) is bad for the company:

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

Goal of the case study:

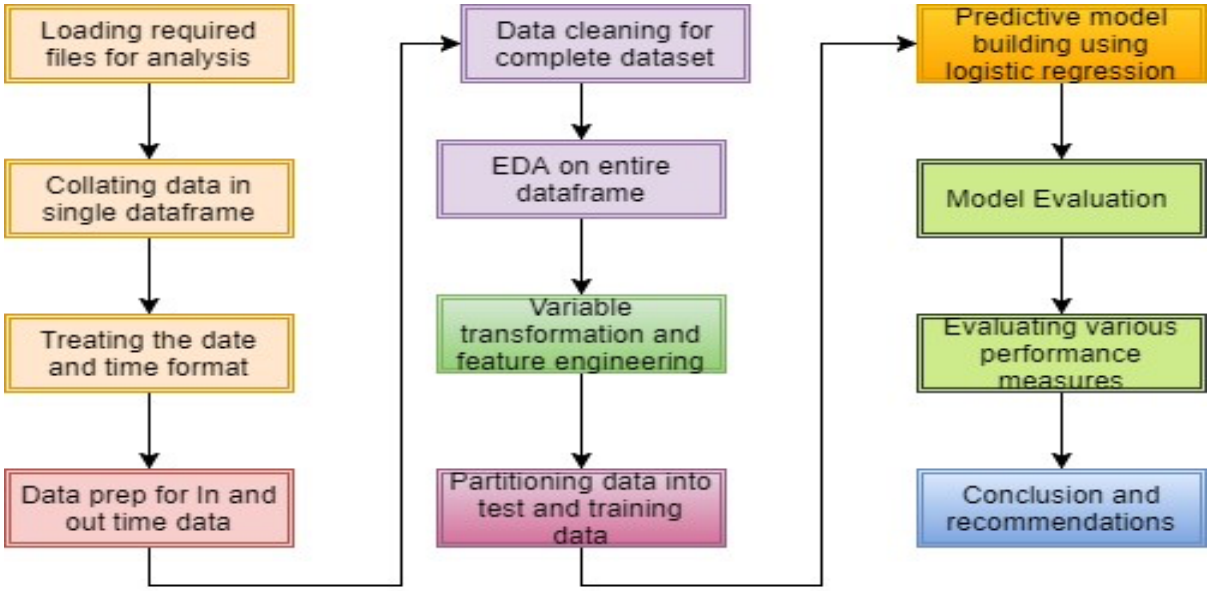
▪ To model the **probability of attrition** using a logistic regression. The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

Data available for the analysis

There are 4410 observations and 41 variables after the Data preparation phase. The list of variables are as below.

```
> names(employee1)
 [1] "Age"                    "Attrition"              "BusinessTravel"
 [4] "Department"             "DistanceFromHome"       "Education"
 [7] "EducationField"         "Gender"                 "JobLevel"
[10] "JobRole"                "MaritalStatus"          "MonthlyIncome"
[13] "NumCompaniesWorked"     "PercentSalaryHike"      "StockOptionLevel"
[16] "TotalWorkingYears"      "TrainingTimesLastYear"  "YearsAtCompany"
[19] "YearsSinceLastPromotion" "YearsWithCurrManager"  "EnvironmentSatisfaction"
[22] "JobSatisfaction"        "WorkLifeBalance"        "JobInvolvement"
[25] "PerformanceRating"      "Jan _hours"             "Feb _hours"
[28] "Mar _hours"             "Apr _hours"             "May _hours"
[31] "Jun _hours"             "Jul _hours"             "Aug _hours"
[34] "Sep _hours"             "Oct _hours"             "Nov _hours"
[37] "Dec _hours"             "no_of_leaves"           "tenureperjob"
[40] "yrswithoutchange"       "comparatio_by_jobrole"
```

Problem Solving Methodology:



Approach for solving case study:

We are breaking down our analysis in following ways:

1.Collate the given files after verifying EmployeeID is the primary key in all the datasets.

2. Since the In time and Out time Data contains the login/logout time stamps of the employees, we first treated this data and aggregated the number of working hours on a monthly basis for all the 12 months and also generated a derived metric called no_of_leaves.

3. We then did the second level of data cleaning on the entire dataset and stored it in a data frame named “employee1”

4. Performed the EDA(Univariate/Bivariate and correlation matrix) on the employee1 data frame to get insights about the data and did the feature engineering (derived metrics) relevant to problem domain 5. Performed the variable transformation(i.e. scaling continuous variables and creating dummy variables)

6. Arrived at the final model iterating through the logistic regression using the glm algorithm.

7. Performed the model evaluation on the test data and determined the optimal cut-off based on the predicted and actual test attrition rates.

8. In addition to the accuracy, specificity and sensitivity performance measures, also used the other performance measures and plots like the AUC,ROC,KS statistic.

9. Generated the lift and gain table and determined the position of the decile which contains the maximum KS statistic.

Points to be noted…

1.For the NA imputation:

For imputing the categorical variables: We replaced the NA's with the mode, i.e. the value that appears most time. In absence of any other information, #this provides the most likely answer

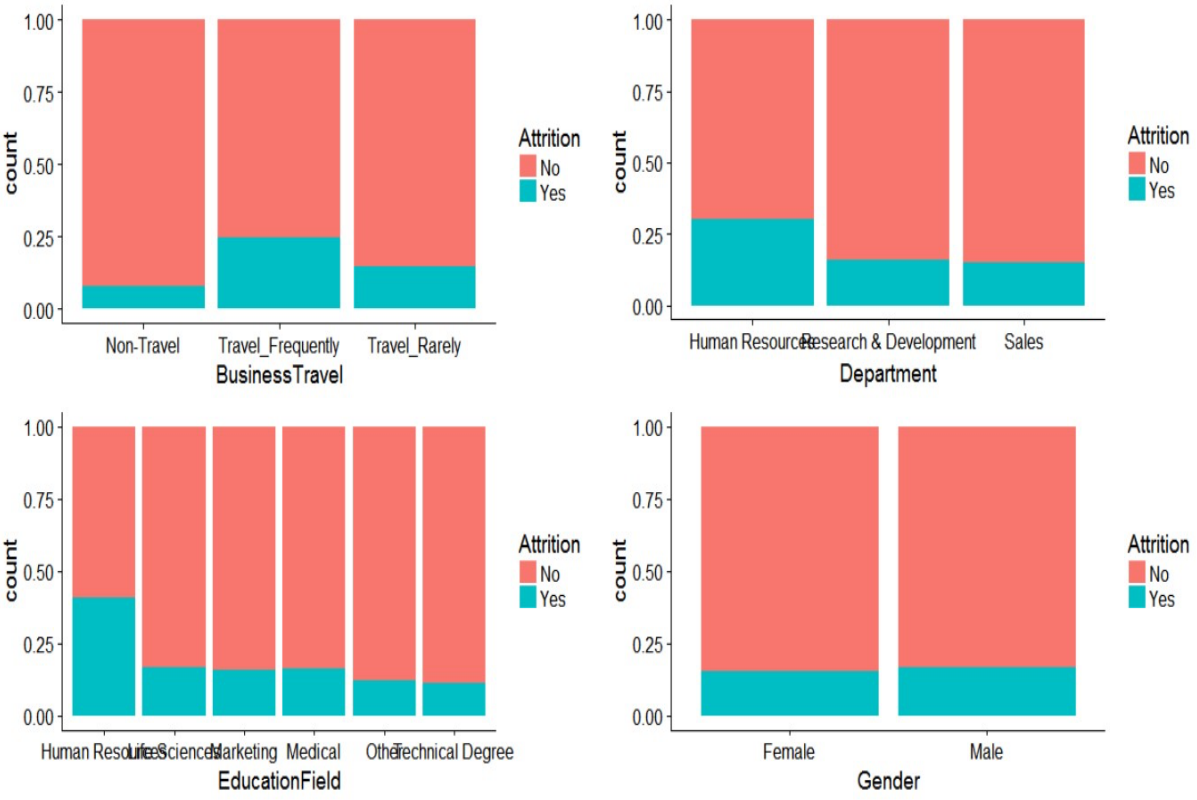
2.We chose the median imputation for NA's in continuous variables. As median imputation will work better because it is a number that is already present in the data set and is less susceptible to outlier errors as compared to mean imputation.

