Human Resource Analytics Case Study

SUBMISSION

Logistic Regression Model to classify employees likely to Attrition and determine factors driving Employee Attrition

1. Case Study Overview

CONTEXT

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.

PROBLEM

The management believes that the level of **attrition** i.e. 15% is bad for the company, and adversely impacts the business and reputation of the organization in the following ways:

- It causes delay in completion of projects formerly undertaken by employees who later attrition making it challenging to meet timelines. This harms the reputation of the organization leading to loss in customer and market share.
- Investment of resources and manpower to maintain a department focused on sourcing and hiring new talent to fill abruptly vacated positions in the organization.
- Requirement to train newly acquired talent to perform effectively in the organization. Delays caused due to time allocated for newly acquired talent for acclimatizing in the new work environment.

OBJECTIVE

The aim is to identify the factors which have the most probability for an employee to attrition. 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.

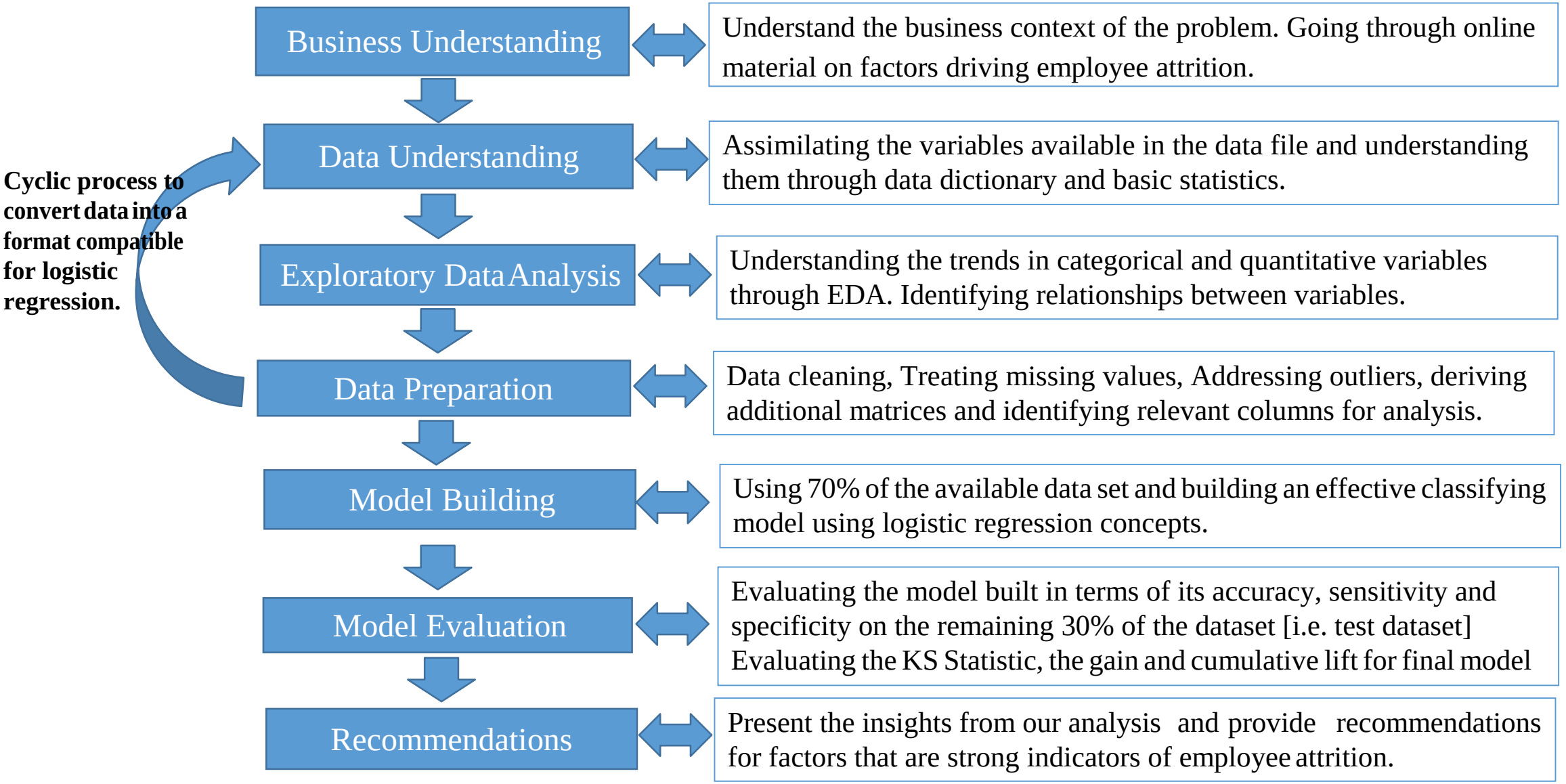
METHODOLOGY AND DELIVERABLES

[1] Determine and understand the driver variables influencing high attrition rate within the organization.

