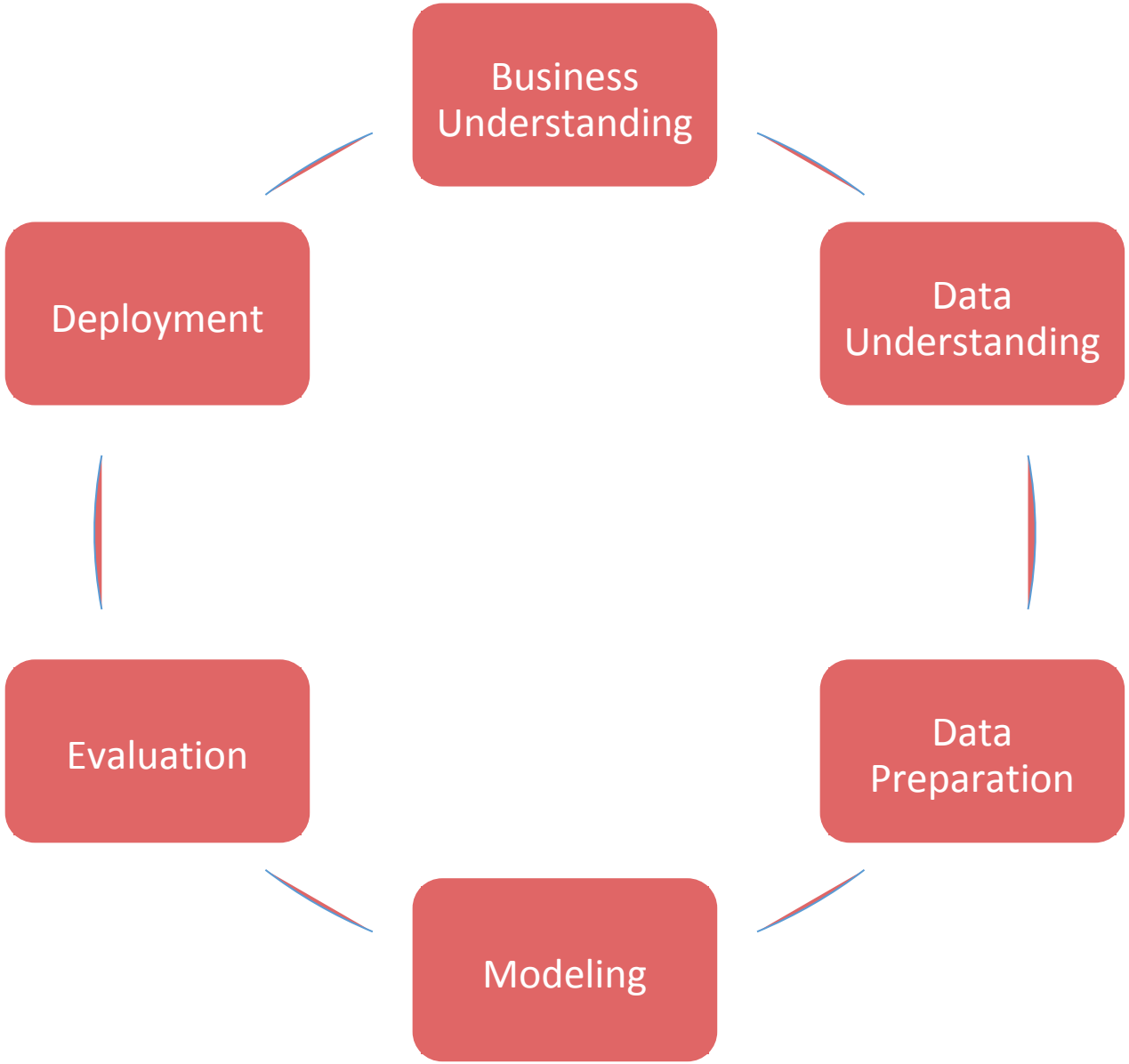
HR Analytics Case Study

Group Name: DataCatalysts

1. Sudarshan Bhattacharjee
2. Sagnik Banerjee
3. Huzaifa Hareeri
4. Yashjit Gangopadhyay

CRISP Data-mining process flow



Business Understanding

The very first step of understanding a business problem is analysing the problem statement & figure out the deliverables of the analysis.

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, because of the following reasons-

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.

Business Understanding

GOALS OF ANALYSIS -

1) To create a model to find the Probability of attrition using logistic regression.

2) The model will be able to effectively factor out strong influencers that finally leads to an employee resigning

3) Suggest corrective measures to the management of XYZ based on the most important factors to reduce high attrition levels

Business Understanding

BUSINESS DOMAIN KNOWLEDGE -

According to Dr. John Sullivan, a Head Recruitment Consultant at Victor Noble Associates in his blogpost -

https://goo.gl/o3y4Ve; he says that just attrition rate is not a very healthy metric to gauge how an organization is doing anymore. If the attrition of a firm is say 10%, but these people leaving the firm are top performers, it's bad for the firm. However, for the same firm, say the attrition rate is 20% but the people leaving the organization are non-performers, that is a really good situation for the firm. For the average performers, to determine the ideal minimum turnover rate, these steps should be considered:

1) Look at what percentage of past turnover among this population is related to a life event (a family move for example) that made continued employment by your firm difficult.