Snapshot of the summary:

```
> summary(employee1)
      Age        Attrition           BusinessTravel                Department
 Min.   :18.00   No :3699   Non-Travel       : 450   Human Resources       : 189
 1st Qu.:30.00   Yes: 711   Travel_Frequently: 831   Research & Development:2883
 Median :36.00              Travel_Rarely    :3129   Sales                 :1338
 Mean   :36.92
 3rd Qu.:43.00
 Max.   :60.00

 DistanceFromHome Education         EducationField    Gender     JobLevel
 Min.   : 1.000   1: 510   Human Resources :  81   Female:1764   1:1629
 1st Qu.: 2.000   2: 846   Life Sciences   :1818   Male  :2646   2:1602
 Median : 7.000   3:1716   Marketing       : 477                 3: 654
 Mean   : 9.193   4:1194   Medical         :1392                 4: 318
 3rd Qu.:14.000   5: 144   Other           : 246                 5: 207
 Max.   :29.000            Technical Degree: 396

                    JobRole     MaritalStatus  MonthlyIncome     NumCompaniesWorked
 Sales Executive          :978   Divorced: 981  Min.   : 10090   Min.   :0.000
 Research Scientist       :876   Married :2019  1st Qu.: 29110   1st Qu.:1.000
 Laboratory Technician    :777   Single  :1410  Median : 49190   Median :2.000
 Manufacturing Director   :435                 Mean   : 65029   Mean   :2.692
 Healthcare Representative:393                 3rd Qu.: 83800   3rd Qu.:4.000
 Manager                  :306                 Max.   :199990   Max.   :9.000
 (Other)                  :645

 PercentSalaryHike StockOptionLevel TotalWorkingYears TrainingTimesLastYear
 Min.   :11.00     0:1893           Min.   : 0.00     Min.   :0.000
 1st Qu.:12.00     1:1788           1st Qu.: 6.00     1st Qu.:2.000
 Median :14.00     2: 474           Median :10.00     Median :3.000
 Mean   :15.21     3: 255           Mean   :11.28     Mean   :2.799
 3rd Qu.:18.00                      3rd Qu.:15.00     3rd Qu.:3.000
 Max.   :25.00                      Max.   :40.00     Max.   :6.000

 YearsAtCompany   YearsSinceLastPromotion YearsWithCurrManager
 Min.   : 0.000   Min.   : 0.000          Min.   : 0.000
 1st Qu.: 3.000   1st Qu.: 0.000          1st Qu.: 2.000
 Median : 5.000   Median : 1.000          Median : 3.000
 Mean   : 7.008   Mean   : 2.188          Mean   : 4.123
 3rd Qu.: 9.000   3rd Qu.: 3.000          3rd Qu.: 7.000
 Max.   :40.000   Max.   :15.000          Max.   :17.000

 EnvironmentSatisfaction JobSatisfaction WorkLifeBalance JobInvolvement
 1: 845                  1: 860          1: 239          1: 249
 2: 856                  2: 840          2:1019          2:1125
 3:1375                  3:1323          3:2698          3:2604
 4:1334                  4:1387          4: 454          4: 432
```

Exploratory Data Analysis:

Fig 1a



Findings:

The plots reveal the proportion of attrition is high :

