HR ANALYTICS CASE STUDY

SUBMITTED BY:

SHIVAM KUMAR

SURABHI SOOD

VIKASH THAKUR

VARGHESE B MATHEW

Problem Statement

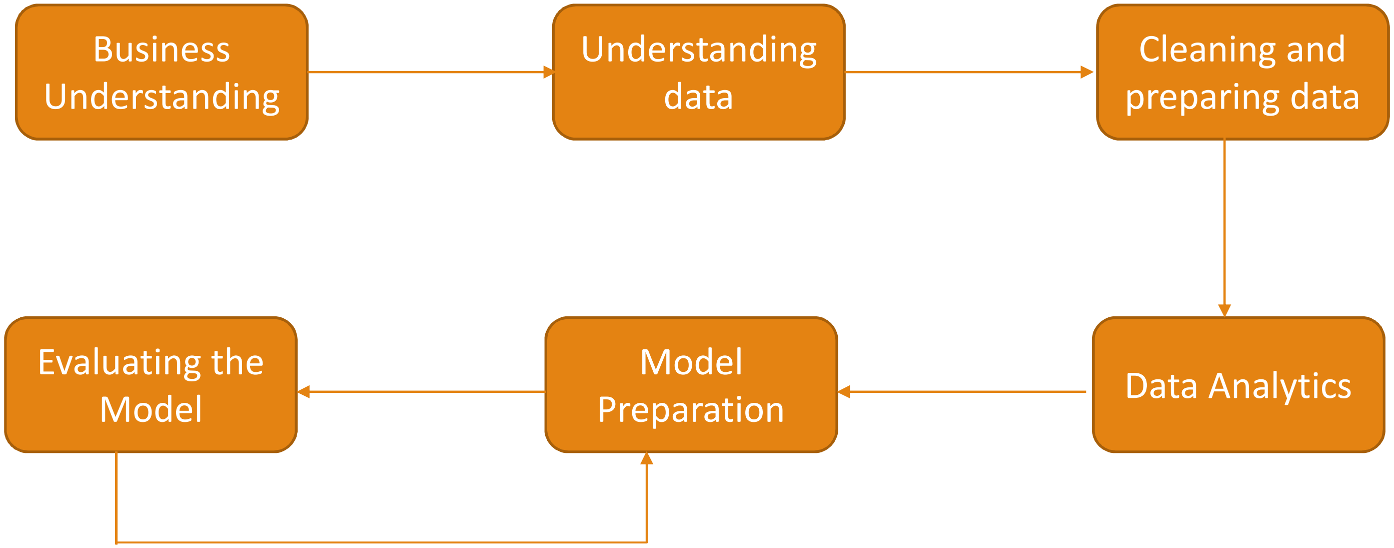
Every year 15%(approx.) of employees of XYZ leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad and negative for the company, because 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 acclimatise themselves to the company

Objective

The objective is to help the management to determine to be focused to curb attrition. In other words what changes are required in the organisation. Also to find the most crucial factors which are to be addressed on priority in order to make their employees stay.

Problem Solving Approach



About the Data

AGE GROUP:
18 to 60

TOTAL EMPLOYEES:
4410

ATTRITION:
711 or 17.29%

FEMALE:
1764

MALE:
2646

ANALYSIS : DATA UNDERSTANDING

Datasets provided for analysis

1. The Manager Survey Data – Collected from a company wide survey.

   The Employee Survey Data – Collected from a company wide survey.

   In Time Data – Collected from company’s attendance Log sheet/ Machine. Out Time Data – Collected from company’s attendance Log sheet.

   General Data – General data includes employees personal data along with education.

2. Attrition from general dataset is the target variable.

3. EnvironmentSatisfaction, JobInvolvement, JobSatisfaction are defined like :1 means 'Low', 2 means 'Medium', 3 means 'High‘ and 4 means 'Very High‘

4. WorkLifeBalance is defined like below 1 means 'Bad', 2 means 'Good', 3 means 'Better’ and 4 means 'Best'

5. Education is defined like : 1 means 'Below College’, 2 means 'College', 3 means 'Bachelor‘ , 4 means 'Master' and 5 means ‘Doctor’

ANALYSIS : DATA CLEANING AND PREPARATION

1. NA’s value treated :

Dealing with NAs in NumCompanies

If the difference between total working years and years at company is equal to zero, then NumCompanies worked =0,

if not zero then NumCompanies worked =1,as no other info is provided about other companies the employee worked.

Dealing with NAs in total working years, by replace it with years at Company

NA’s EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance is treated by calculating Mode.

2. Single Value columns removed such as EmployeeCount, StandardHours and Over18 for having no significance.

For Attrition as this has 2 levels, being realigned as numerical Yes == 1 and No == 0.

3. Average working hours are calculated based on in time and out time datasets for each employee as part of derived metrics.