Business Understanding

2) Subtract from that percentage the number of turnovers that could have been avoided by making simple accommodations (a flexible schedule for example).

3) The result should be the ideal turnover rate among average performers. Also, a healthy attrition rate for each organization is relative it has been said.

However, in this article https://goo.gl/iTR2o1 Meredith Mejia puts a number to the question - 10%. Considering this, there is definitely a reason for concern for XYZ as the current attrition rate is well above 10%.

To put into perspective, the analysis & model can suggest probabilistically what are the most influential factors are for an employee to leave the firm.

Data Understanding

Input files used:

1. generic_info.csv - general information which includes demographic informations

info on Age, working experience educational background, department he/she working in, monthly income,salary hike received etc.

2. employee_survey.csv - information on survey done for employee regarding job satisfaction, environment satisfaction etc.

3. manager_survey.csv - information on survey given by manager regarding performance of the employee job involvement level

4. in_time.csv - information on in time of employee, helps for calculation of working hours.

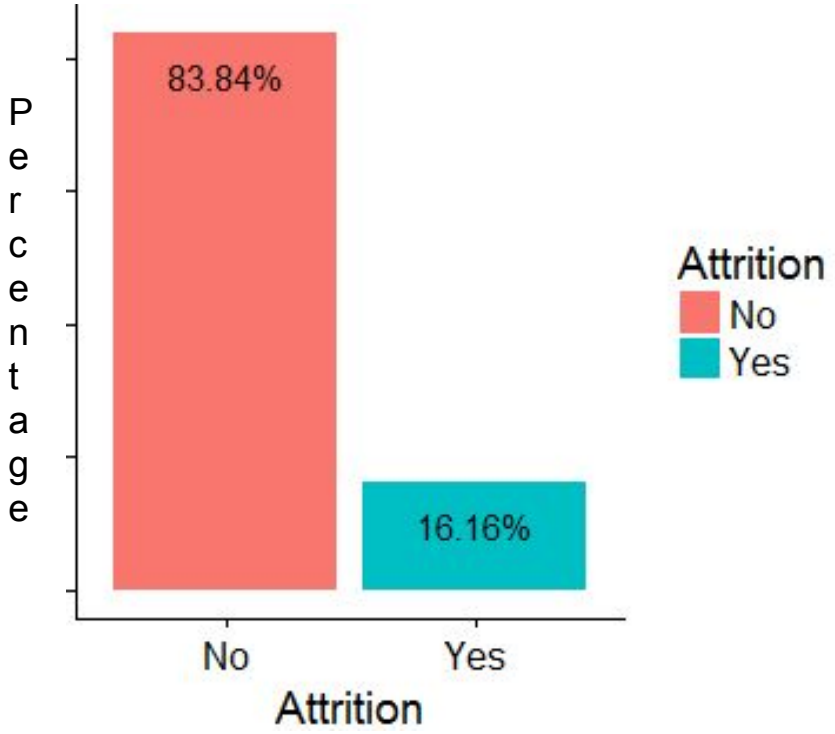
5. out_time.csv - information on out time of employee, helps for calculation of working hours of employee

6. Data dictionary - column wise information of all the files

Data Understanding

Of the 4410 employees present in the firm, last year 16.1% of the employees left their job. This translates to about 710 employees. It has been established that at any point in time, the firm must maintain a workforce of 4000 employees.

According to the figures, last year there was a deficit of 300 employees in the company.



From business perspective, this transpires to a recurring expenditure of paying hiring managers, agencies; loss in terms of time and experienced resources leaving the house.
The deliverables might also get affected due to this unhealthy trend.

Data Cleaning & Preparation

- No concrete information was provided on which is the primary key field in between all the different files & also the files in_time & out_time had a missing primary key column heading.
- From the data dictionary & the name of the column, we assumed Employee ID / Employee number is the primary key column. Then some checks were performed to find if actually the assumption is true or not.

```
if(length(unique(employee_survey$EmployeeID)) = = nrow(employee_survey)){paste("EmployeeID is primary key for the file employee_survey")}

sum(duplicated(employee_survey$EmployeeID))
```

  It is known that a primary key field is unique & cannot be repetitive. The above check performed on each of the files confirmed that indeed EmployeeID is the primary key field.

- Next, NA check is done. 19 NA's in the field NumCompaniesWorked and 9 in the field TotalWorkingYears are found. For NumCompaniesWorked:
- NA - can be replaced with 2 if TotalWorkingYears and YearsAtCompany is not same and difference between them is just one years or else can be replaced as 1 where TotalWorkingYears = YearsAtCompany

Data Cleaning & Preparation

- 2. zero's are replaced with 2 if difference between TotalWorkingYears and TotalWorkingYears is 1 , assuming that it was an issue of wrong data input & also an employee wouldn't be possibly changing the company more than once in an year as most of the companies have probationary period of 1 year in the contract agreement.
- For TotalWorkingYears :

