HR
CASE STUDY

ADITHYA K
PRIYA LOBBO

PROBLEM STATEMENT

ABSTRACT

☐ 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 IS BAD 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 ACCLIMATIZE THEMSELVES TO THE COMPANY

CASE STUDY OBJECTIVE

- FACTORS THEY SHOULD FOCUS ON, IN ORDER TO CURB ATTRITION.
- MOST IMPORTANT ASPECTS THAT CAN BE ADDRESSED RIGHT AWAY.

METHODOLOGY – CRISP DM FRAMEWORK

1. Business Understanding

2. Data Understanding

3. Data Preparation

4. Modelling

5. Model Evaluation

6. RECCOMENDATION BASED ON MODEL

DATA UNDERSTANDING

DATA SOURCE – 5 DATA FILES

- GENERAL_DATA.CSV: Contains employee's demographic data and other work details like job role, travel, etc (4410 observations)

- EMPLOYEE_SURVEY_DATA.CSV: Contains survey data of employees, like satisfaction, etc (4410 observations)

- MANAGER_SURVEY_DATA.CSV: Contains survey data based on manager's perspective of employees performance, etc (4410 observations)

- IN_TIME.CSV: Contains office in time data of the employee for a year, in YYYY-MM-DD date time format

- OUT_TIME.CSV: Contains office in time data of the employee for a year, in YYYY-MM-DD date time format

Data Cleaning

Ensure no duplicate data exists

Remove columns where for ALL rows the value is NA, 1, 0, "", identical values.

Convert numeric fields required for analysis like rate and term to numeric values

Change Date and time to proper format, required for analysis
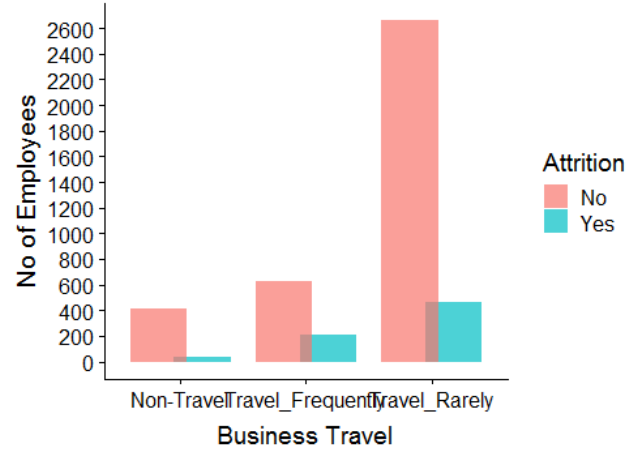
Convert to factors

Outliers Analysis was done and due to negligible number of outliers , they were not removed from data
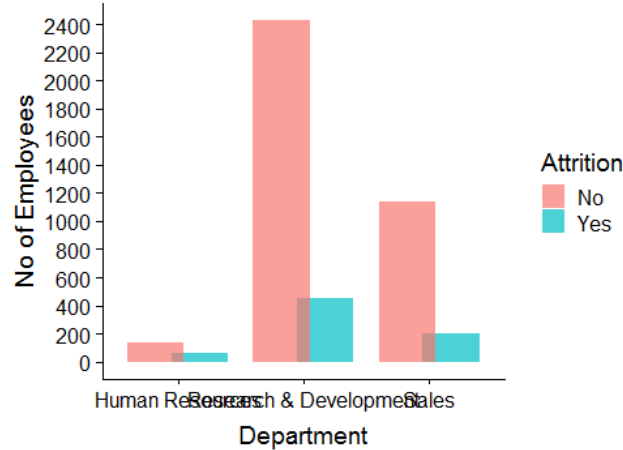
DATA ANALYSIS 1

OBSERVATION

- "Frequent" – "Business Travel" has higher percentage of Attrition (~ 25%)

- The "HR department" appears to have a comparatively higher Attrition rate

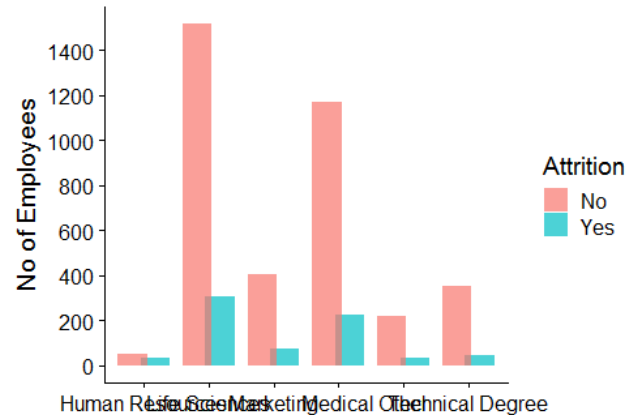- Attrition among Males/Female seems to be balanced.