4. Dummy variables for following categorical predictors having more than 2 levels are created

5. All continuous variables are scaled.

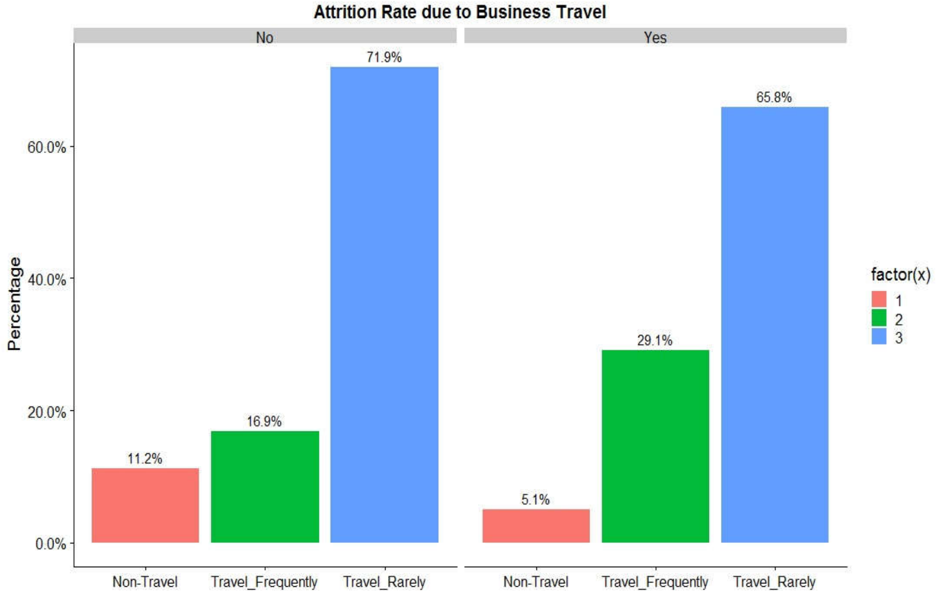
EDA : EXPLORING CATEGORICAL VARIABLES

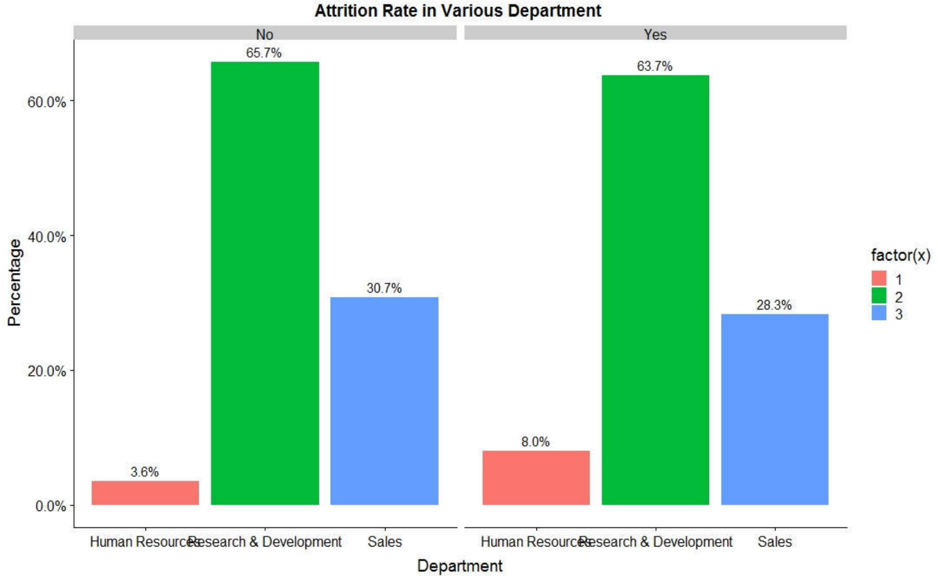
1. Attrition rate due to business travel:

Travel Rarely attrition rate is highest

2.Attrition rate in various department :