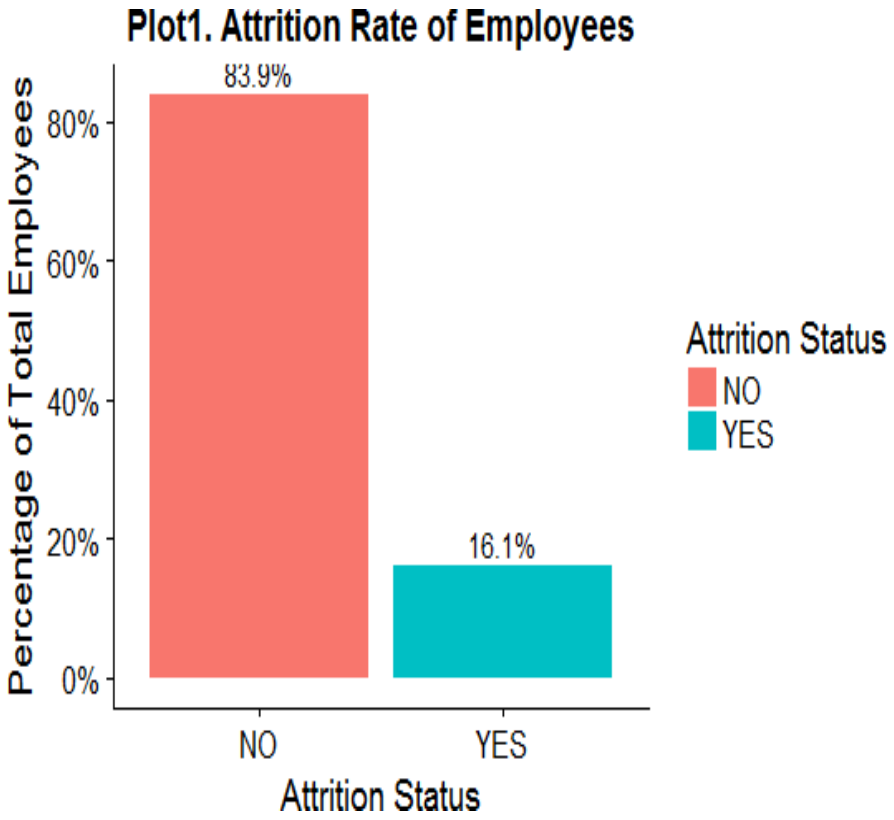
[2] Build a model to estimate the probability of an employee to attrition.

[3] Provide actionable insights for inducing changes within the organization so as to curb attrition and encourage employees to stay.

Problem Solving Methodology



2. Business Overview



Insight1: From the above plot it is Clear that 16.1% of the 4410 employees from the dataset have attrition. This is a cause for concern.

4. Data Understanding

There are five tables available for analysis:

[1] general data= This contains personal details, demographics, educational data, career relevant information and wages for each employee.

[2] employee survey= This contains information regarding a few scaled categorical parameters which were filled by the employees within the organization.

[3] manager survey= This contains information regarding scaled parameters used by the managers within the organization to rate the employees

[4] in punch= This contains time and date information for the year 2015 registered when the employee began work for the day

[5] out punch= This contains time and date information for the year 2015 registered when the employee ended work for the day

The in punch and out punch datasets have been used to derive 5 metrics:

[1] Tot.logged.hours- Total hours worked in 2015.

[2] Avg.logged.hours- Average working hours logged per day.

[3] Tot.leaves.taken- Total leaves taken

[4] Tot.excess.logged- Total hours worked in excess of standard [i.e, 8hrs/day]

[5]Avg.excess.logged- Average hours worked in excess of standard [i.e,8hrs/day]

5. Assumptions and Data Handling

[1] EmployeeID- Has been used as the primary key to merge the 5 individual tables. For the **in_punch** and **out_punch** tables the first column name is missing. It has been replaced with EmployeeID

[2] In_Punch and Out_Punch- All the date and timestamp records of in_punch and out_punch have been converted into the standard yyyy-mm-dd HH:MM:SS format and the difference of out_punch and in_punch for a given day is stored as total working hours for that day in the table worked_hours. Columns with all missing values have been treated as holidays and been removed using the **datachop function.** Five metrics have been derived from this table as relevant for analysis Tot.logged.hours, Avg.logged.hours, Tot.leaves.taken, Tot.excess.logged and Avg.excess.logged.