DATA ANALYSIS 2

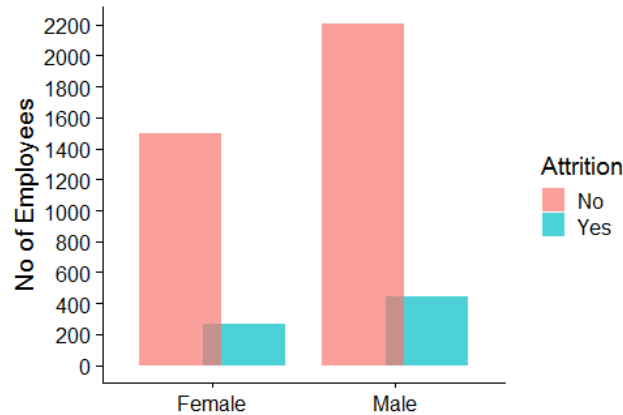JobRole & Attrition

MaritalStatus & Attrition

StockOptionLevel & Attrition
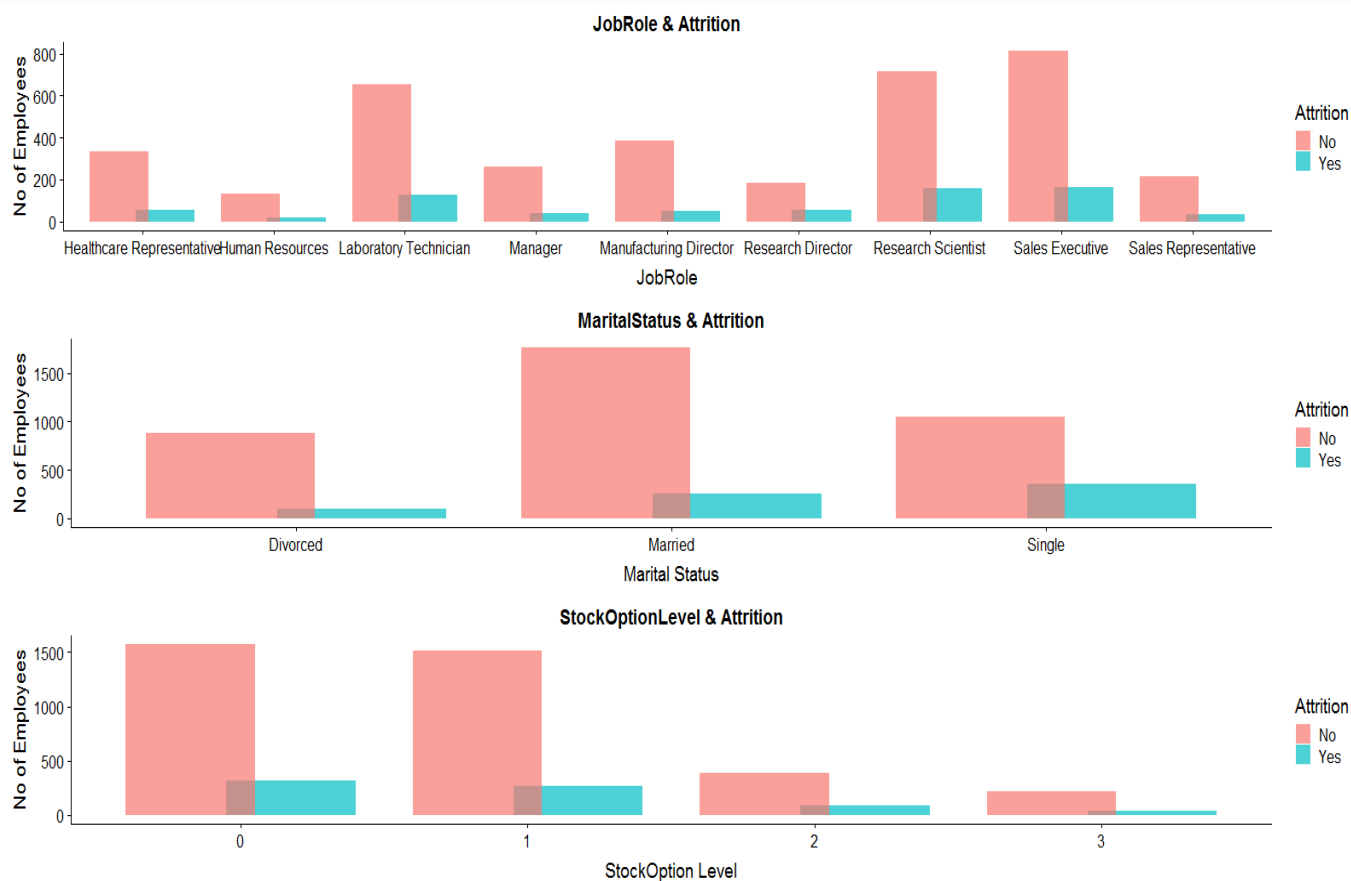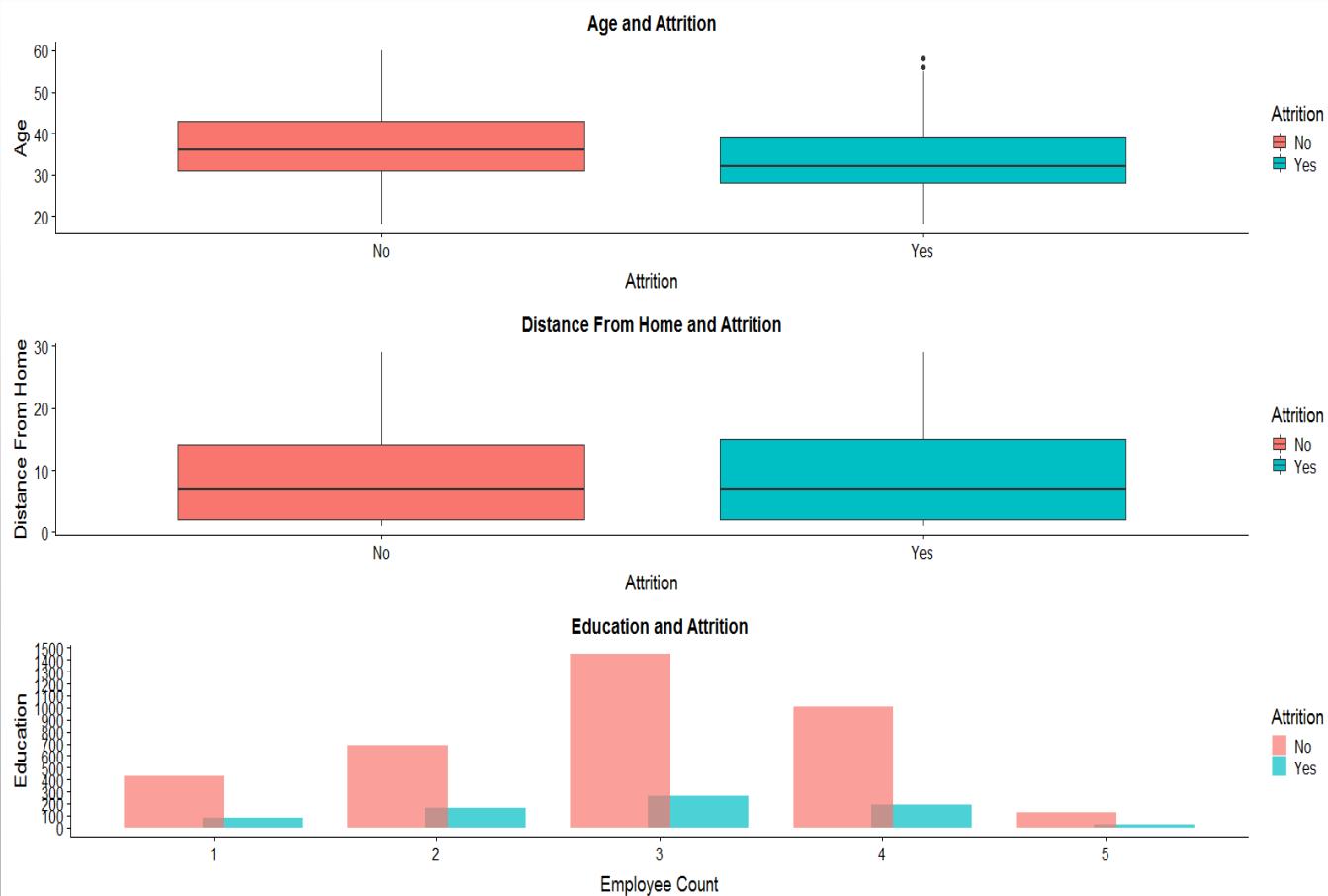
OBSERVATION