Research and development , attrition rate is highest





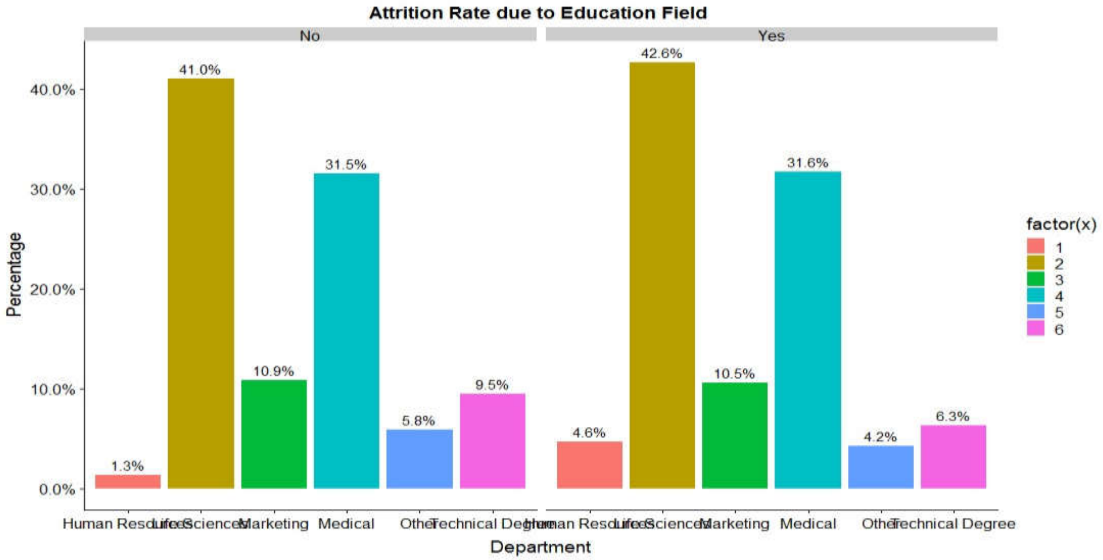
EDA : EXPLORING CATEGORICAL VARIABLES

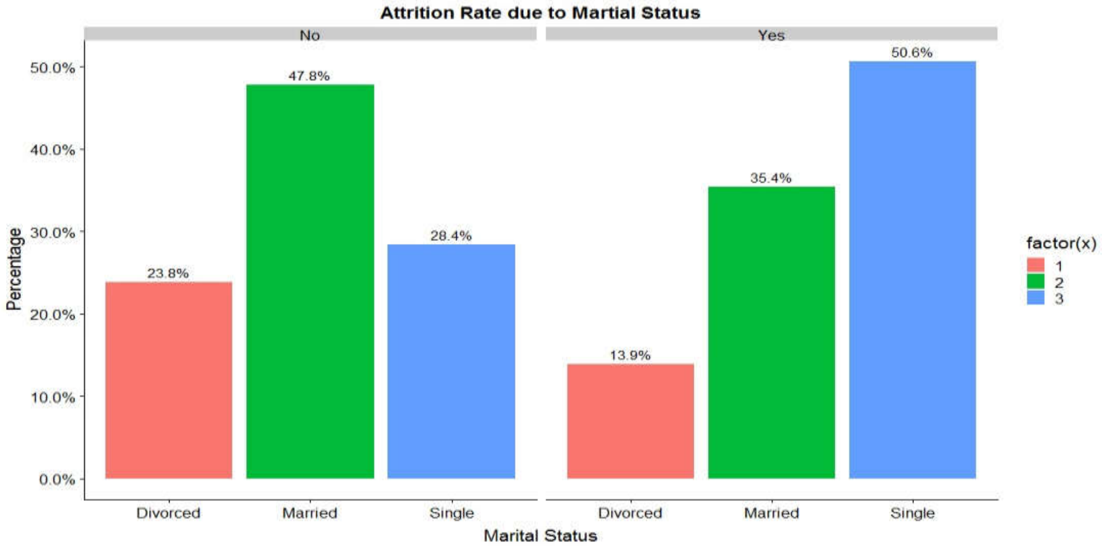
1.Attrition rate due to education field

Life Science field attrition is highest

2. 1.Attrition rate due to marital status

Single ones attrition is highest





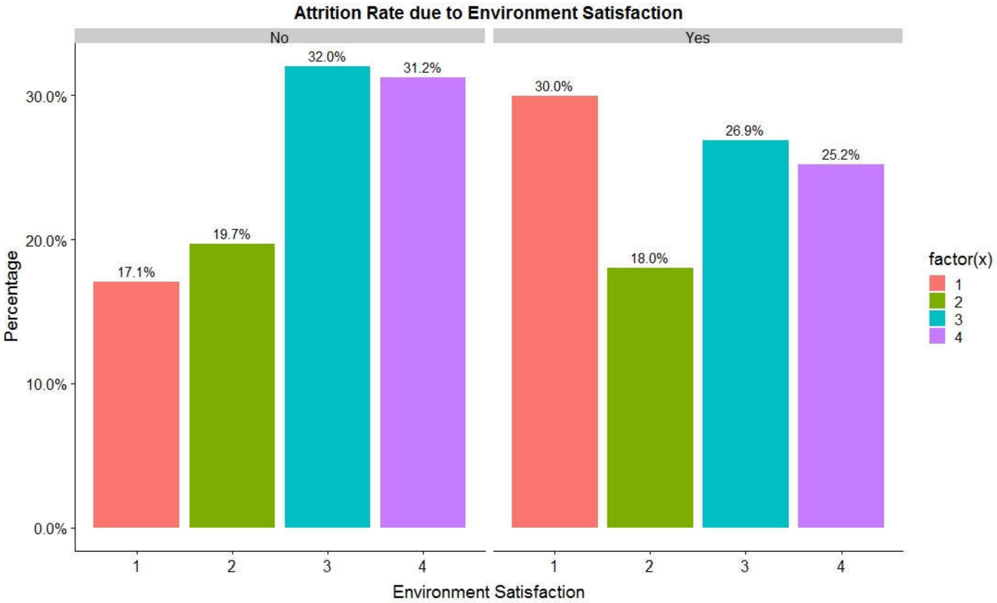
EDA : EXPLORING CATEGORICAL VARIABLES

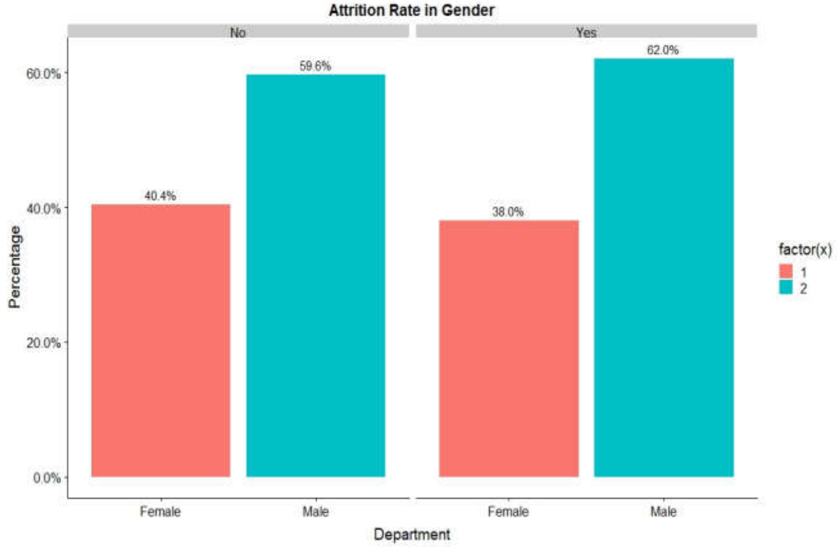
1.Attrition due to Environmental satisfaction