  1. NA - replace it with YearsAtCompany if NumCompaniesWorked is 0 or 1.

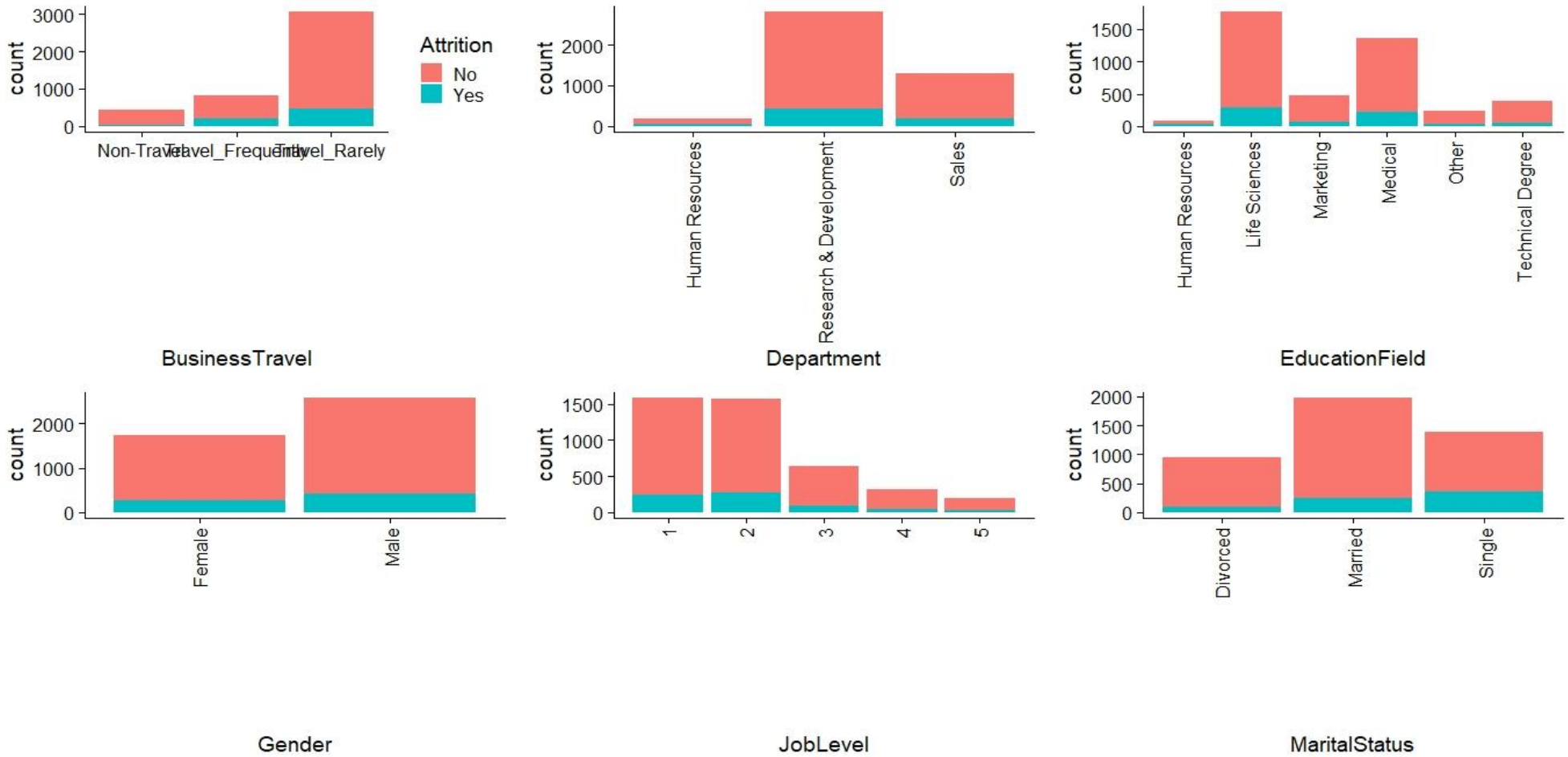
- After all the manipulation was done, marginal number of NAs - 5,9, 16 were found which were ignored.
- Average working hours of each of the employees for the previous year from the in_time & out_time files. This was one of the derived metrices.
- Lastly, all the data from different files was merged into a single data frame called base_data. The data was grouped by the primary key Employee ID
- NA check on the final data frame shows only 99 NA values are present in a data frame of 4410 observations amounting to only 2.24% of the entire data points. Hence, these blank values were omitted.