- Employees with "Single", "Marital Status" has higher Attrition rate

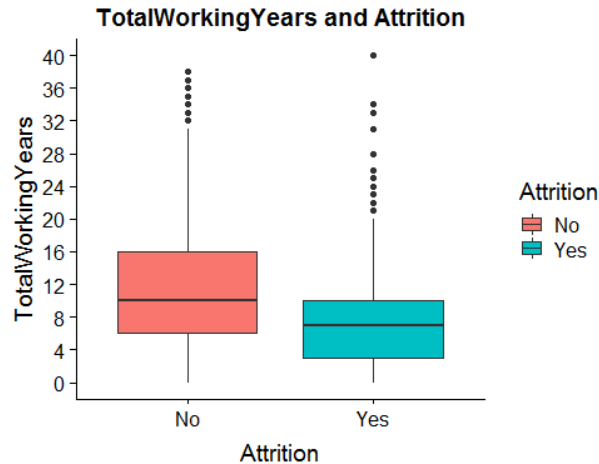- No significant impact of "Job Role" & "Stock option" observed on Attrition

DATA ANALYSIS 3

OBSERVATION

- Lower "Age" Employees tend to have higher Attrition rate

- "Distance From Home" and "Education" do not seem to effect Attrition

DATA ANALYSIS 4
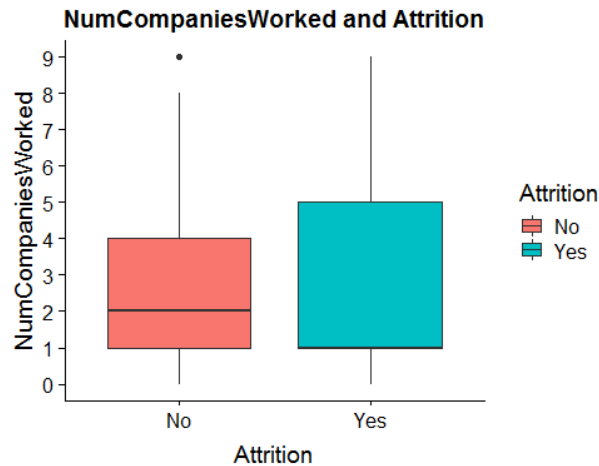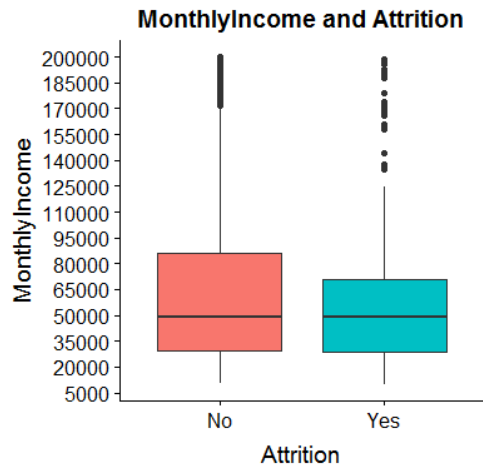
OBSERVATION