Grade 1 more likely to leave

2. Attrition in gender

Males have highest attrition rate





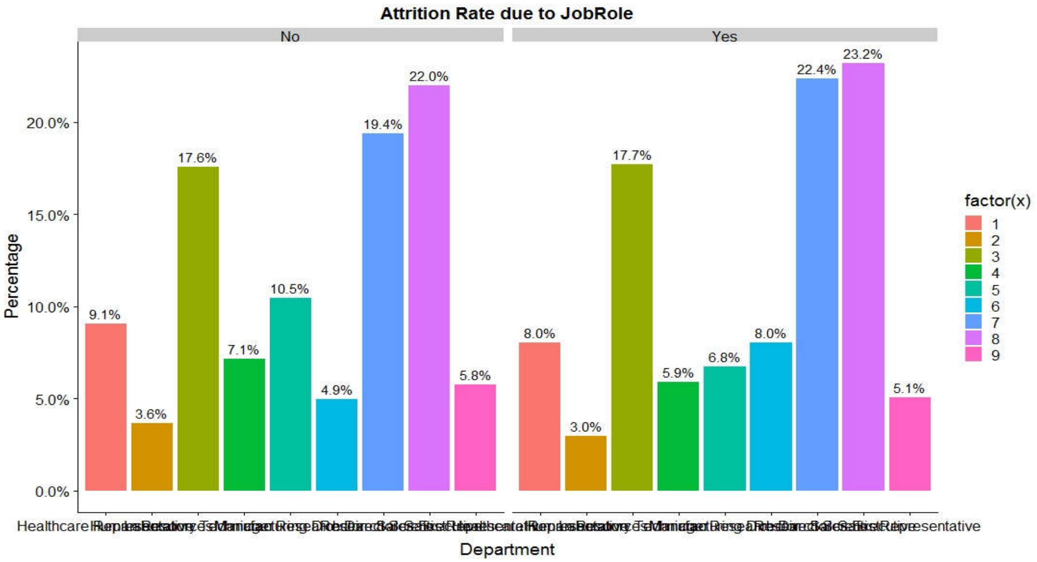
EDA : EXPLORING CATEGORICAL VARIABLES

1. Attrition due to job role:

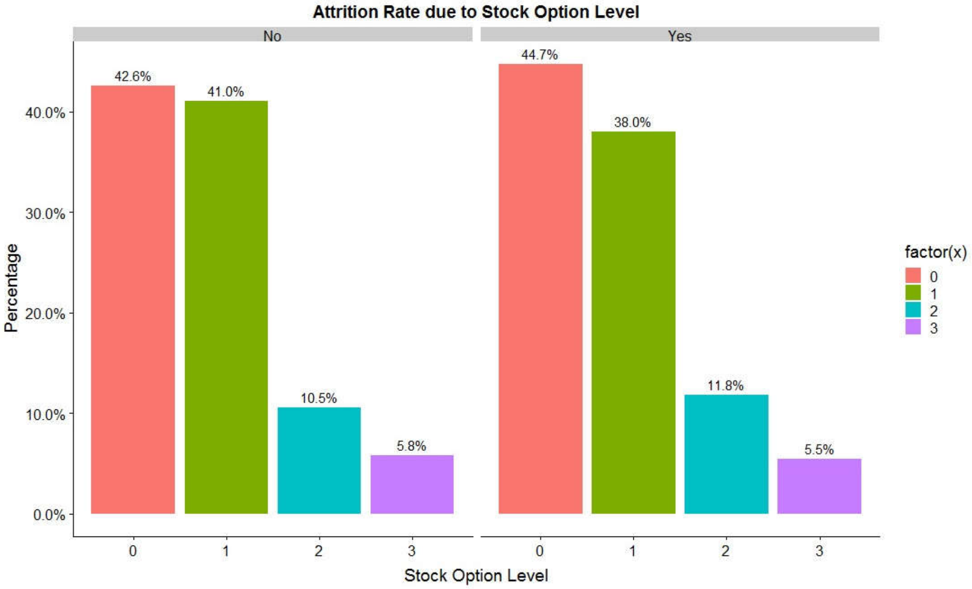
Sales Executive with 23.2% more likely to leave

2. Attrition due to stock option

Stock option 0 has highest attrition .