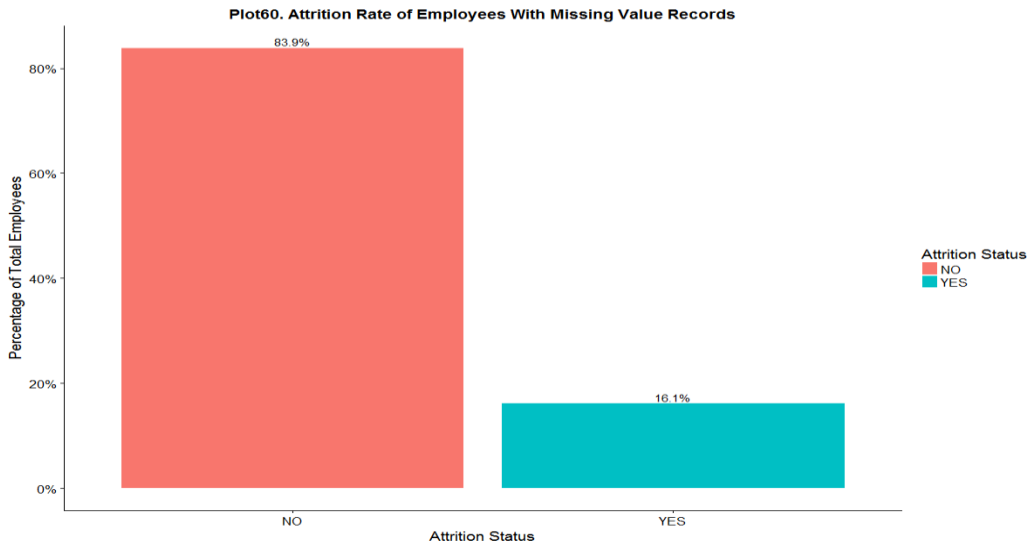
[3] Data Cleansing- Removed EmployeeCount, Over18 and Standard Hours columns which are all NAs or with only one unique value as the lack of variability will not contribute to any useful insights. The ***datachop function*** checks all the records of each column for more than 1 unique value, if there is only one unique value or NA it deletes that column from the main dataset.

[4] Duplication Checks- Data Duplication checks have been performed.

[5] We will convert all character attribute records to upper case to avoid any case sensitive inconsistencies and data entry discrepancies. The case conversion will be done using a custom defined *function caseconversionfun*.

[6] Outlier Treatment- Outliers for all numerical variables have been performed by observing the percentile variation. Outlier records have been capped appropriately.

[7] Missing Value Treatment- The 110 records with missing values have been removed from the dataset. **The missing records account for less than 2.5% of the dataset and does not impact the homogenous nature of the dataset. The plots below show that the attrition status varies by 0.1% after removing the missing records. This is justifiable.**



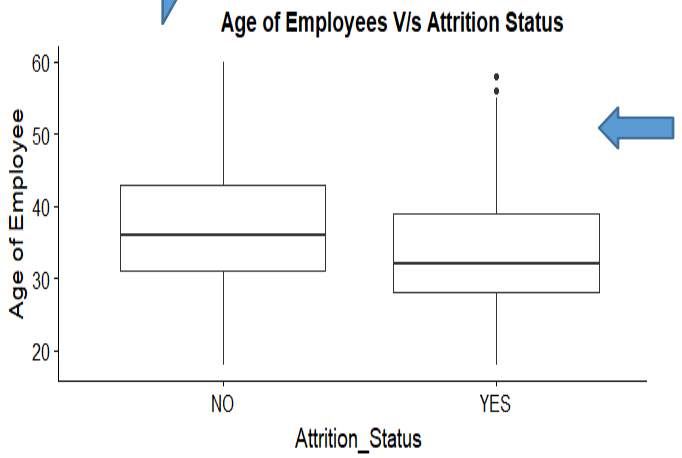
6. Final Model

Varible Type	Variable	Coefficients	Std. Error	z value	Pr(>|z|)	VIF
Dependent Variable	**Attrition**					
Intercept	(Intercept)	-2.37743	0.12955	-18.352	< 2E-16	
Independent Variables	1. Age	-0.29295	0.07883	-3.716	0.000202	1.803
	2. NumCompaniesWorked	0.30926	0.05881	5.259	1.45E-07	1.236
	3. TotalWorkingYears	-0.56516	0.10482	-5.392	6.97E-08	2.419
	4. TrainingTimesLastYear	-0.21102	0.05789	-3.645	0.000267	1.025
	5. YearsSinceLastPromotion	0.63178	0.0766	8.248	< 2E-16	1.877
	6. YearsWithCurrManager	-0.52902	0.08594	-6.155	7.49E-10	1.844
	7. Tot.logged.hours	0.64836	0.05309	12.214	< 2E-16	1.058
	8. BusinessTravel.xTRAVEL_FREQUENTLY	0.66585	0.13127	5.072	3.93E-07	1.020
	9. JobRole.xMANUFACTURING.DIRECTOR	-0.82601	0.21633	-3.818	0.000134	1.025
	10. MaritalStatus.xSINGLE	0.91825	0.11481	7.998	1.27E-15	1.056
	11. EnvironmentSatisfaction.xLOW	0.80542	0.13088	6.154	7.57E-10	1.037
	12. JobSatisfaction.xLOW	0.58144	0.13747	4.229	2.34E-05	1.143
	13. JobSatisfaction.xVERY_HIGH	-0.54993	0.13689	-4.017	5.89E-05	1.138
	14. WorkLifeBalance.xBETTER	-0.38657	0.11302	-3.42	0.000625	1.023