- Employees who have worked in more "Number of companies" have higher Attrition rate

- The "Total Working Years" of an employee also impact Attrition inversely

- "Monthly Income" and "Percentage Salary Hike", surprisingly don't seem to impact Attrition

DATA ANALYSIS 5

TrainingTimesLastYear and Attrition

YearsAtCompany and Attrition
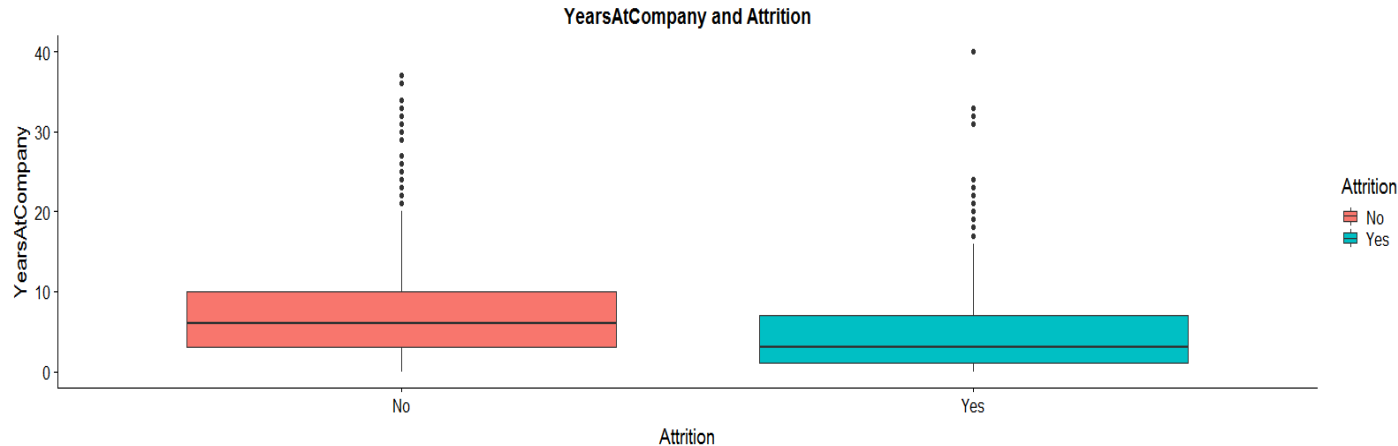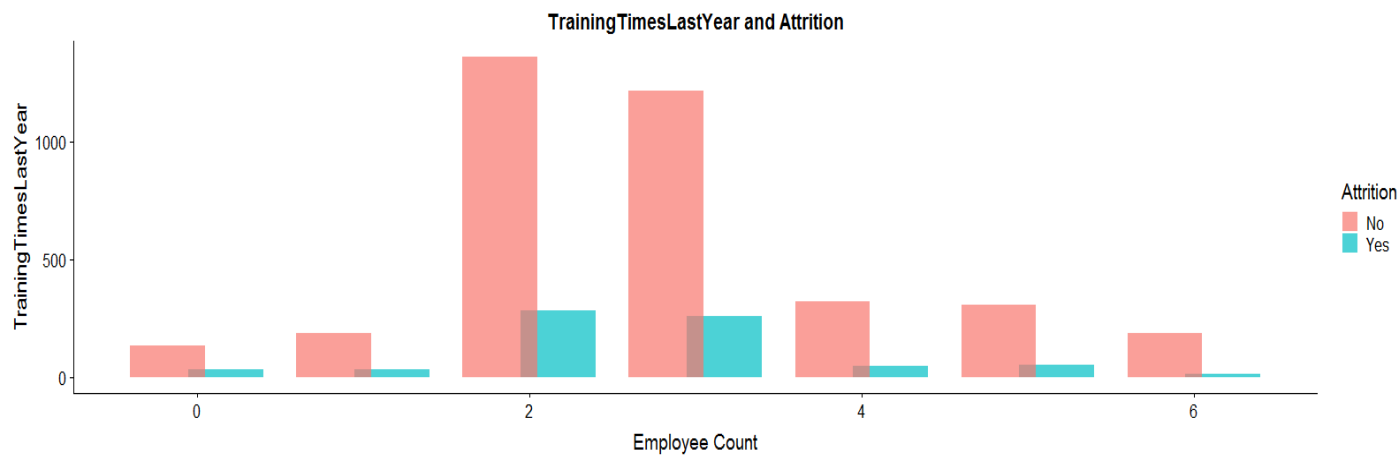
OBSERVATION

- Employees who have lesser number of "Years at the Company" have higher Attrition rate

- "Training Times Last Year" does not show any significant impact on Attrition

PREPARING FOR MODELLING

Preparing Variables for Modelling

    a.   "**Attrition**" is the **Target Variable** used for this model, as the objective of this analysis to find the factors that can be used to curb Attrition in the company

    b.   A total of **24 variables,** collected from the 5 Excel data source files provided, were used to predict the probability of an employee leaving the company

    c.   Continuous variables are **normalised using the "scale" function** to ensure a more standardised set of values in the model

    d.   Categorical variables are **converted to numeric using dummy variables** for regression modelling