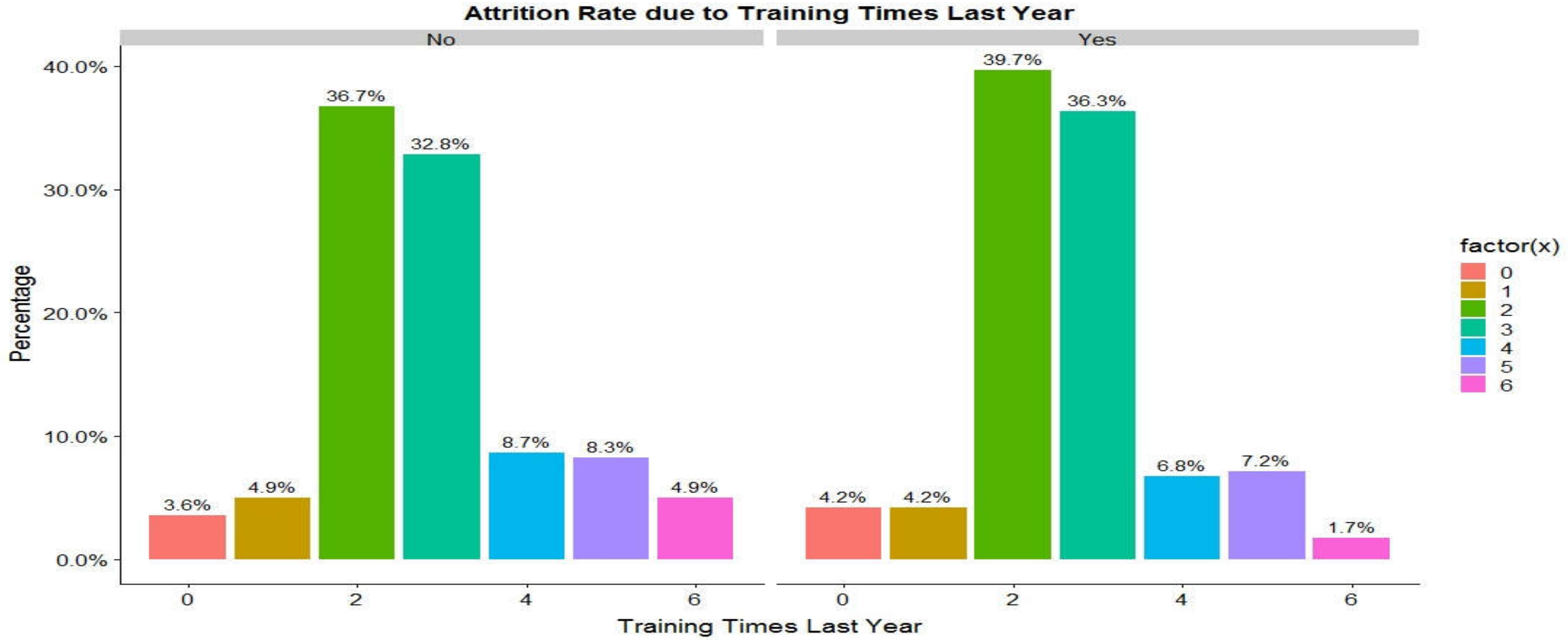


Attrition Rate due to JobRole



Attrition Rate due to Stock Option Level

EDA : EXPLORING CATEGORICAL VARIABLES



Attrition rate is highest for Training time last year 2 followed by 3

EDA : EXPLORING NUMERIC VARIABLES

In the graphs:

Blue zone represents Attrition rate,

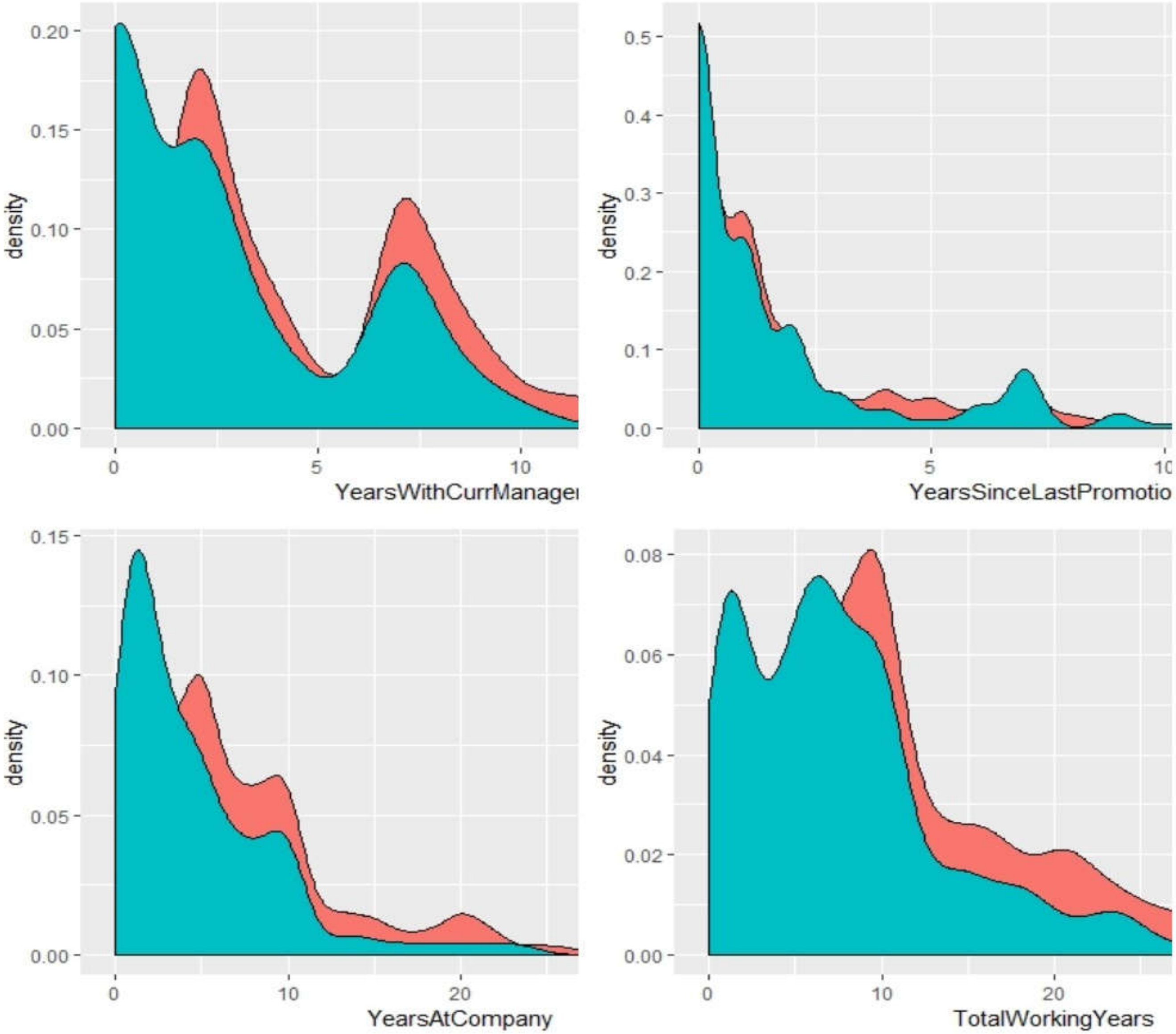
Red zone marks for no attrition .