Insights

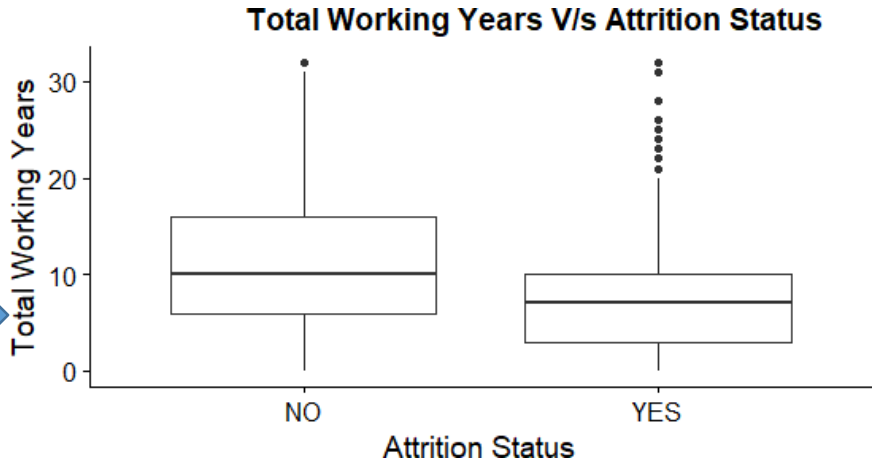
- After building 25 models we arrived at the final model having 14 independent variables and 1 intercept to estimate the probability of Attrition.
- The VIF value for all the independent variables is less than 3. This implies low co-linearity between predictor variables
- The Significance value of each variable is high and therefore no more variables can be removed without impacting the performance of the model.
- We will evaluate this final model with the preceding model_23 show the variation in Evaluation Metrics.

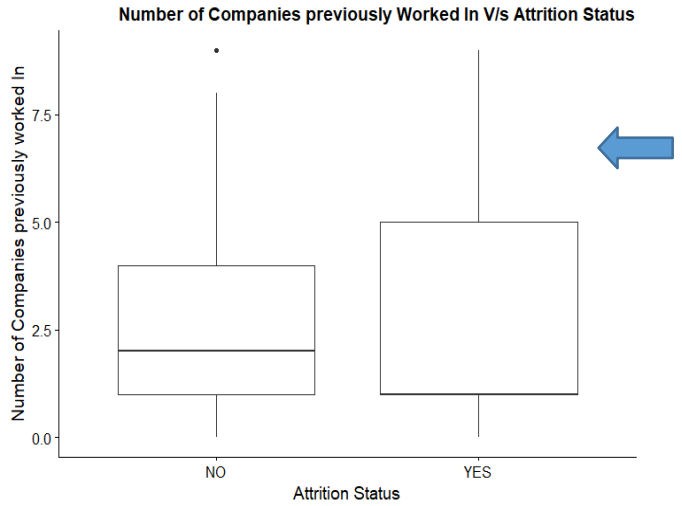
7. Understanding the significant variables and their Impact on Employee Attrition



Attrition rate is higher with employees have lesser ***age*** as shown in the graph to the left.

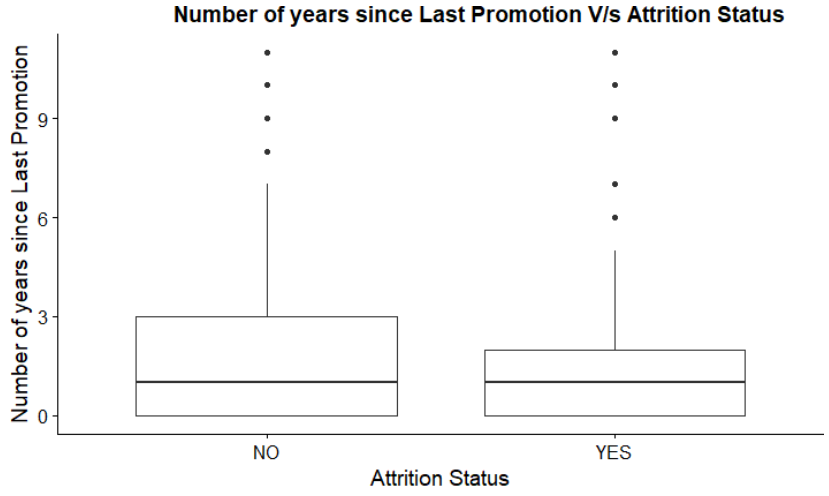
TotalWorkingYears vs attrition graph shows that if TotalWorkingYears increases the likelihood of attrition will decrease. Senior employees are less likely to attrition.