Preparing Data Sets for Modelling

    The master data set is split into **Test data and Train data** in the ratio 7:3 respectively

MODELLING

1. **Logistic Regression Modelling** is used as the target variable "Attrition" is of categorical type

2. The **glm (Generalised Linear Model) function** is used to build the model on the Train Data Sets for Modelling

3. **Step AIC function** is used on the first model to obtain an optimum model based on Akaike Information Criteria(AIC). This suggests variables significant for predicting "Attrition"

4. **VIF (Variable Inflation Factor)** is used to check for multicollinearity between variables

5. An **Iterative approach** is used to building models using the glm function, where in each iteration, variables are **removed if they have 'less significance' (p-value > 0.05) and high multicollinearity (VIF > 3)**, until a final model is obtained

6. **The final model selected has 14 variables with high significance (negligible p-value) and low multicollinearity (VIF<3)**

7. **The Logistic Regression Model thus built, correctly identifies 77% of employees that are likely to churn And identifies employees that are not likely to churn, with 77% accuracy**

FINAL MODEL – 14 SIGNIFICANT VARIABLES

Coefficients

```
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -1.47845    0.14363 -10.294  < 2e-16 ***
Age                                -0.30841    0.07880  -3.914 9.09e-05 ***
NumCompaniesWorked                  0.31057    0.05821   5.336 9.52e-08 ***
YearsSinceLastPromotion             0.59023    0.07518   7.851 4.13e-15 ***
YearsWithCurrManager               -0.51118    0.08440  -6.057 1.39e-09 ***
timespend_median                    0.65077    0.05231  12.441  < 2e-16 ***
TotalWorkingYears                  -0.52654    0.10646  -4.946 7.58e-07 ***
EnvironmentSatisfaction.x4         -0.46178    0.12533  -3.685 0.000229 ***
JobSatisfaction.x2                 -0.59157    0.17007  -3.478 0.000505 ***
JobSatisfaction.x3                 -0.51884    0.14729  -3.522 0.000428 ***
JobSatisfaction.x4                 -1.12897    0.15971  -7.069 1.56e-12 ***
WorkLifeBalance.x3                 -0.40639    0.11199  -3.629 0.000285 ***