YearsAtCompany: More attrition among those with low no of years at Company

YearsSinceLastPromotion: More attrition among those who promoted recently

TotalWorkingYears: More Attrition Among those with low no. of working years

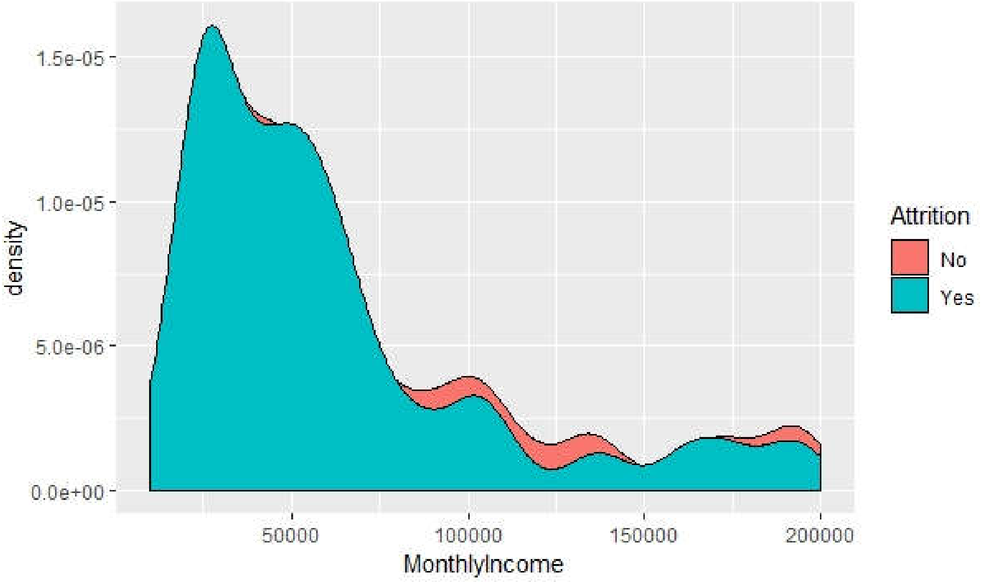
Yearswithcurrent manager: More attrition among those with less no. of years under current manager

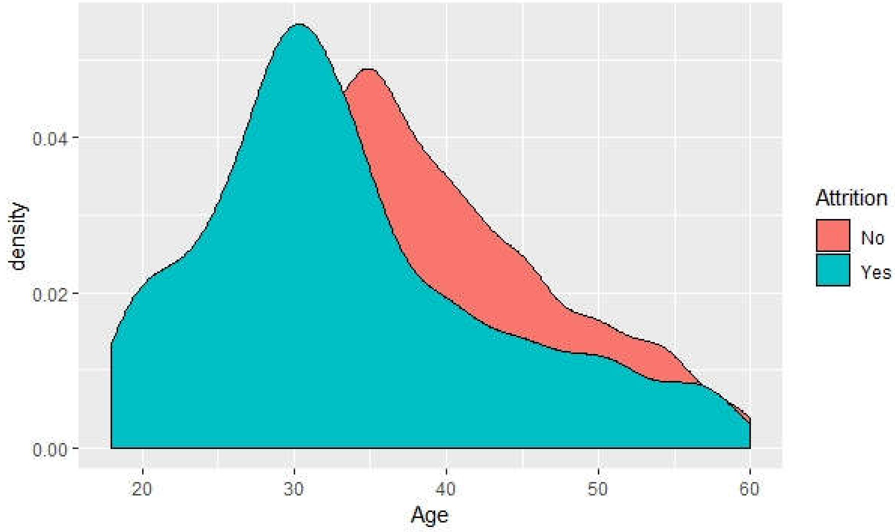


EDA : EXPLORING NUMERIC VARIABLES

1.Monthly income: Attrition higher among low monthly income

2.Age: Attrition higher among low age





MODEL BUILDING USING LOGISTIC REGRESSION

1.Our Response variable is “Attrition” (1 == Yes, & 0 == No)

2.Rest all non constant numeric variables are scaled to aid in regression modelling.

3.Splitting data into training and test data set.

4.For creating Train and test datasets from final data set:

5.We fixed seed to 100 and Used split ratio of 0.7 for training dataset and remaining data has been assigned to test dataset

6.Initial model has been conceived with glm function, then StepAIC has been applied to arrive at standard model which yielded on iterative predictor selection without major reduction in AIC Score.

7. Removed the variables having high VIF value and low significance i.e. if p-value > 0.05

8.Checked the correlation among variables appropriately and removed from the model accordingly.