1.For those who travelled frequently

2.In HR department

3.Have educational background in HRM

Fig 1b

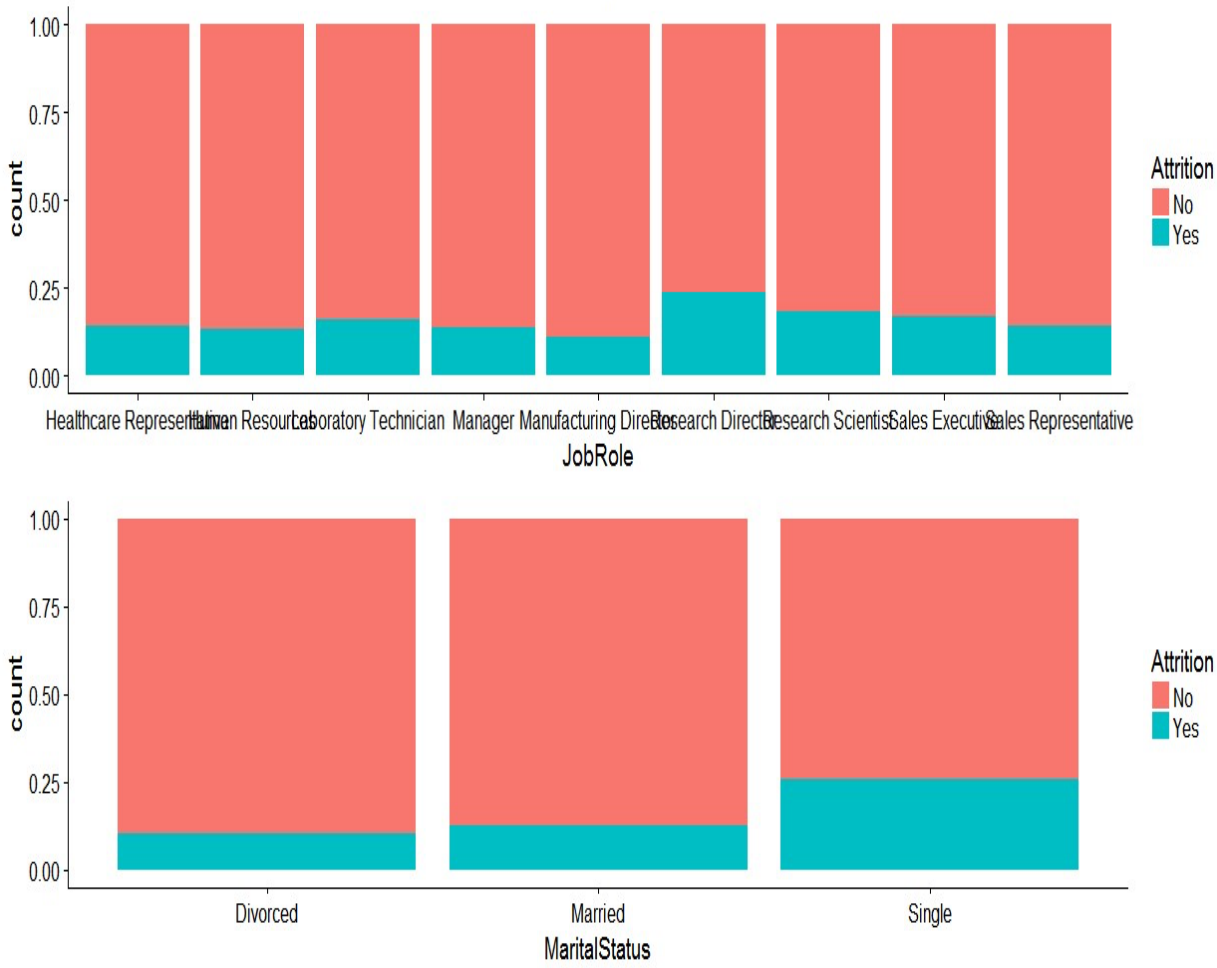
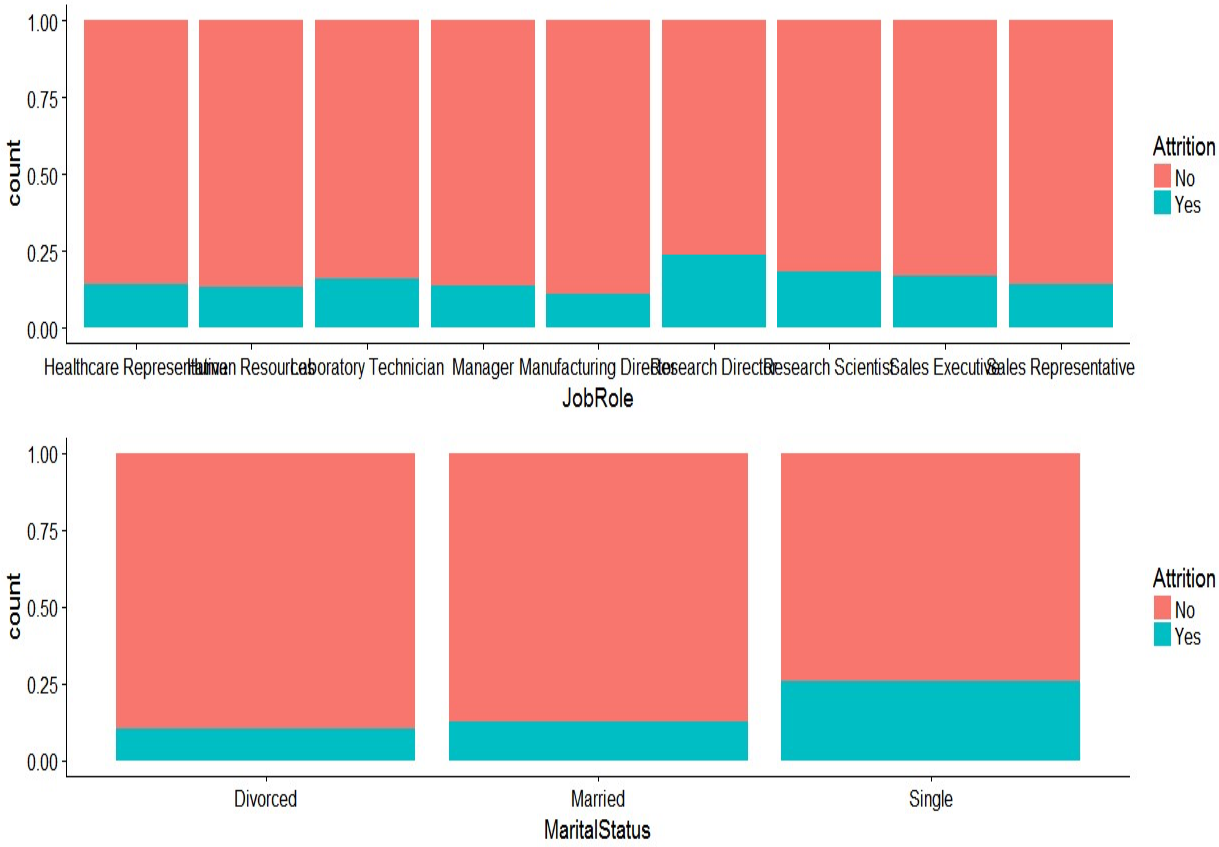


Fig 1b

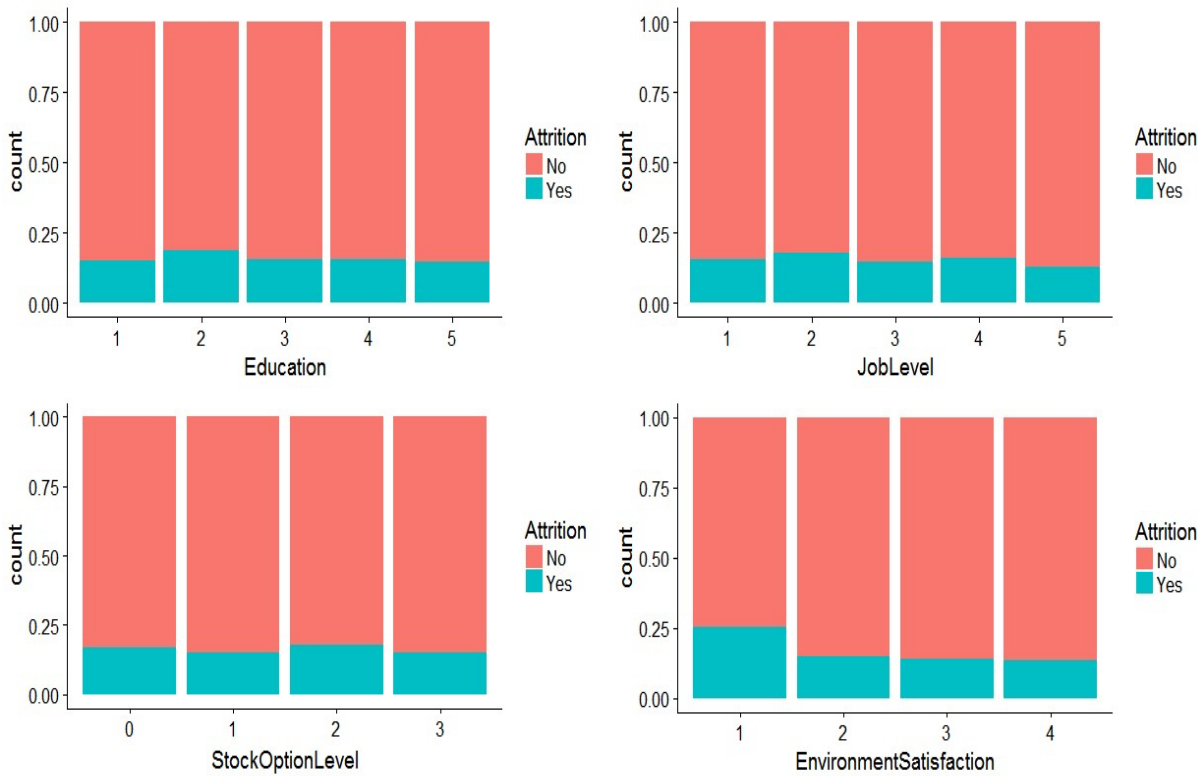


Findings:

The plots reveal the proportion of attrition is high :

1. For job role as research director
2. For those with Marital status as single

Fig 1c



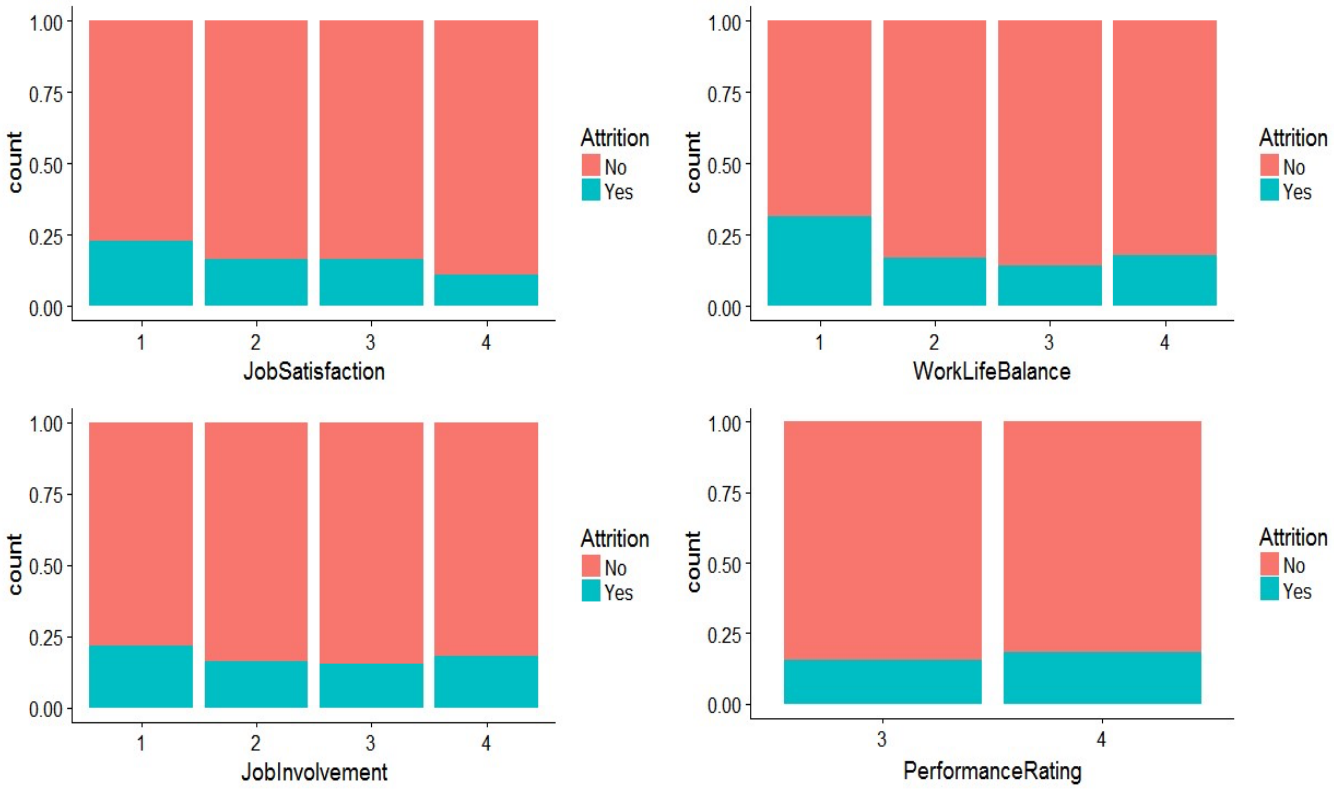
plots reveal the proportion of attrition is high :

1. For education level 2
2. For job level 2
3. Stock option level 2
4. Environment satisfaction level

Where

1=low,2=Med,3=High,4=Very The high

Fig 1d



Findings:

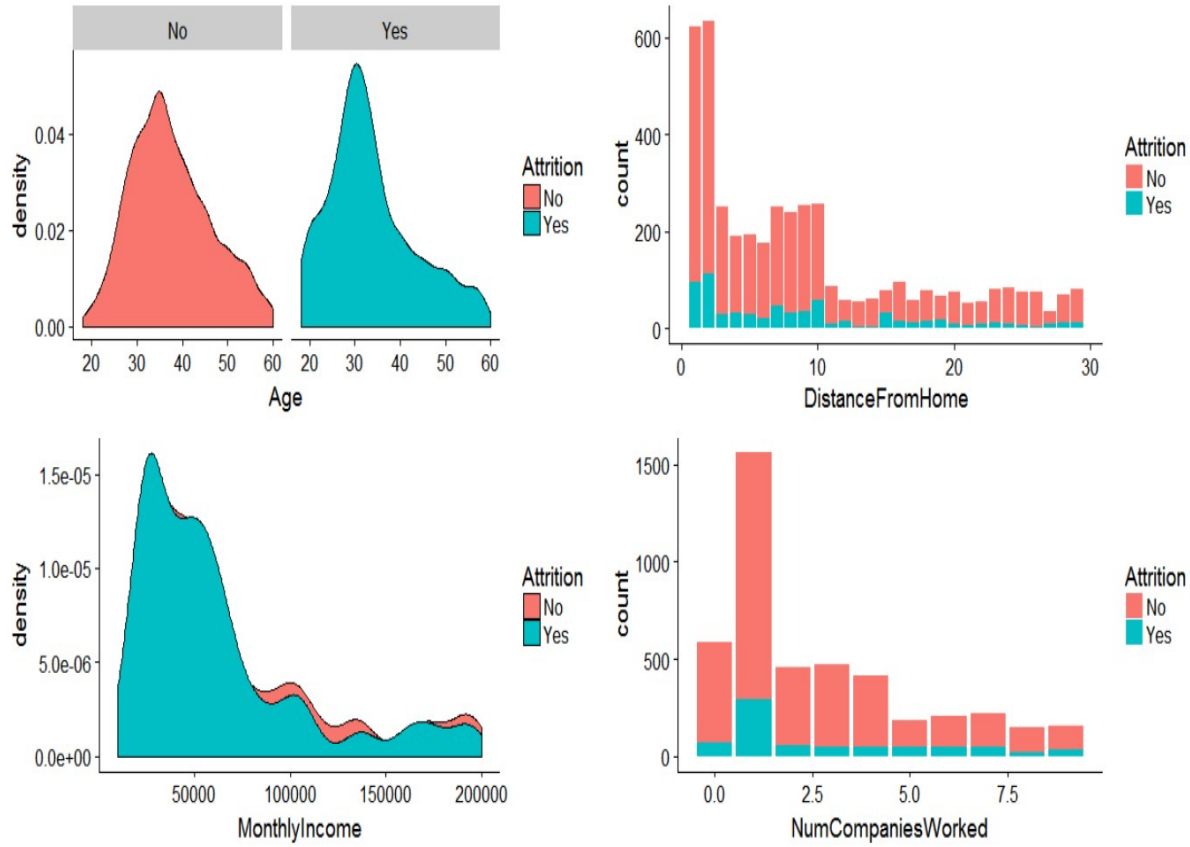
The plots reveal the proportion of attrition is high :