BusinessTravel.xTravel_Frequently   0.68691    0.13084   5.250 1.52e-07 ***
JobRole.xManufacturing.Director    -0.83088    0.21308  -3.899 9.64e-05 ***
MaritalStatus.xSingle               0.90873    0.11386   7.981 1.45e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
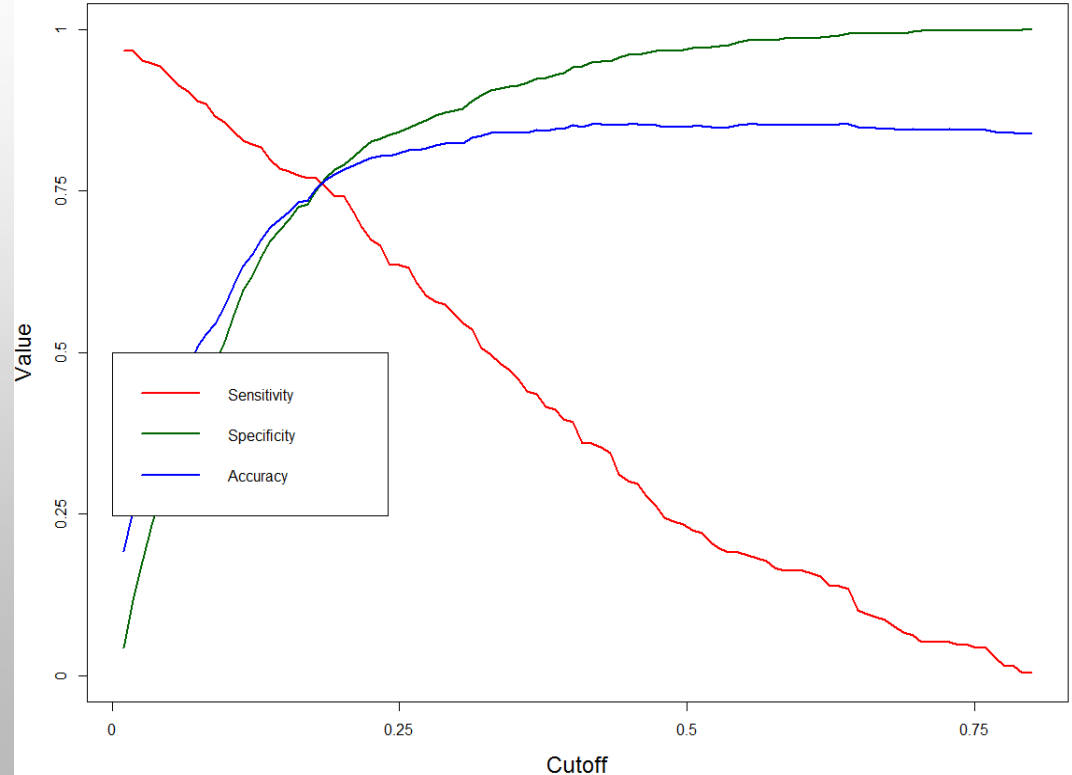
MODEL EVALUATION - CONFUSION MATRIX

1. The Final Model was **run on the Test data set** to predict Attrition

2. This model was validated using the **Confusion matrix:**
Accuracy = 0.751938
Sensitivity = 0.77 03349
Specificity = 0.7483811

3. Cut off at intersection of Accuracy, Sensitivity and Specificity gives a **balanced model**

4. The **optimum cut-off probability is found to be .19**

MODEL EVALUATION – KS STATISTIC

- The **KS Statistic of the model is found to be .519 or 51.9**%

- It **lies in the 3rd Decile (top 30**%**),** implying that it is beneficial to target 30% of the employees most likely to leave, and work on making them stay (Targeting fewer employees will not identify enough employees likely to leave or Targeting more employees will be inefficient)

- KS Statistic is considered **Optimal** if it is between 40-60 & lies within the first 3 Deciles, both of which are **satisfied in this model**

MODEL EVALUATION – GAIN & LIFT CHARTS
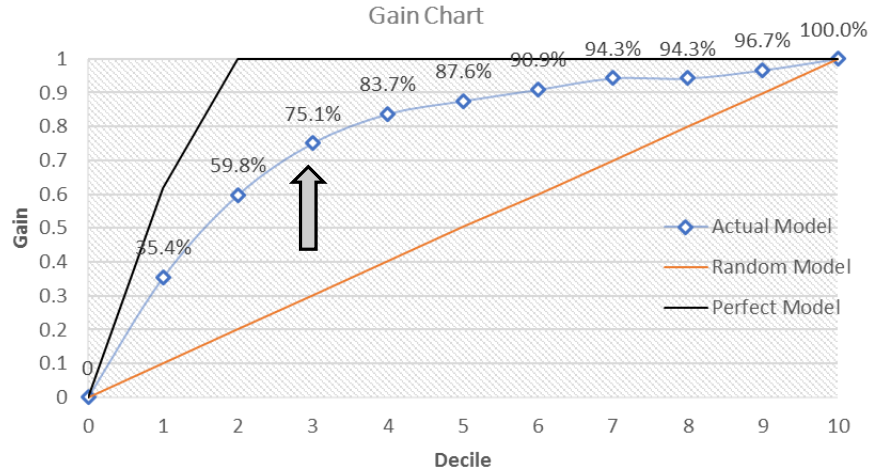
GAIN is 75% in the 3rd Decile

- This model is able to identify 75% of the employees likely to leave in the first 3 deciles (top 30% of a list sorted according to probability)
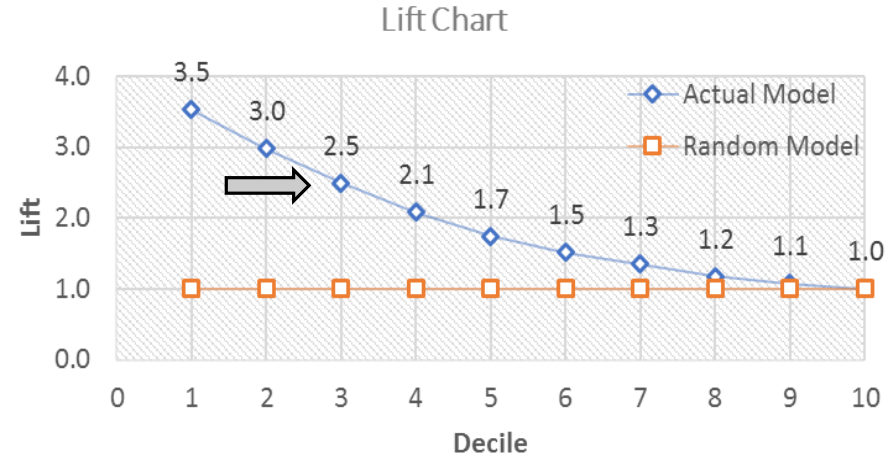
LIFT is 2.5

- This model offers a "lift" of 2.5, implying that the model identifies 2.5 times more attrition than a random model

RECOMMENDATIONS TO CURB ATTRITION (1/4)

<u>Age</u>

- Employees aged **36 years and above are more likely to stay*** & Employees aged **32 years and below are more likely to leave***

- *Recommend that the company invests in retention opportunities like Staggered Rewards scheme for Good Employees & considers Employment Bonds for younger employees when the company invests in them*
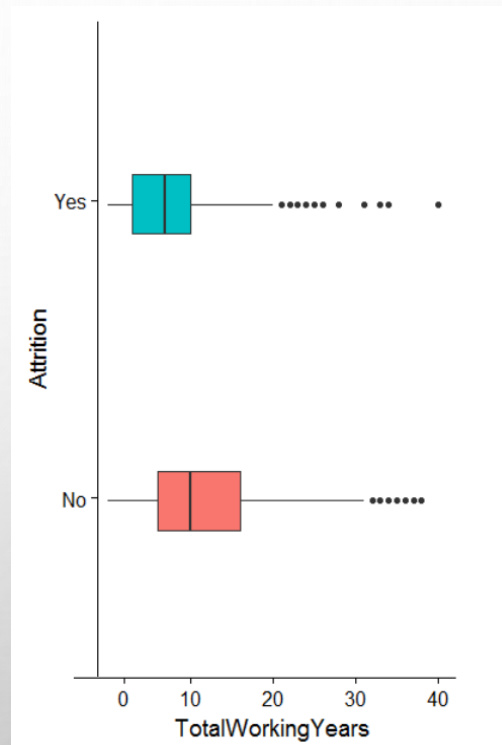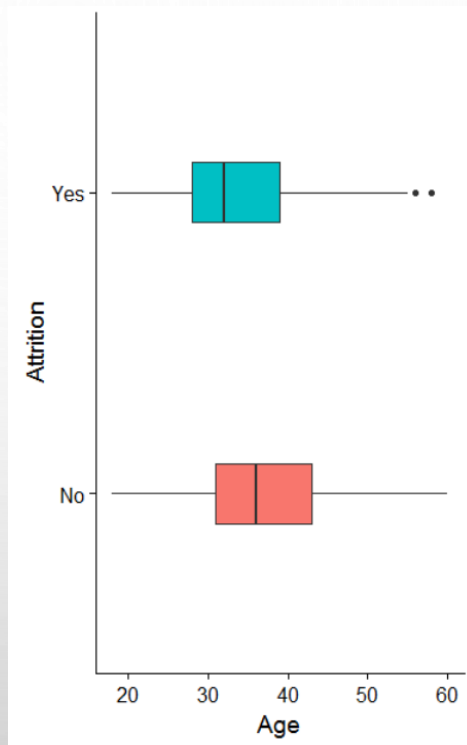