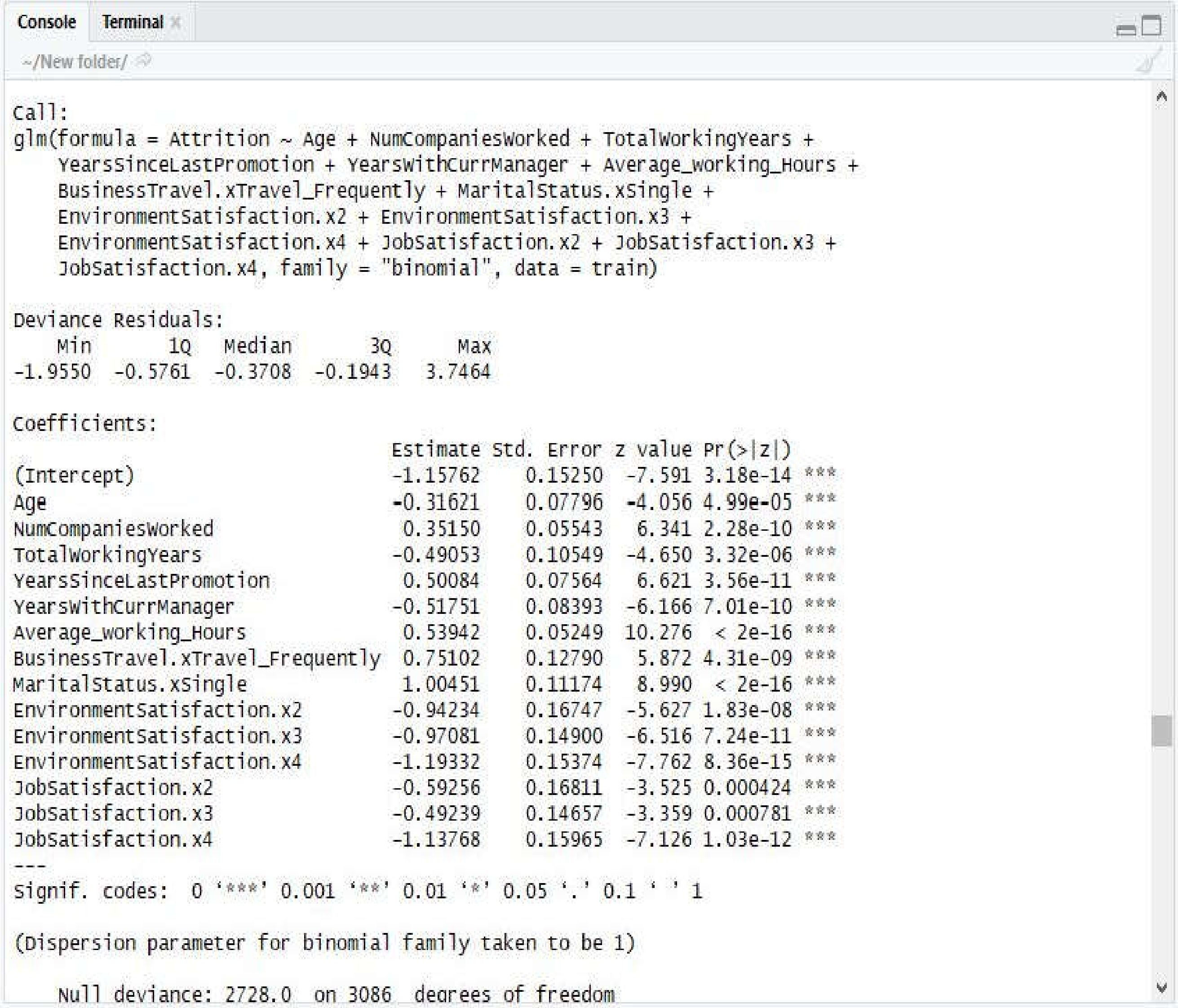
9.Then based on VIF (variance inflation factor) and P - value (with significance) predictors have been filtered and after another 28 iterations we could achieve our final model with almost all predictors being significant with lowest VIF are present.

FINAL LOGISTIC MODEL - SIGNIFICANT VARIABLES

In the final model we have 14 significant variables with positive and negative coefficients.

Positive coefficient means that if positively more the value of the variable more will be the chance of attrition and

Negative coefficient means they will be affecting attrition rate negatively



```
Call:
glm(formula = Attrition ~ Age + NumCompaniesWorked + TotalWorkingYears +
    YearsSinceLastPromotion + YearsWithCurrManager + Average_working_Hours +
    BusinessTravel.xTravel_Frequently + MaritalStatus.xSingle +
    EnvironmentSatisfaction.x2 + EnvironmentSatisfaction.x3 +
    EnvironmentSatisfaction.x4 + JobSatisfaction.x2 + JobSatisfaction.x3 +
    JobSatisfaction.x4, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9550  -0.5761  -0.3708  -0.1943   3.7464

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -1.15762    0.15250  -7.591 3.18e-14 ***
Age                               -0.31621    0.07796  -4.056 4.99e-05 ***
NumCompaniesWorked                 0.35150    0.05543   6.341 2.28e-10 ***
TotalWorkingYears                 -0.49053    0.10549  -4.650 3.32e-06 ***
YearsSinceLastPromotion            0.50084    0.07564   6.621 3.56e-11 ***
YearsWithCurrManager              -0.51751    0.08393  -6.166 7.01e-10 ***
Average_working_Hours              0.53942    0.05249  10.276  < 2e-16 ***
BusinessTravel.xTravel_Frequently  0.75102    0.12790   5.872 4.31e-09 ***
MaritalStatus.xSingle              1.00451    0.11174   8.990  < 2e-16 ***
EnvironmentSatisfaction.x2        -0.94234    0.16747  -5.627 1.83e-08 ***
EnvironmentSatisfaction.x3        -0.97081    0.14900  -6.516 7.24e-11 ***
EnvironmentSatisfaction.x4        -1.19332    0.15374  -7.762 8.36e-15 ***
JobSatisfaction.x2                -0.59256    0.16811  -3.525 0.000424 ***
JobSatisfaction.x3                -0.49239    0.14657  -3.359 0.000781 ***
JobSatisfaction.x4                -1.13768    0.15965  -7.126 1.03e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2728.0  on 3086  degrees of freedom
```

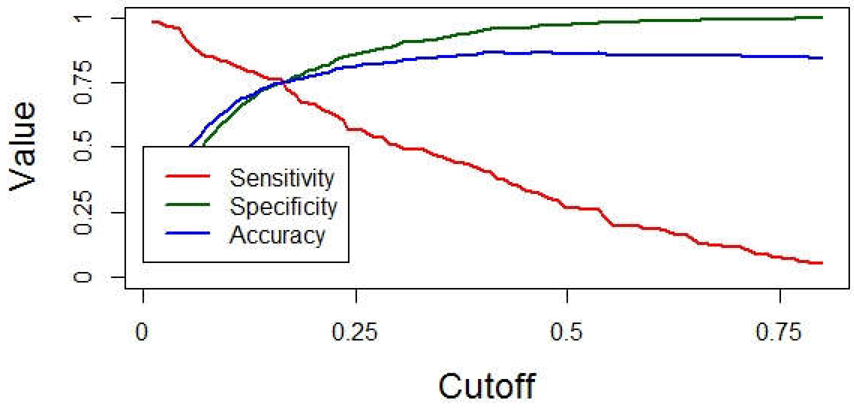
To evaluate the final model, we execute the model on the test dataset and performed the following model validations:

Finding Accuracy, Specificity and Sensitivity through Confusion Matrix

In order to find a suitable probability cut-off, we checked the Accuracy, Sensitivity and Specificity from 1% to 80% probability values

The optimum cut-off probability is the one where the value of specificity and sensitivity are close to each other. Here we have taken a safe range of 0.01. Cut-off probability was found to be ~0.1616162

MODEL EVALUATION



Cut off probablity	value	percentage
Accuracy	0.7241119	72%
Sensitivity	0.7323944	73%
Specificity	0.7225225	72%

MODEL EVALUATION CONTD.