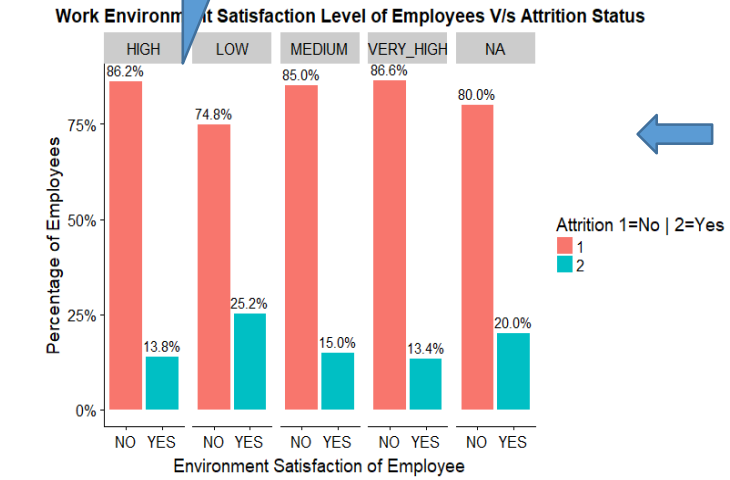




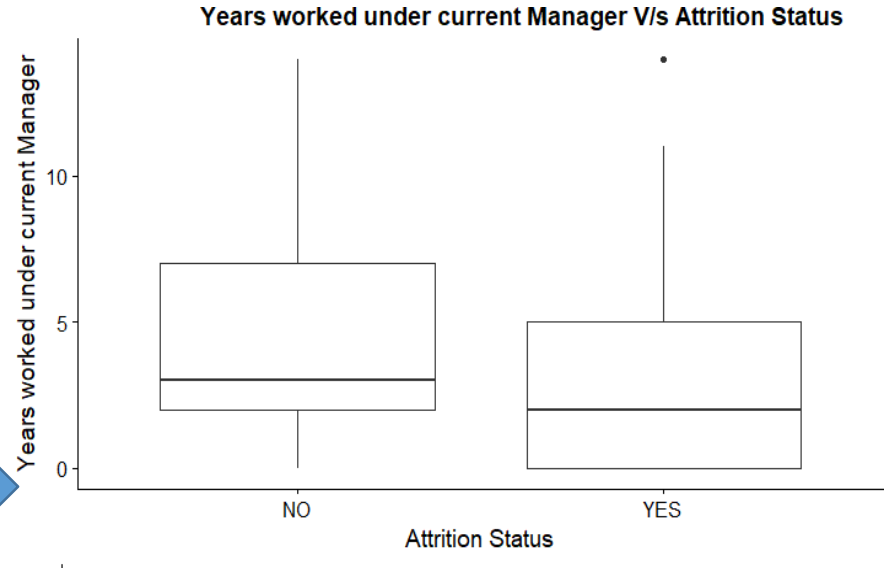
Number of Companies vs attrition graph imply that if an employee has previously worked in several companies they are more likely to attrition.

YearsSinceLastPromotion graph implies that if an employee has not been granted a promotion recently they are at a higher risk of attrition.

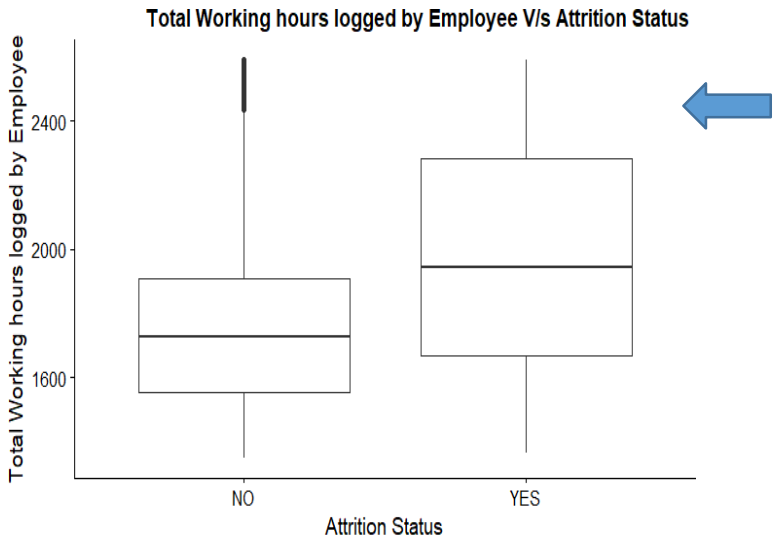




A low ***EnvironmentSatisfaction*** is strong factor implying that employees with a low work environment satisfaction are a very high risk of attrition.