<u>Experience/ Total Working Years</u>

- Employees that have worked for a total of **10 years or more are more likely to stay*** & those who worked for a total of **7 years or less are more likely to leave***

- *Recommend that the company respects & rewards the work experience and harvests the employees skills well to encourage employees to increase their experience in the current company. Milestone service years may be rewarded*

*Coefficients of the variables Age and TotalWorkingYears are significant. Among attritions, median age = 32 and median exp. = 7. Among non-attritions, median age = 36 and median exp. = 10
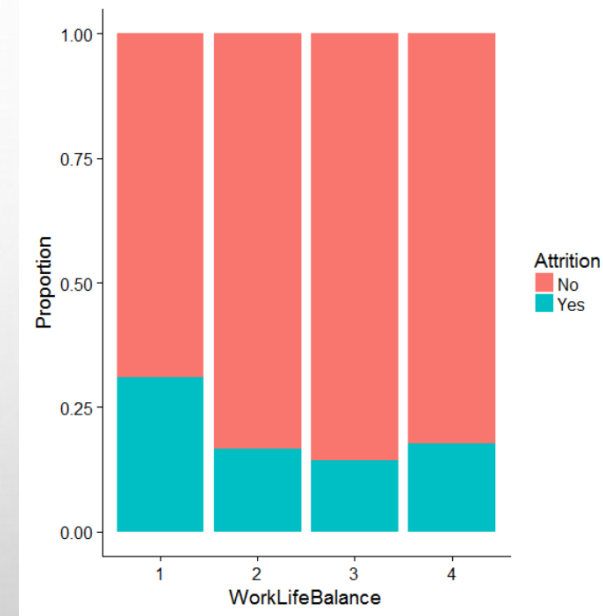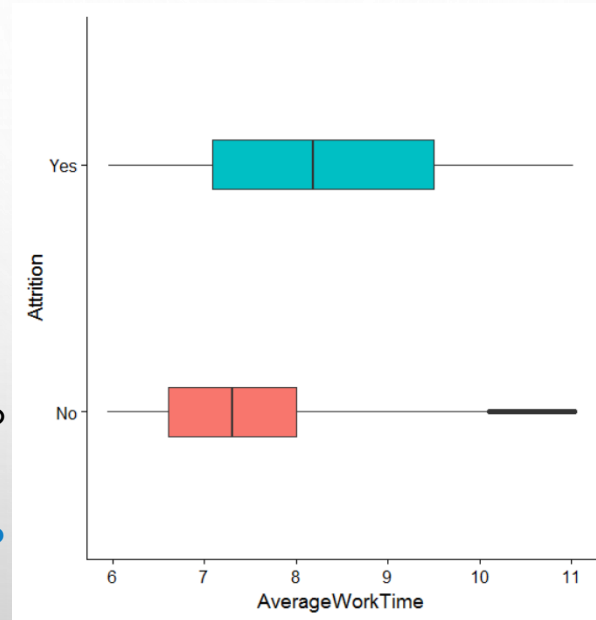
Average Work Hours

- Employees that, on **average work for 7.3 hours or less, are more likely to stay*** & those that worked **8.2 hours or more, are more likely to leave***

Work Life Balance

- Employees that rated their **work life balance as good, better or best, are more likely to stay**** & those that rated their work life balance as bad, are more likely to leave****

- *Recommend encouraging 8 hour work days to increase productivity & force employees to use their vacation time. Work should be uniform distributed by the manager to ensure no individual is overloaded*

*Coefficients of the variables AverageWorkTime and WorkLIfeBalance are significant. Average work hours data is based on means/medians etc. Employees were asked to report their level of work life balance in a survey.

RECOMMENDATIONS TO CURB ATTRITION (3/4)
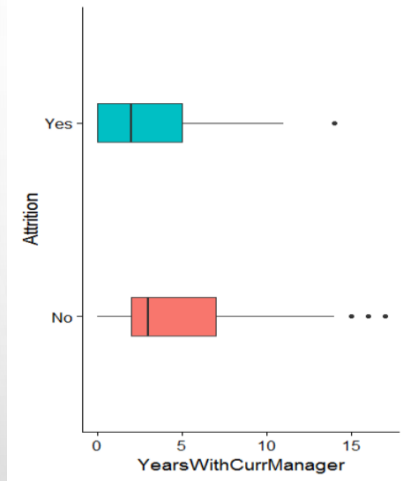
Marital Status

- Employees with **Single Marital Status are more likely to leave***

Years with Current Manager

- Employees that **spent 3 years or more under the same manager are more likely to stay*** & those that spent less than **2 years are more likely to leave***

- *Recommend that managers be trained & evaluated based on their team members individual growth and satisfaction.*

Frequent Business Travel

- **Employees that travelled more frequently are more likely to leave***

- *Recommend that managers ensure employees travel does not impact him/her negatively and try to align with the employees travel requirements*