Calculating KS Statistics for test data

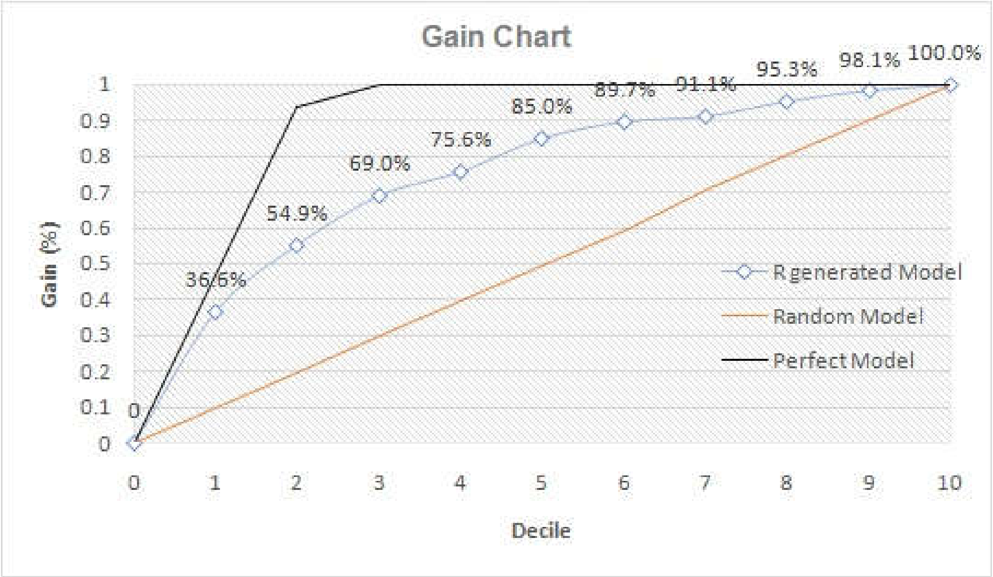
KS statistics measures the degree of separation between positive and negative distribution.

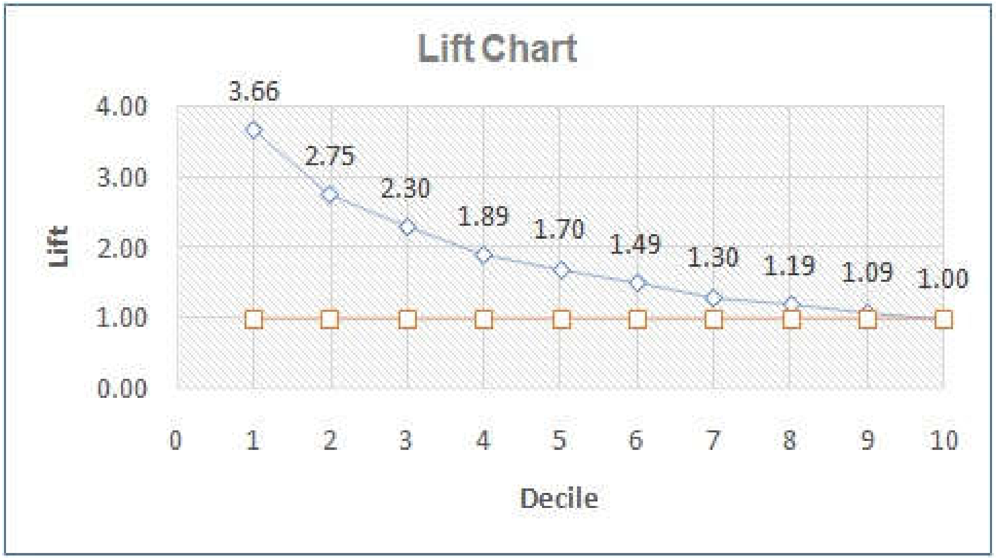
The optimal value of KS statistics for a good model should lie between 40-60 and should be within first 3 Deciles.
KS-Statistics for our final model is 0.46

Lift and Gain Chart

It helps to measure the effectiveness of the model by calculating the percentage of events captured in each decile.

Here, we can see that even by targeting top 30% of the employees, we can predict attrition approx 70% of the time





CONCLUSION

FACTORS	RESULTS/SUGGESTIONS
Age	Employees with lower AGE group are more prone to quit the company
AverageWorkinghours	Attrition Rate is higher for those working more than 8 hours
YearsSinceLastPromotion	If Employees are getting frequent promotions then lesser chance of quitting the organization when compared to employees whose promotions are delayed
YearsWithCurrManager	Employee who works with same manager for longer time has less chances of quitting the organization
NumCompanies worked	No of companies worked is more, Attrition rate increases
EnvironmentSatisfaction, JobSatisfaction	The better these are for employees the lesser Employee will quit the organization
MaritalStatus	Those who are single tend to leave the company more
BusinessTravel	Frequent Business Travel is a reason for attrition
TotalWorkingYears	As the number of years increases, attrition tend to be lower

THANK YOU