Deriving Metrics

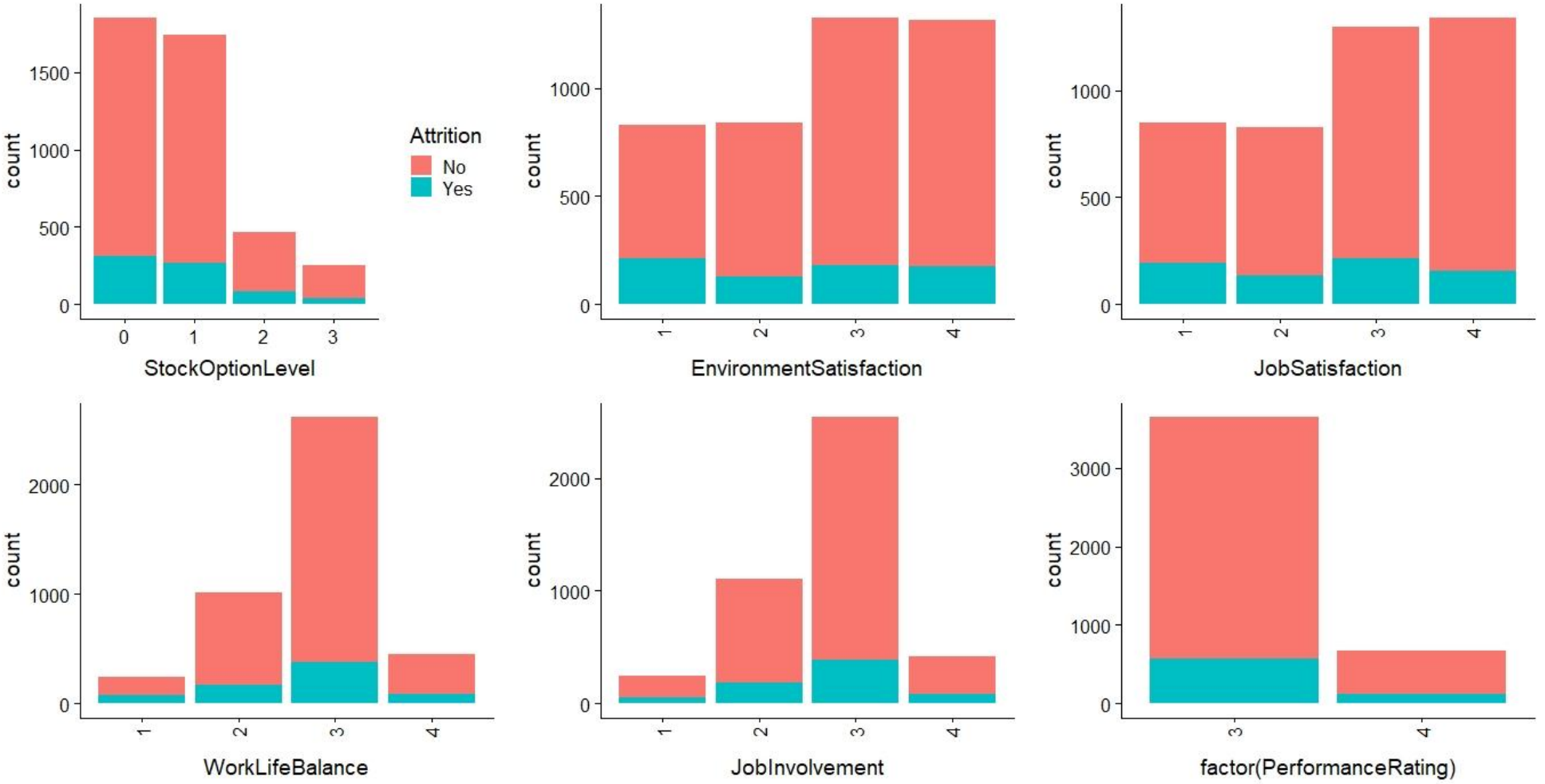
In light of the data provided to us already and gathering some insights about business use cases, we figured metrics like average working hours & overtime done by the employees are also very important factors & should be counted in. This is why we derived these metrics out of the in_time & out_time files provided to us.

Univariate Analysis

We found the proportion of employees who have left the company against those who are still in the company for each category. Here is the visualization of all categorical variables against attrition.



Univariate Analysis



Univariate Analysis - Conclusions

1) Based on travel, it can be said the section of people who get to travel a lot, are the ones who leave the company least frequently.

2) People from the HR department have the lowest rate of attrition and the R&D department is the one where people are leaving their jobs the most.

3) People with specialisation in Life sciences are most probable of leaving the organisation.

4) Male employees are more susceptible to leaving females seem to be more loyal to the organisation

5) Among men, those who are single are more probable of leaving the organisation. Understanbly as they are not settled.

6) StockOptionLevel - Employess with lower levels of Stockoption tend to leave the organisation more

7) EnvironmentSatisfaction - employees who have voted their work environment satisfaction to be of a low category, are the ones who leave the organisation the most. This is quite understandable.