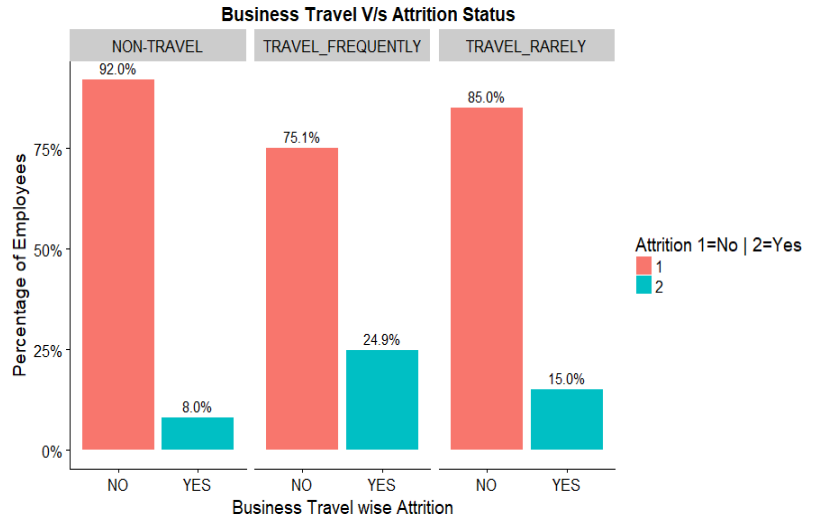


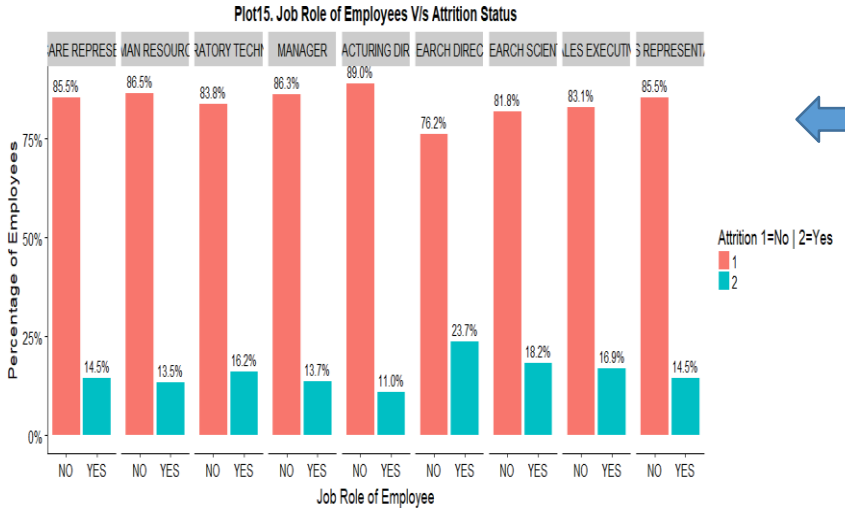
YearsWithCurrManager vs attrition implies that if there is a frequent change in management then there will be an increased risk of attrition.



Tot.logged.hours is an interesting parameter that shows if an employee is over-worked or not granted enough leaves he is more likely to attrition.

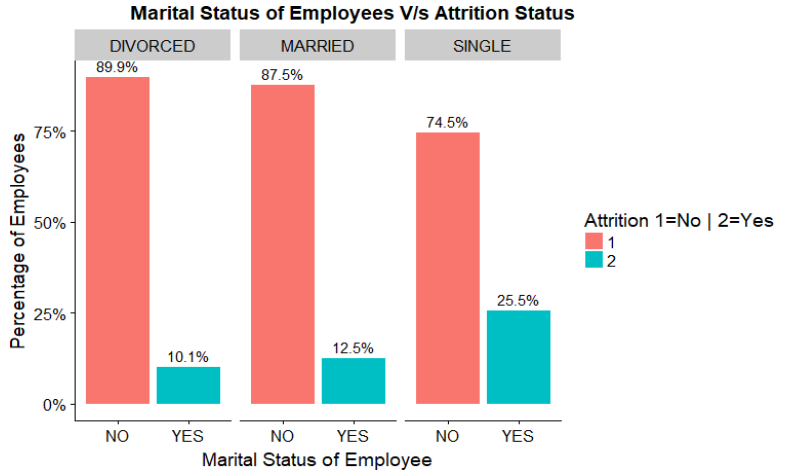
Under ***Business Travel*** Employees who *TRAVEL_FREQUENTLY* are the ones who are likely to attrition.