1. For Job satisfaction level 1
2. For Work life balance level 1
3. Job involvement level 1
4. Performance rating 4

Where

1=low,2=Med,3=High,4=Very high

Fig 2a



Findings:

The plots reveal:

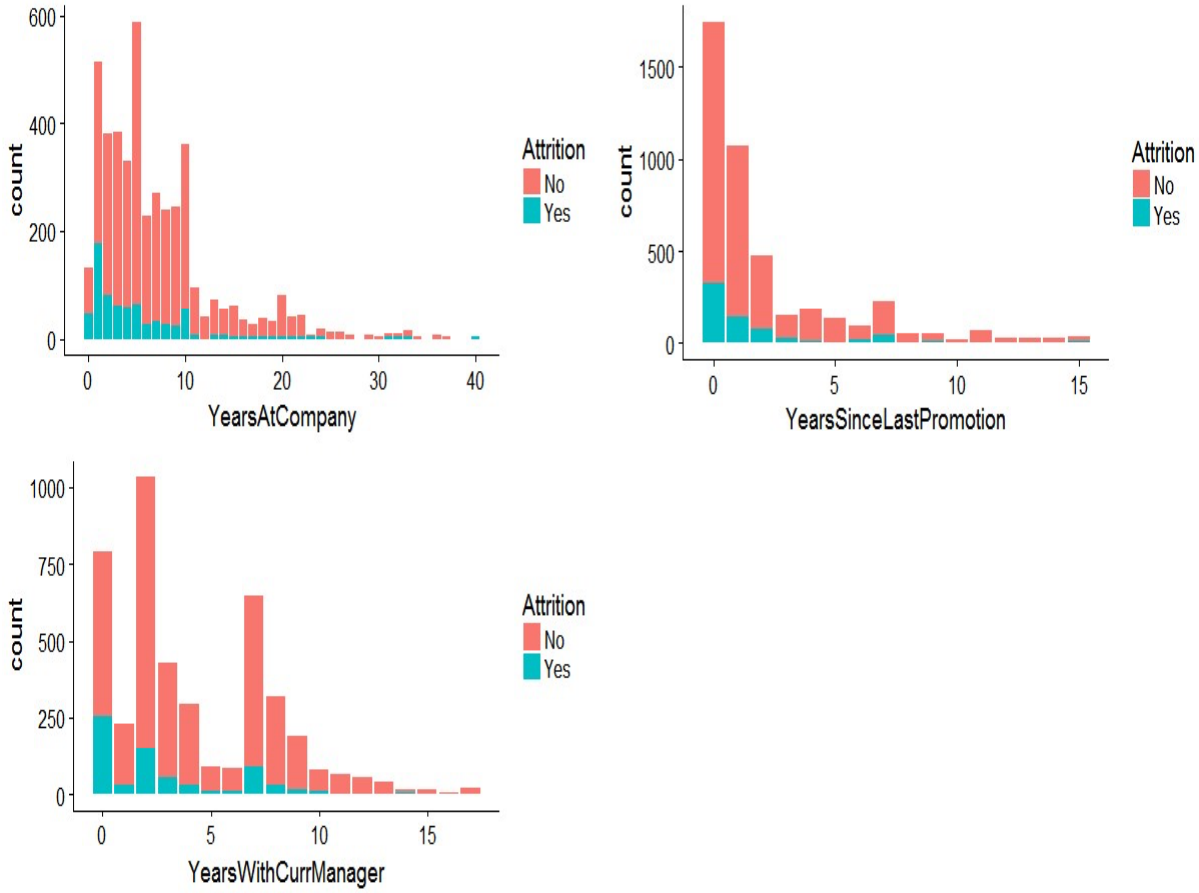
1. Age: Majority of employees leaving the organization are aged around 30 years
2. Distance from home:

Interestingly, employees who have left the company stay closer to office

3. Monthly income: Attrition is high in employees of lower income groups

Number of companies worked: Clearly people who have worked in one company before have higher attrition rate.

Fig 2c



Findings:

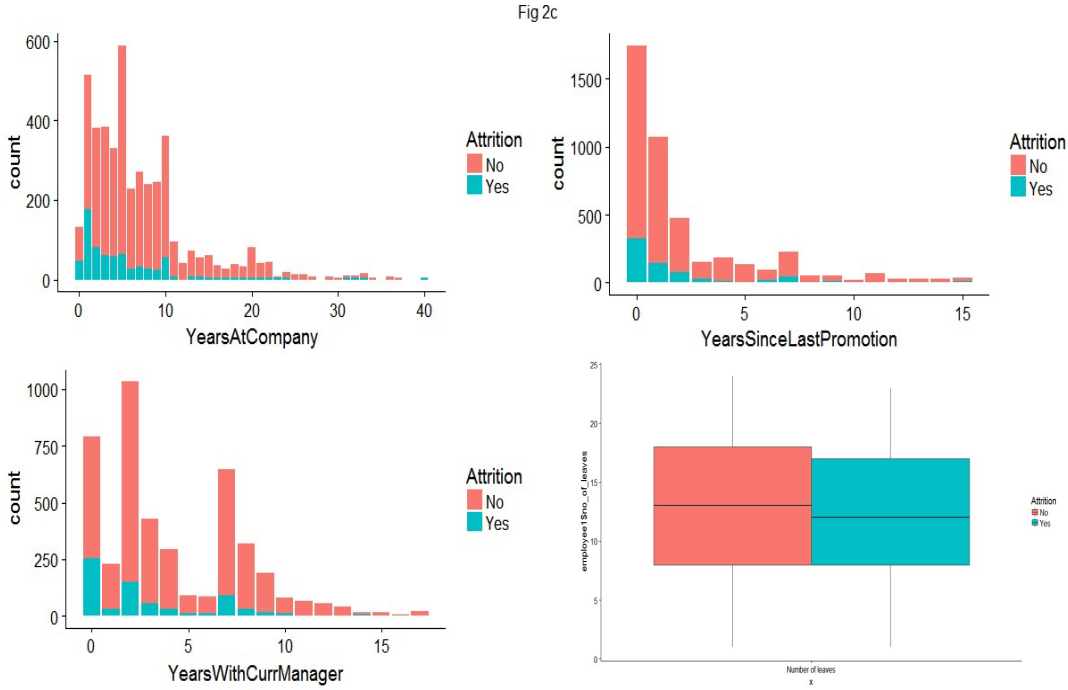
The plots reveal:

1.Years at Company: Larger proportion of new comers are quitting the organization. Which sidelines the recruitment efforts of the organization.