Univariate Analysis - Conclusions

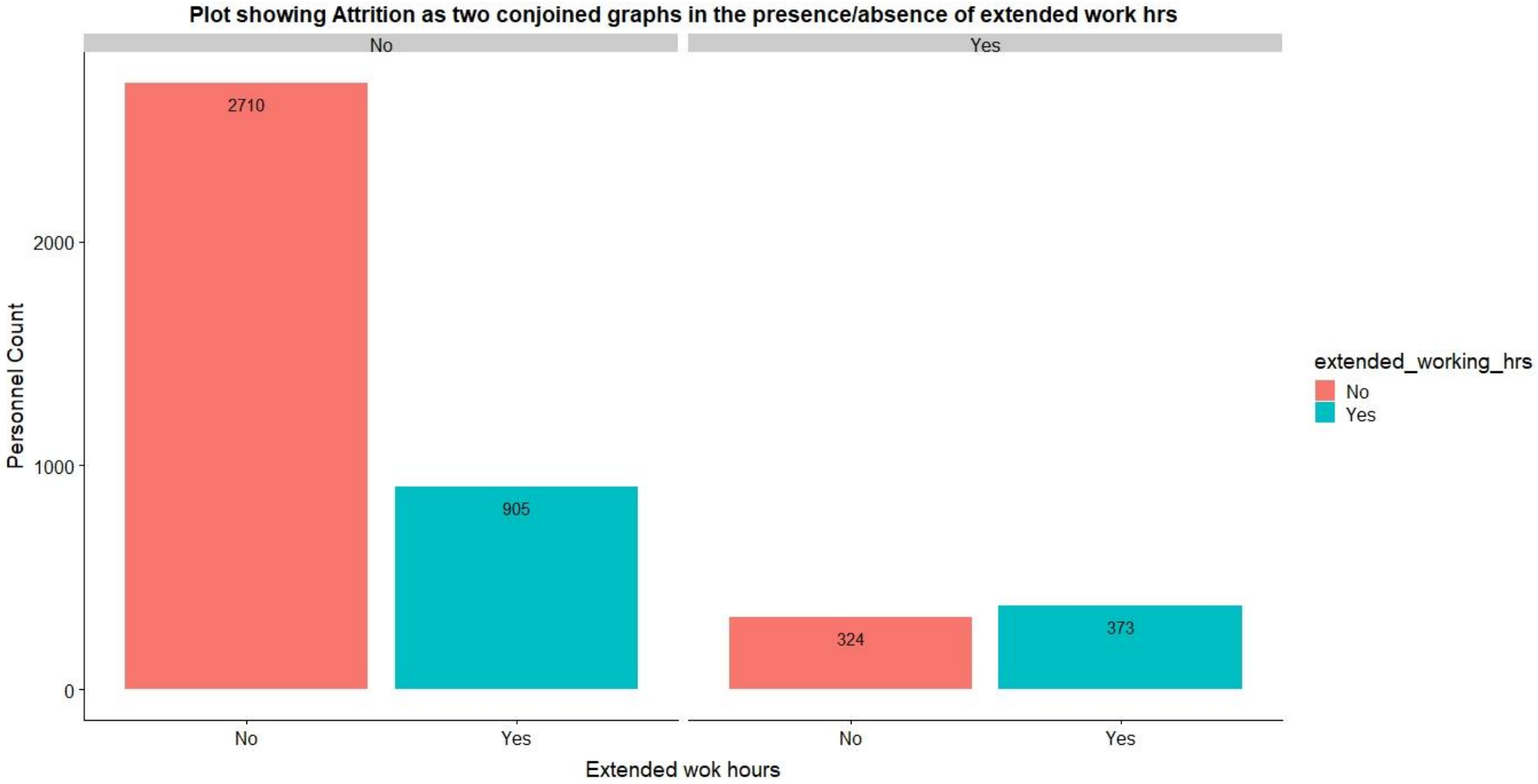
8) JobSatisfaction - surprisingly, the people who have rated the job satisfaction as being 'High' are the ones who leave the organisation most. People rating their Jobsatisfaction level to be Medium are the ones who stay on.Though this is strictly in terms of numbers & not as a percentage of the total people in that level. Percentage wise it is people with low jobsatisfaction who are leaving the most.

8) WorkLifeBalance - if considering the proportion of data, attrition is more in level 1 which is as expected.

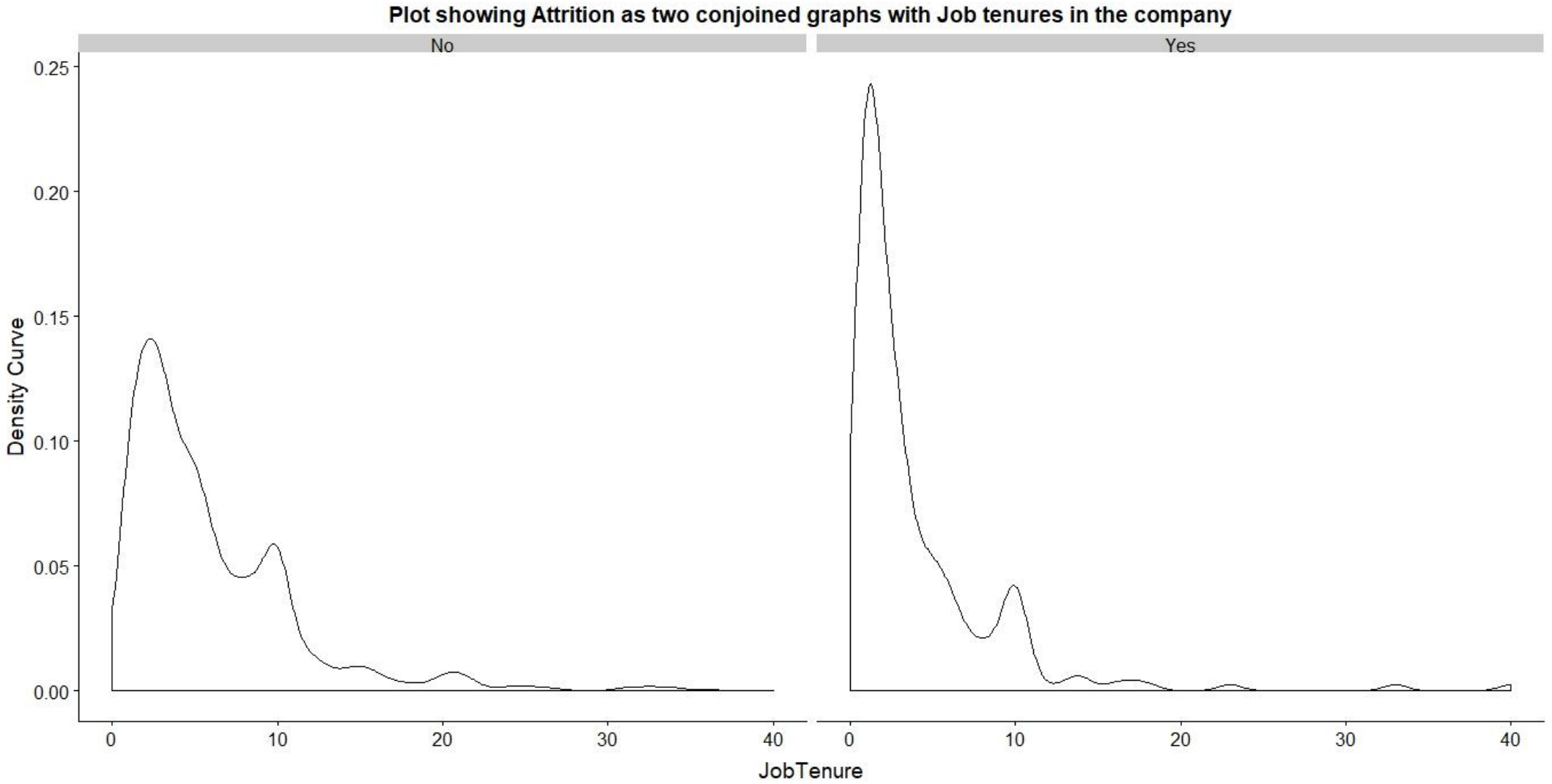
9) JobInvolvement - Attrition is more with involvement level 3