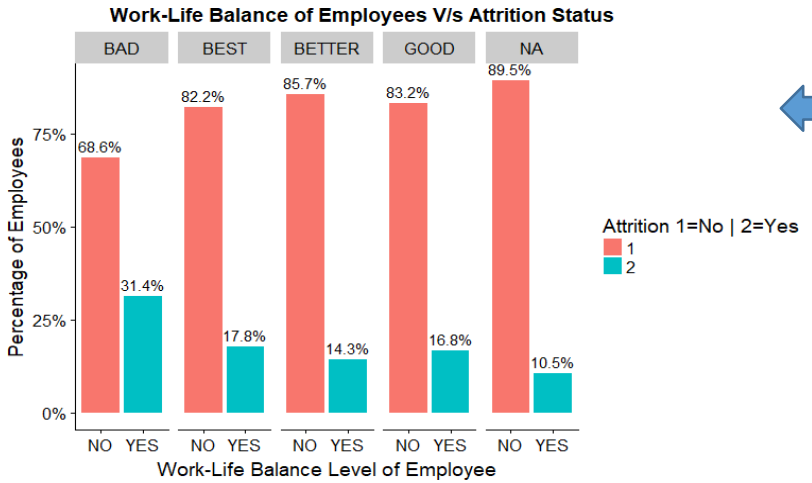




Under ***JobRole*** MANUFACTURING.DIRECTOR are least likely to attrition.

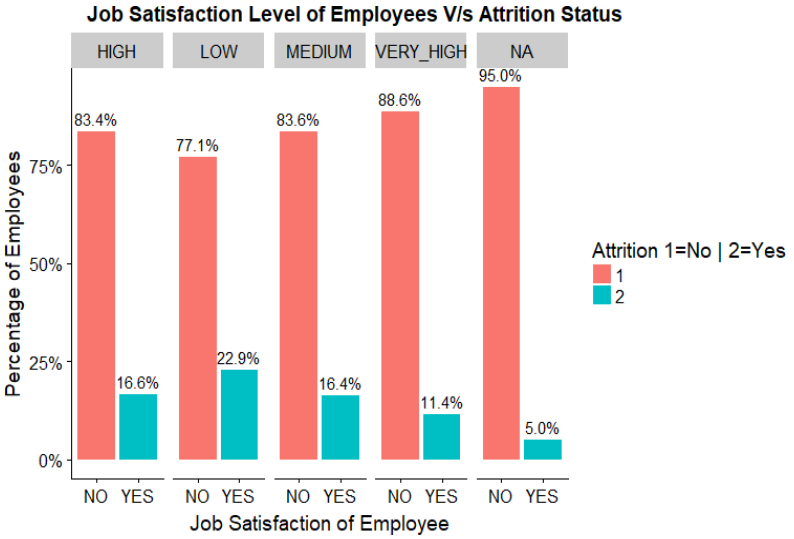
Under ***MaritalStatus*** employees who are *SINGLE* are at a higher risk of attrition.





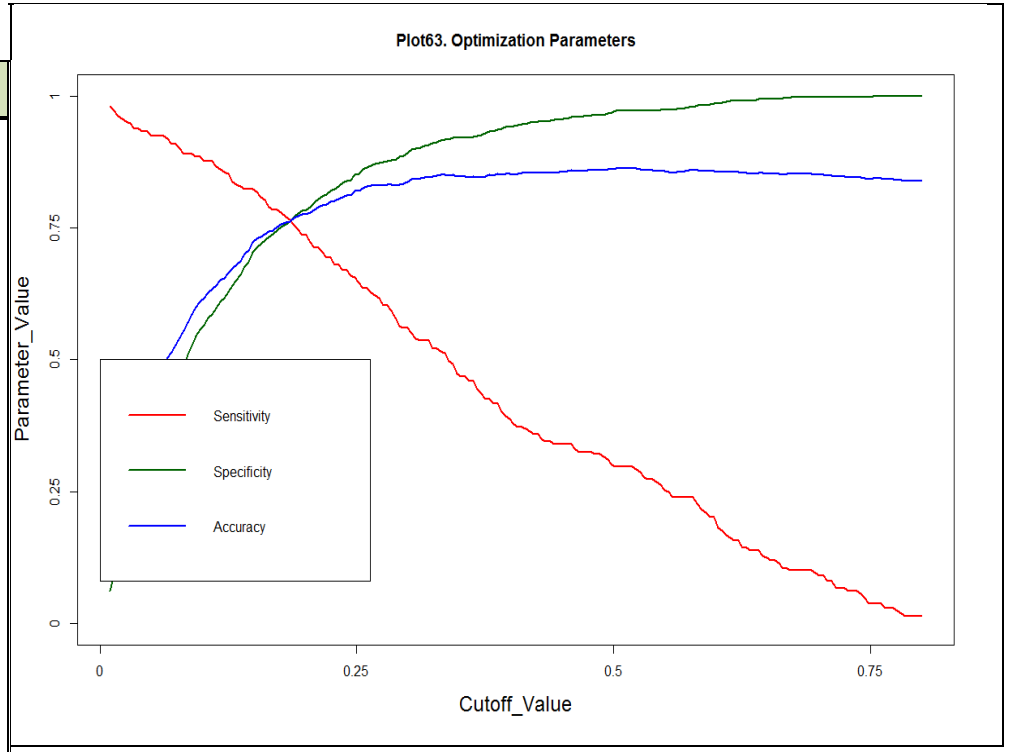
Under ***WorkLifeBalance*** employyes having *BETTER* are the least who attrition and the employees with BAD work life balance are the highest who attrition.