2.Years Since Last Promotion: Larger proportion of people who have been promoted recently have quit the company.

3.Years With Current Manager: As expected a new Manager is a strong cause for quitting.

Fig 2c



Findings:

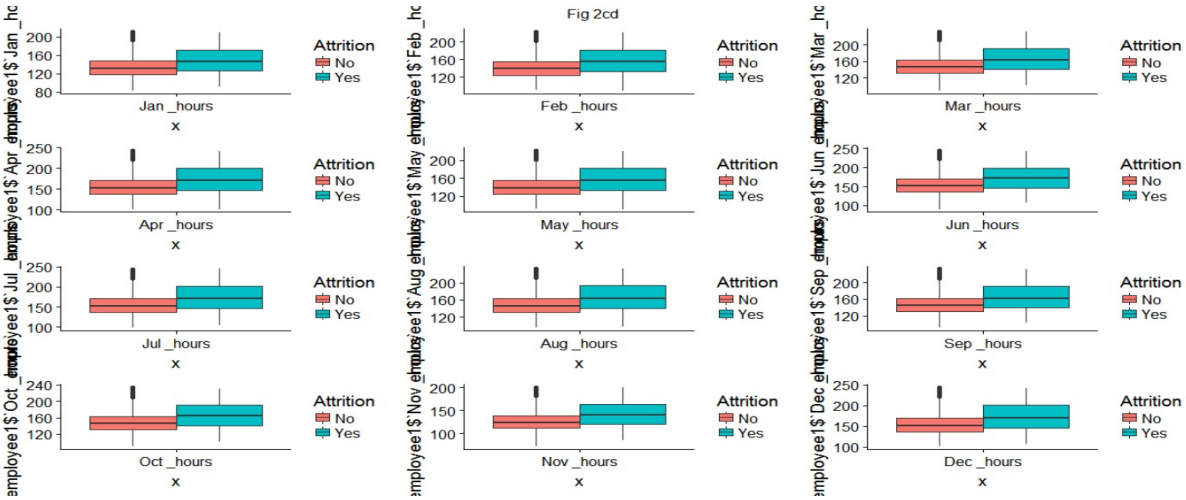
The plots reveal:

1.Years at Company: Larger proportion of new comers are quitting the organization. Which side-lines the recruitment efforts of the organization.

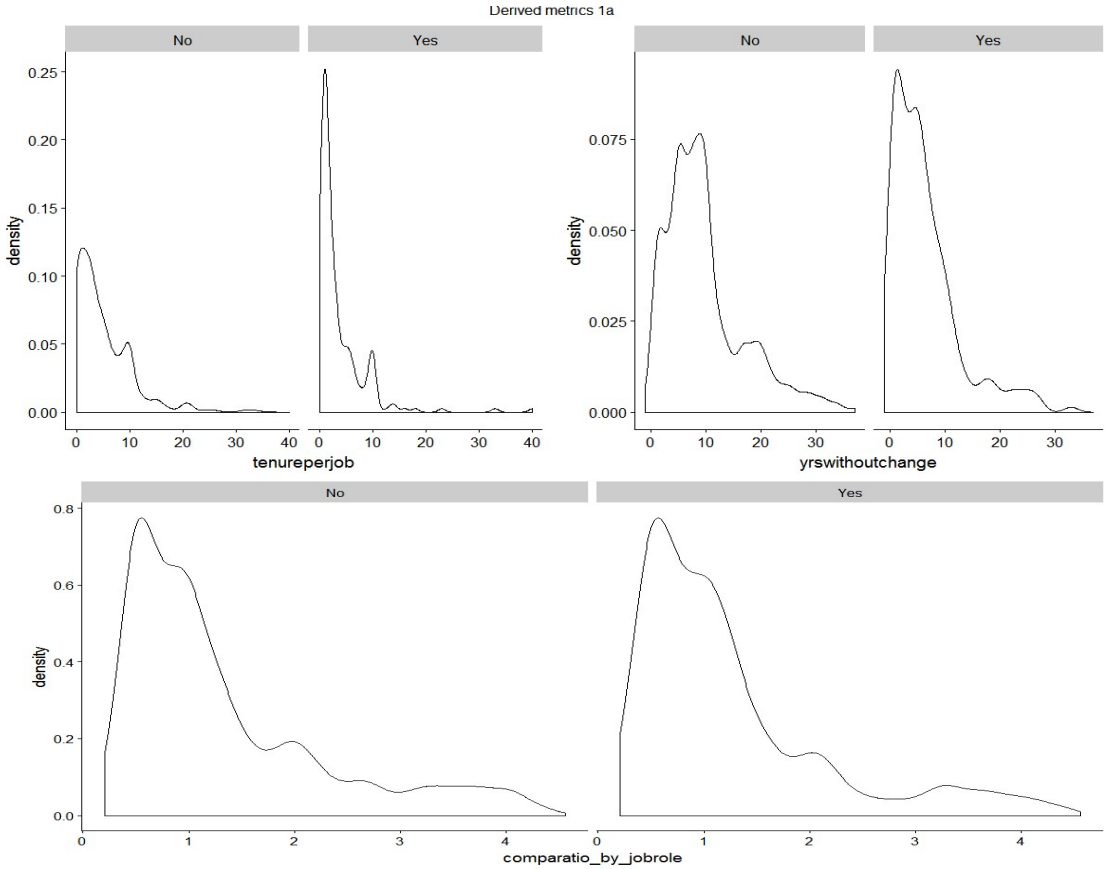
2.Years Since Last Promotion: Larger proportion of people who have been promoted recently have quit the company.

3.Years With Current Manager: As expected a new Manager is a strong cause for quitting.

4. Contrary to the belief, the median number of leaves by proportion of people who have left the company is lower than people who have not left.



Findings: While the individual months doesn't give any insight, one thing that's evident is people who #left the company have a higher median working hours per month than those who have.