10) PerformanceRating - compared to rating 4, rating 3 has high attrition amount.

Analysis of continuous Variables - Graph1



Analysis of continuous Variables - Graph2



Analysis of continuous Variables

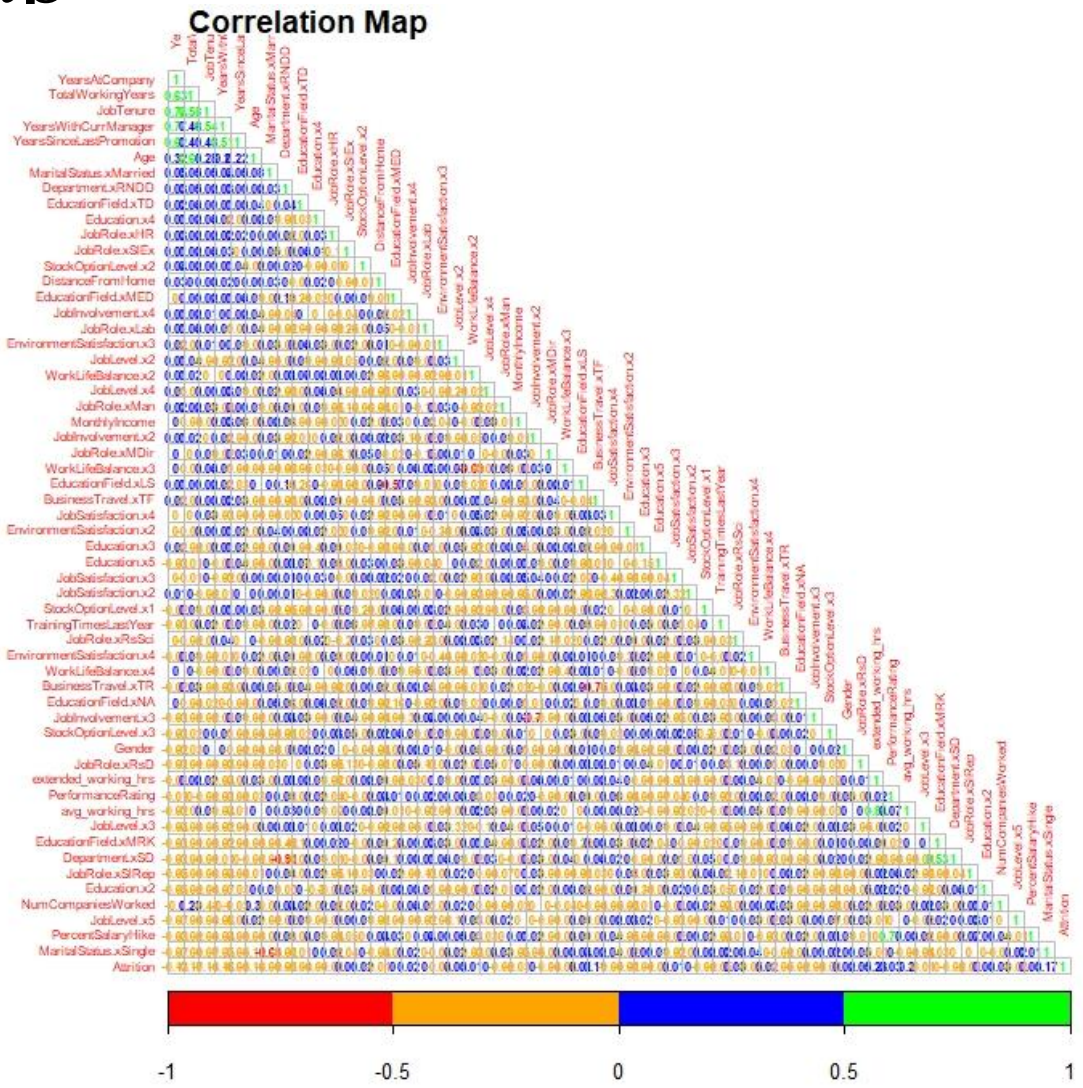
Graph -1

373 of the people who left the company, had to work for extended work hours. that is more than 53% of the people who left the company last year. This shows a clear trend that extra working hours is clearly not good if you are looking to retain people.