Employees with ***JobSatisfaction*** *VERY_HIGH* are ones who are less likely to attrition.



8. Model Evaluation

Parameter	Model 23	Final Model [Model 25]
Variables	. (Intercept) 1. Age 2. NumCompaniesWorked 3. TotalWorkingYears 4. TrainingTimesLastYear 5. YearsSinceLastPromotion 6. YearsWithCurrManager 7. Tot.logged.hours 8. BusinessTravel.xTRAVEL_FREQUENTLY 9. JobRole.xMANUFACTURING.DIRECTOR 10. MaritalStatus.xSINGLE 11. EnvironmentSatisfaction.xLOW 12. JobSatisfaction.xLOW 13. JobSatisfaction.xVERY_HIGH 14. WorkLifeBalance.xBEST 15. WorkLifeBalance.xBETTER 16. WorkLifeBalance.xGOOD	. (Intercept) 1. Age 2. NumCompaniesWorked 3. TotalWorkingYears 4. TrainingTimesLastYear 5. YearsSinceLastPromotion 6. YearsWithCurrManager 7. Tot.logged.hours 8. BusinessTravel.xTRAVEL_FREQUENTLY 9. JobRole.xMANUFACTURING.DIRECTOR 10. MaritalStatus.xSINGLE 11. EnvironmentSatisfaction.xLOW 12. JobSatisfaction.xLOW 13. JobSatisfaction.xVERY_HIGH 14. WorkLifeBalance.xBETTER
Cutoff Value	0.1852	0.1847
Accuracy	0.7674	0.7620
Sensitivity	0.7751	0.7656
Specificity	0.7660	0.7641
KS Statistic	0.5411	0.5269
Gain till 4th Decile	83.2536	82.2967
Cumulative Lift till 4th Decile	2.0813	2.0574



Insight: We conducted model evaluation between Model 23 and our Final Model. Model 23 has 16 independent predictor variables while the Final Model has 14 independent variables.

The difference in evaluation parameters such as Accuracy, Sensitivity and Specificity is negligible. The KS Statistic varies by less than 1.5%. While gain and lift parameters for the 4th Deciles is similar. **Therefore we selected Model 25 as the final model as it has fewer predictor variables and almost negligible loss in Performance Metrics.**

The above graph was used to determine the Optimal Cutoff value for the final model as 0.1847.

9. Insights and Suggestions

1. 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. 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.

2. MaritalStatus.xSINGLE has the strongest positive coefficient with respect to Attrition. This implies that employees who are single are at a higher risk of Attrition. **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.**

3. 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. **The HR Department can look into this factor and regulate the frequency and terms of Business Travel**.

4. Tot.logged.hours has a positive coefficient with respect to Attrition. This implies that employees who log excess working hours with respect to the standard are more likely to Attrition. **This is an indication of overutilization of employees or irregular distribution of workload. Employees who are forced to work longer hours throughout the calendar year are at a high risk of Attrition. HR should look into the working hour's information to find the reasons behind excess working hours and regulate it by spreading the workload or providing monetary incentives for complying with excess workload if it is un-avoidable.**

5. YearsSinceLastPromotion has a positive coefficient with respect to Attrition. This implies that as the number of years since previous promotion increases the risk of attrition increases. **This is an indicator of an employee feeling de-motivated for not receiving a promotion for a significant duration leading to the decision to attrition. The HR should look revising the employee appraisal process so as to justly evaluate the employee's performance. Providing periodic promotions or economic incentives might encourage employees to stay.**

6. 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.

7. Better Work-Life Balance has a positive coefficient with respect to Attrition. **Implying that employees who have a work-life balance of better or best are less likely to attrition**. HR could look into why employees rate work-life balance as bad. Possible remedies could be to arrange for outdoor leisure sessions at local resorts or nearby vacation locations. The other possibilities could be discouraging employees from working excess hours frequently so that they may have spare time to spend with family and friends.

8. YearsWithCurrManager has a negative coefficient with respect to Attrition. Implying that employees develop a working bond with their managers and are less likely to Attrition if the manager is not replaced frequently. **The HR should refrain from frequent management changes or team leader changes as employees might find it difficult to acclimatize with frequent changes of their management.**

9. 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.

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

11. Training Times Last Year has low but significant negative coefficient with respect to Attrition. Implying that if an employee is sent for frequent training sessions they are less likely to attrition. The HR could look into this parameter and conduct training sessions for those employees who have not been given frequent training sessions. This might motivate the employees to find better job satisfaction as they will gain new skills from these training sessions.

Thank You!