Findings:

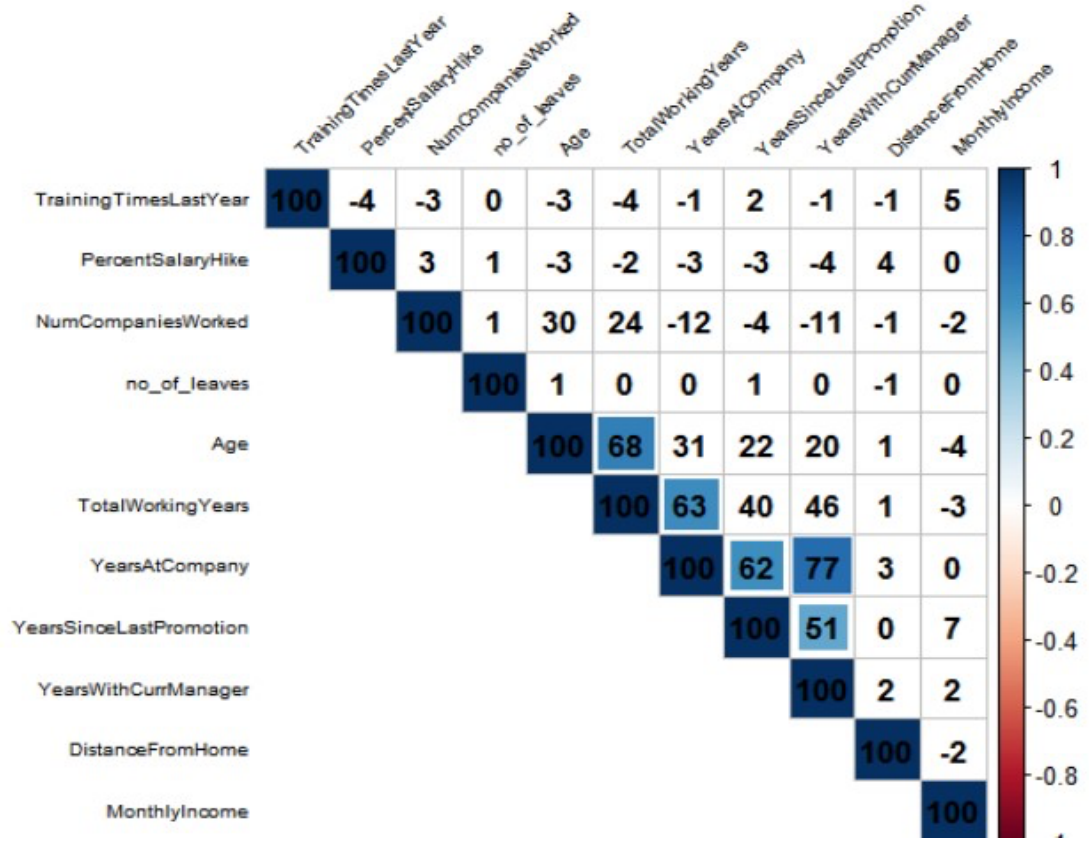
The plots reveal:

1.Attrition is higher for the group of people whose tenureperjob is in the bracket of 0 to 10

2.Attrition is higher for years without change in bracket of 0 to 10

3.People with lower compa ratio have a high attrition rate indicating dissatisfaction towards the salary being drawn

Correlation Analysis.



Findings:

The plots reveal:

1.Age is correlated to Total working years (68% correlation)

2. Total working years is correlated to years at company (61% correlation)
3. Yearsatcompany is correlated to yearssincelastpromotion (62% correlation)
4. Yearsatcompany is correlated to yearswithcurrentmanager (77% correlation)
5. Yearssincelastpromotion is correlated to yearswithcurrentmanager (51% correlation)

Final logistic regression model:Explanation

1.We had 15 iterations for the modelling, and the final model is the 15th model.

2.There are total of 11 independent variables in our model and the coefficients, standard error, Z value and P values are tabulated below.

3.Negative coefficients indicates a negative correlation with the dependent variable and vice versa.

	Estimate	Std	Error Z	Value	Pr(>|z|)
(Intercept)	-2.70424	0.09666	-27.977	< 2e-16	***
Age	-0.38839	0.08072	-4.812	1.50E-06	***
NumCompaniesWorked	0.33465	0.05856	5.715	1.10E-08	***
TotalWorkingYears	-0.42701	0.10843	-3.938	8.21E-05	***
YearsSinceLastPromotion	0.5492	0.07345	7.477	7.58E-14	***
YearsWithCurrManager	-0.53412	0.08426	-6.339	2.31E-10	***
EnvironmentSatisfaction	-0.376	0.05595	-6.721	1.81E-11	***
JobSatisfaction	-0.34356	0.05487	-6.261	3.82E-10	***
WorkLifeBalance	-0.22875	0.05498	-4.161	3.17E-05	***
`Oct _hours`	0.63607	0.05418	11.74	< 2e-16	***
BusinessTravel.xTravel_Frequently	0.87474	0.12758	6.856	7.07E-12	***
MaritalStatus.xSingle	1.05283	0.11366	9.263	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The deviance residuals ,Null deviance and residual deviance and AIC are as follows:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9923  -0.5692  -0.3663  -0.1857   3.7588

    Null deviance: 2678.3  on 3048  degrees of freedom
Residual deviance: 2135.2  on 3037  degrees of freedom
AIC: 2159.2