*Coefficients of the variables TrainingTimesLastYear and YearsWithCurrManager are significant. Rest of the data is based on means/medians etc.

RECOMMENDATIONS TO CURB ATTRITION (3/4)

Job Satisfaction

- Employees that have **medium, high or very high levels of job satisfaction, are more likely to stay*** & those that have low levels of job satisfaction, are more likely to leave*

Environment Satisfaction

- Employees that have **medium, high or very high levels of environment satisfaction, are more likely to stay*** & those with low levels of environment satisfaction are more likely to leave*
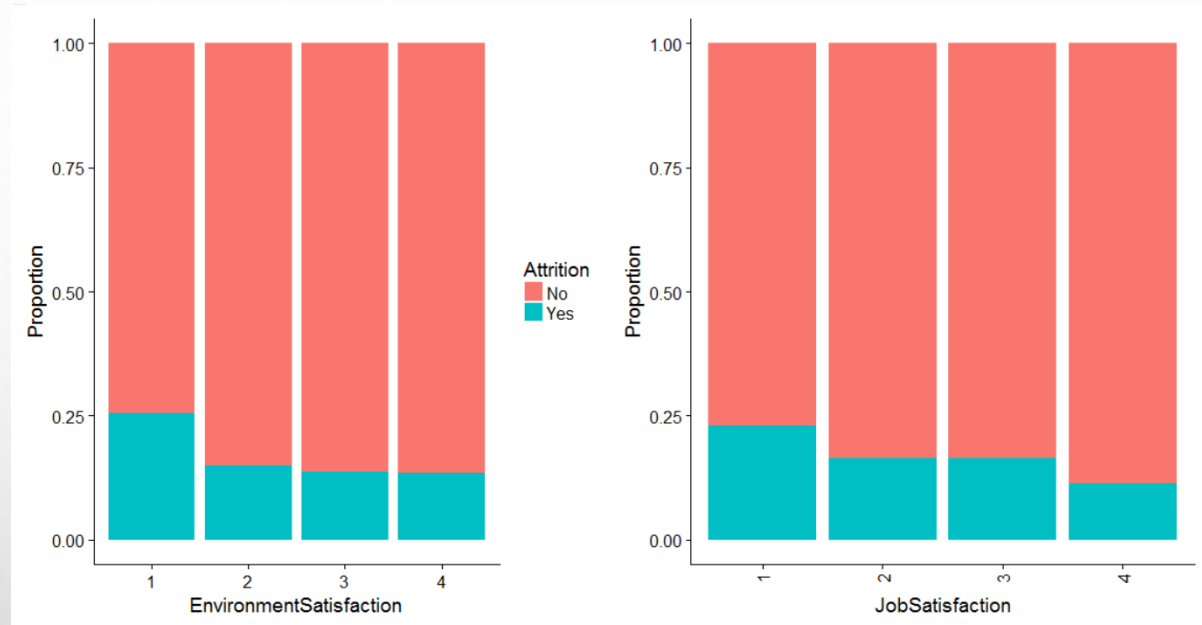
- *Recommend that the company invests in retention opportunities like Staggered Rewards scheme for Good Employees & considers enforcing Employment Bonds for younger employees when the company invests in them.*

*Coefficients of the variables JobSatisfaction and EnvironmentSatisfaction are significant. Employees were asked to report their job satisfaction and work environment satisfaction levels in a survey.

CONCLUSION

FACTORS INFLUENCING ATTRITION

- A Predictive model was developed to predict Employee Attrition
- 14 driver variables were identified by the model which directly had significant impact on the attrition of employees in XYZ company
- These driver variables identified by the Model were same as those variables identified during initial Data Analysis (seen in the bar charts & plots)
- A set of recommendations have been suggested to curb attrition based on these driver variables impacting Attrition
- Driver Variables include:
  - Age,
  - Number of Companies Worked,
  - Years Since Last Promotion and Years With Current Manager,
  - Average Work Time Spent in a Day,
  - Total Working Years of Experience,
  - Environment Satisfaction, Job Satisfaction, Work Life Balance,
  - Frequent Business Travel,
  - MaritalStatus – Single
  - JobRole as Manufacturing.Director

METRICS of MODEL

- Accuracy: 0.751938
- Sensitivity: 0.77 03349
- Specificity: 0.7483811
- Optimum Cut off Probability: 0.19
- KS Statistic: 0.519
- Gain: 75% in 3rd Decile
- Lift: 2.5