Graph -2

The density curve clearly shows the trend that more is the number of years spent in the company, less is the probability of the person leaving the organization. Once people have spent significant number of years working at a place, they find a good comfort zone there.

Correlation Map



Model Building

Pre-Modeling Stage:

•In this stage, we prepared the input variables which were to be given to the model.

•We normalized the continuous variables. We scaled them using scale function in R.

•We found the nature of the target variable to decide the model to be used.

•Here our target variable is Attrition which is categorical in nature and hence we used logistic regression model
•Attrition rate found in our dataset is 16% only.

•We created dummies for categorical variable as part of variable reduction technique

•We split the data into train and test data in ratio 7:3

Modeling Stage

•We used generalized linear model to build our model to predict the attrition

Model Building

- We used step AIC to get the starting model. Here we are predicting how Attrition behaves with respect to other variables. Since step AIC suggests the optimal model based on Akaike Information Criteria (model having lowest AIC value is preferred), there are some insignificant variables included in the model
- We used vif (Variable Inflation Factor) to check the correlation between the variables
- Finally we used p-values to remove insignificant variables.

Finalised Model

Coefficients:

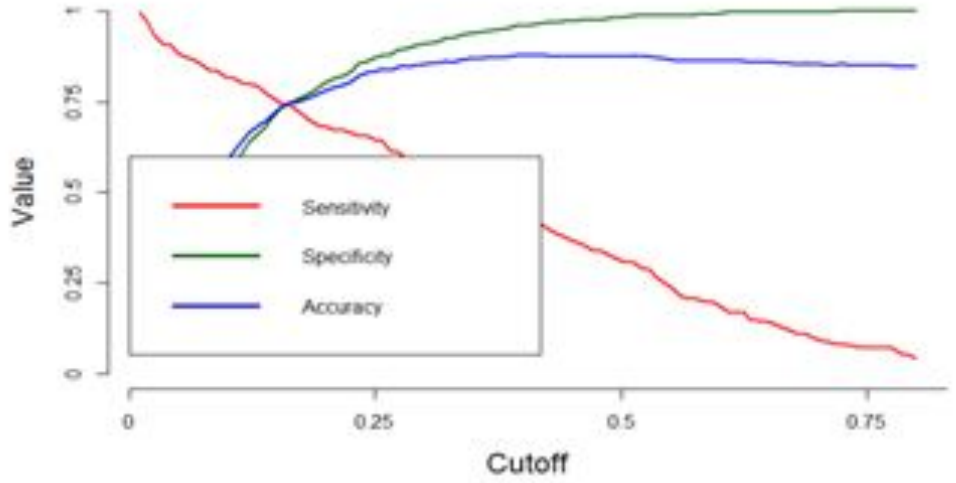
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.93256	0.13595	-14.215	< 2e-16	***
NumCompaniesWorked	0.34664	0.05625	6.163	7.16e-10	***
YearsSinceLastPromotion	0.55979	0.07684	7.285	3.21e-13	***
YearsWithCurrManager	-0.47339	0.08554	-5.534	3.13e-08	***
Age	-0.36046	0.07977	-4.519	6.22e-06	***
TotalWorkingYears	-0.52870	0.10655	-4.962	6.98e-07	***
BusinessTravel.xTF	0.66311	0.13290	4.989	6.06e-07	***
JobRole.xMDir	-0.75620	0.21820	-3.466	0.000529	***
MaritalStatus.xSingle	0.89456	0.11466	7.802	6.10e-15	***
EnvironmentSatisfaction.x2	-0.67864	0.16483	-4.117	3.84e-05	***
EnvironmentSatisfaction.x3	-0.91374	0.15371	-5.945	2.77e-09	***
EnvironmentSatisfaction.x4	-1.25831	0.16152	-7.791	6.67e-15	***
JobSatisfaction.x4	-0.87025	0.13478	-6.457	1.07e-10	***
extended_working_hrs	1.56824	0.11518	13.616	< 2e-16	***

- After tweaking each iteration of the model, we finally come to the end of this process on the 27th iteration. We chose to go ahead with this particular model as each of the variable involved seems to have very low p values.
- Also, upon checking the VIF score for the variables in the model, we see that the VIFs are also pretty low too.