Number of Fisher Scoring iterations: 6
```

Computing the statistical significance of the Null and Residual deviance using the chi square statistic:

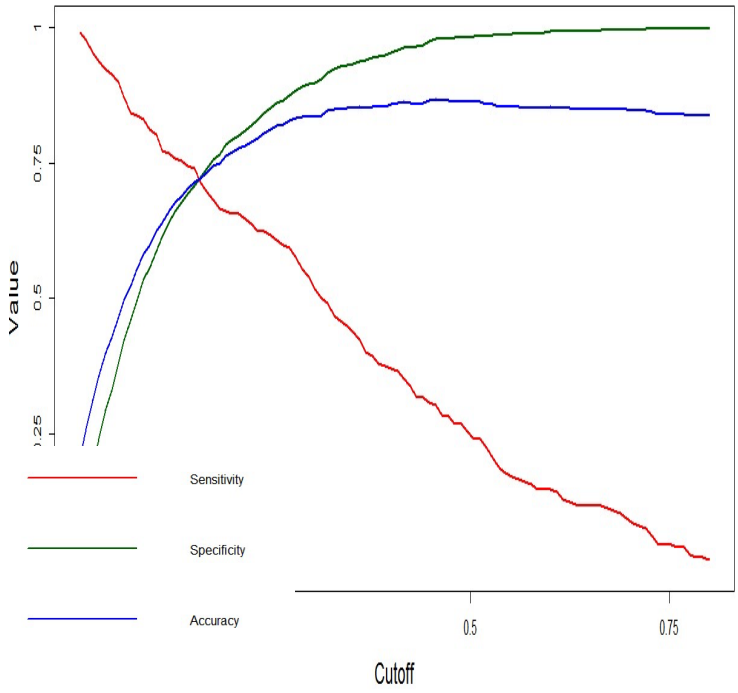
```
> 1-pchisq(2678.3 ,3048) #Null Deviance
[1] 0.9999996
> 1-pchisq(2135.2 ,3037) #Residual Deviance
[1] 1
> 1-pchisq(2678.3-2135.2 ,3038-3037) #The difference between Null and Residual deviance
[1] 0
```

This implies the deviance of the model only with constant term and deviance of the model with 11 independent variables added to it are the same, and we can reject the null hypothesis.

Model Evaluation: Determining the optimal cutoff(on test data)

Below is the performance measures(Accuracy, sensitivity and specificity for 3 cut off levels. Clearly , as the cut off decreases the sensitivity improves. The computed optimal cut off for our final model as calculated on the test data is **0.16.**

Cutoff	Accuracy	Sensitivity	Specificity
0.5	0.86	0.25	0.98
0.4	0.85	0.37	0.95
0.16	0.72	0.71	0.72



The plot on the right displays the accuracy, sensitivity and specificity based on the optimal cut-off value of 0.16. The colours of the different curves are denoted by the legends on the plot.

Model Evaluation: Few more performance measures

1. **KS Statistics**: KS Test measures to check whether model is able to separate events and non-events. *KS*

= Maximum difference between Cumulative % Event and Cumulative % Non-Event. Ideally, it should be in first three deciles and score lies between 40 and 70.For our model , 0.43 is the Max k statistic for test data

2. **Lift and Gain Table on test data:**

a*)We can identify 75% of employees who are going to leave the company by contacting 40% of total*

employees.

b) *The Cum Lift of 3.71 for top decile, means that when selecting 10% of the records based on the*

model, one can expect 3.71 times the total number of targets (events) found by randomly selecting 10%of-file without a model.

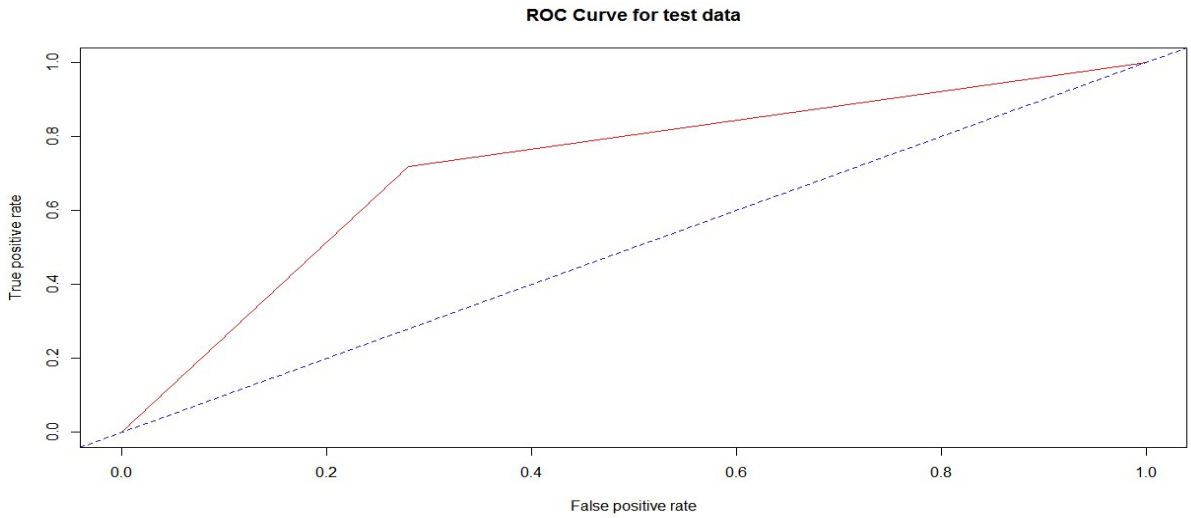
c) *The maximum KS statistic is 46.53 and is in the second decile which is a very good indicator for the robustness of the model.*

bucket	total	totalresp	Cumresp	Gain (% of cum_respondants)	Cumlift	total_nonresp	Cum_nonresp	percent_cum_nonresp	KS_statistic
1	137	83	83	37.05	3.71	54	54	4.749340369	32.30
2	136	49	132	58.93	2.95	87	141	12.40105541	46.53
3	136	17	149	66.52	2.22	119	260	22.86719437	43.65
4	136	21	170	75.89	1.90	115	375	32.98153034	42.91
5	136	12	182	81.25	1.63	124	499	43.88742304	37.36
6	136	11	193	86.16	1.44	125	624	54.88126649	31.28
7	136	11	204	91.07	1.30	125	749	65.87510994	25.20
8	136	6	210	93.75	1.17	130	879	77.30870712	16.44
9	136	9	219	97.77	1.09	127	1006	88.47845207	9.29
10	136	5	224	100.00	1.00	131	1137	100	0

Model Evaluation: Few more performance measures cont.

1. **AUC (Area Under the curve) :** *Our model has **an AUC of 0.72**. AUC value of >0.70 is a an indication that the model is good.*

2. **ROC Curve:**



Conclusions and Recommendations

Based on the variables in the final model, we have the following interpretations and recommendations to the management:

NumCompaniesWorked *:* Attrition is higher for a person who has worked in multiple companies. This could be an input to the hiring team to be wary of the people who have changed companies frequently. At least, they should interview that person in greater detail to find out if they had genuine reasons for changing.

TotalWorkingYears - Lesser experienced people have a higher attrition rate possibly due to them not seeing growth opportunities. It's recommended to focus HR and management's energy on these individuals and show them a growth path in the organization. Giving access to senior staff who are willing to volunteer their time and energy to mentor the new entrants might help with this.

YearsSinceLastPromotion - Clearly, the longer an employee has gone without a promotion, higher are the odds of them not staying in the company. For such employees, maybe a job rotation where their skills are better suited and chances of them growing are higher could be something HR can look into.

Oct _hours*:* It could be that, employees are forced to put in long hours maybe even picking up slack for others especially during the last quarter. This leads to frustration followed by quitting. One feedback to HR and management is to introduce intermediate progress checking throughout the year rather than at the end.

Conclusions and Recommendations cont.

YearsWithCurrManager - When the manager changes, comfort level between employee and immediate supervisor is likely to get reduced. In many companies, it's very common for new managers to have a HR facilitated employee feedback session in the first six months of taking up the role. Such an exercise will help the employees point out any blind spots the manager can be having from the team's perspective and help improve the relationship with the manager and their team.

EnvironmentSatisfaction/WorklifeBalance/Jobsatisfaction - If actionable items can be derived from the surveys, they should be published and management should be held accountable to take those actions and improve the satisfaction or employee perception. If employees don't see any changes inspite of giving feedback especially constructive feedback, they will move out.

BusinessTravel.xTravel_Frequently - Put a cap on the travel or have mandatory breaks between travel, make provisions for virtual meetings, offer to rotate employee roles so they get a break from traveling-

-

MaritalStatus.xSingle - This could be because such employees are not tied down to the locations, maybe have a little more time on their hands. Provide opportunities that involve business challenges, travel, contributing back to society, personal growth etc for such employees to retain them.