TotalWorkingYears	YearsSinceLastPromotion	Age	YearsWithCurrManager
2.515633	1.871381	1.814571	1.766447
EnvironmentSatisfaction.x3	EnvironmentSatisfaction.x4	EnvironmentSatisfaction.x2	NumCompaniesWorked
1.578173	1.574430	1.491201	1.215362
extended_working_hrs	MaritalStatus.xSingle	JobSatisfaction.x4	JobRole.xMDir
1.075733	1.037950	1.023255	1.015869
BusinessTravel.xTF			
1.014974			

Evaluating the Model

We ran the model against the test dataset. We performed the following model validations - 1. Finding Accuracy, Specificity and Sensitivity through Confusion Matrix. In order to find a suitable probability cut-off, we checked the Accuracy, sensitivity and specificity for 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 ~0.18

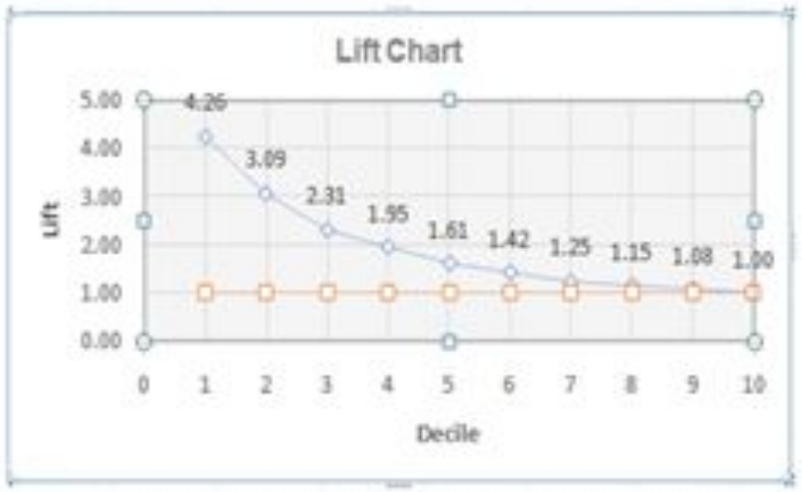


Predictior	No	Yes
No	809	54
Yes	276	155

Evaluating the Model

Calculating KS Statistics •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 model is 0.513

Gain and Lift Charts



Drawing Conclusions

1. **NumCompaniesWorked** - This seems to have a highly positive correlation value to attrition. Thus, as the number of companies worked for increases, the chances of the individual leaving the job increases too.
   Suggestion: It seems the new joinees who are being hired from another organisation, their expectation setting is not happening correctly. All new joiners must be oriented to the goal of the company & setting expectations in terms of compensation, work load & roles & responsibility must be done correctly.
2. **YearssinceLastPromotion** - As the gap between two promotions tend to increase, people tend to leave out the company.
   Suggestion: In non-promotion cycles, high performing employees must be shown appreciation. This can be done in a number of ways like arranging a high performer award, alloting points to the employee against which they can redeem some gifts. High performers should also be given higher bonuses.
3. **Yearswithcurrentmanager** - the variable is negatively correlated with Attrition. The more the number of years an employee spends under their current manager, the better the chances of the employee being retained.

   Suggestion: In order to ensure people stop moving out of the organisation, all the managers must be trained to be approachable & garner people skills. This should be more so of in the case of managers of single employees.

Drawing Conclusions

4. Age - As the age of an employee increases, their chances of leaving the organisation decreases. This is a little obvious as with age, people want to settle down in life.
Suggestion: retaining an employee in the initial 3-4 years of their employment is the key. Since according to the analytics, post that they would most probably retain. Hence, attractive packages, prospective growth should be projected early on to a fresher joining the company.

5. TotalWorkingYears - negatively correlated to attrition. More the number of years of experience, lesser the chances the employee would leave.
Suggestion: while recruiting newer candidates, a little more preference should be given to those with more experience. Those with few years of job experience can be observed to be prone to attrition.

6. BusinessTravel - Employees who travel frequently for Business have a positive coefficient with respect to Attrition. The company should look into the reasons why this occurs. Perhaps the employees are not compensated for their travel expenses or are not provided with satisfactory amenities during travel. It could also mean that these employees are finding opportunities to network with the business partners and are offered a job to join the partner organizations.
Suggestion: The HR Department can look into this factor and regulate the frequency and terms of Business Travel.

Drawing Conclusions (Contd.)

7. M.Director - Employees with the Manufacturing Director job role are least likely to Attrition due to the strong negative coefficient with Attrition.
Suggestion: HR could provide awards for Manufacturing Directors for their long service with the company so as to motivate employees from different job roles to continue working in the company.

8. MaritalStatusxSINGLE - has the strongest positive coefficient with respect to Attrition. This implies that employees who are single are at a higher risk of Attrition.
Suggestion: As this is not a parameter that can be regulated by the company we cannot provide an actionable insight on this parameter. However, the company may look into possibilities of introducing a term of employment bond with single employees to discourage them from attrition.

9. Environment Satisfaction - Low Work Environment Satisfaction has a strong positive coefficient with respect to Attrition. When employees feel that the working environment is not conducive or not meeting their expectation they are at a high risk of attrition.
Suggestion: The company may undertake a study to understand why employees have a low work environment satisfaction and implement remedial measures to improve the work environment. Additional team building sessions and amenities like canteen, indoor sports room or a leisure room can be integrated to improve the work environment.

Drawing Conclusions (Contd.)

10. Job Satisfaction - Low Job Satisfaction has a positive coefficient with respect to Attrition. This implies that employees who have a low job satisfaction are at a high risk of Attrition. The company must look into the reasons why employees are not satisfied with the current job role. Possible remedies could be integrating additional responsibilities for the individuals with low job satisfaction or introducing department level competitions for performance related activities and periodically present awards to employees based on different parameters. Employees with a very high Job satisfaction are less likely to Attrition due to the negative co-efficient. HR could study the traits responsible for very high job satisfaction and assimilate the same for the employees with low job satisfaction.

THANK YOU!