HR Analytics

HR Dataset from ‘XYZ’

Group Name:

1. Manish Jha
2. Namrata Khatri
3. Prashant Agrawal
4. Rahul Shukla

Objectives & Goals of HR Dataset Analysis

Business Objectives:

Model the probability of attrition in the company. Identify key variables that are to be addressed.

Strategy: The results 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.

Overall Structure of the presentation:

- **Problem Statement** –Identify the driving factors behind attrition of employees i.e. the variables which are strong indicators of attrition.
- **Results of Modelling**– Explain results from model in business terms.
- **Visualization**– Support the data analysis using visualization charts.

Guidelines & Assumptions

Guidelines:

- Dataset File name need not be changed.
- All code to be written in R

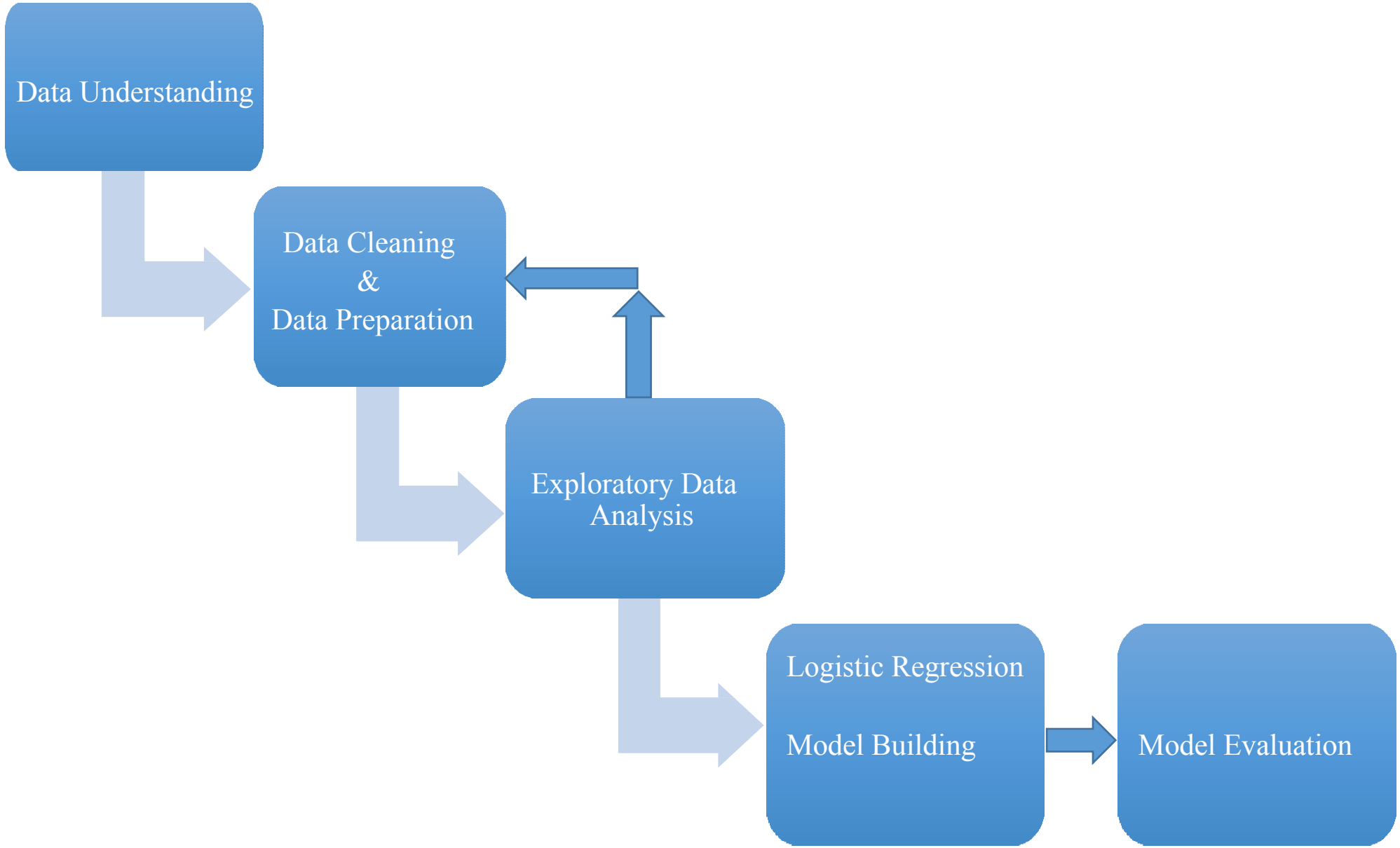
Data Management Framework & Technology:

- CRISP Data Management Framework
- R & R Visualizations.

Assumptions:

- **Baseline Code–** The code used for analysis is to be baselined and it can be used in future if follow-up questions are there from Chief Data Scientist.
- **Visualization-** More charts/visualization options can be made available for exploration subject to level of details asked during presentation.
- **Data** - All blank values are considered as NA while importing. NA values are subsequently ignored on case to case basis.
- **R Code-** Chief Data Scientist can ask to demonstrate additional charts and show-case the model using the R Code.

Approach – High Level



Data Understanding & Preparation

Data Understanding:

1. Merge the Datasets provided and create a master data frame ‘hrdb’
2. In_time & Out_time files to be used to create derived metrics for employees
3. Total no of employee observations in the dataset: 4410
4. Duplicate Ids are not there

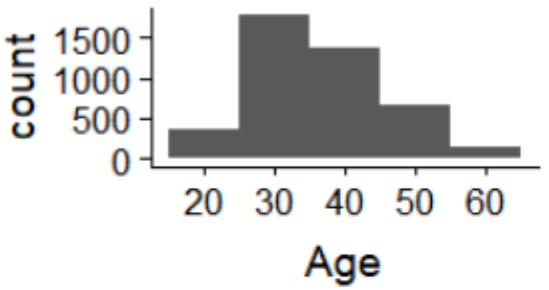
Data Cleaning & Preparation:

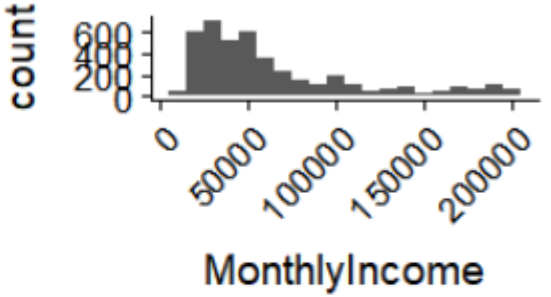
1. Remove variables where unique value is 1 (EmployeeCount, Over18 & StandardHours).
2. Convert dates from character to date format.
3. Remove holidays from the dataset
4. Derive new metrics for employee leaves & no of working hours based on in/out time
5. Handle outliers
6. Convert categorical columns into factors
7. Factor variables with two levels – convert values to 0 and 1
8. Creating dummy variables for all factors with more than 3 levels
9. Feature standardization -- scaling

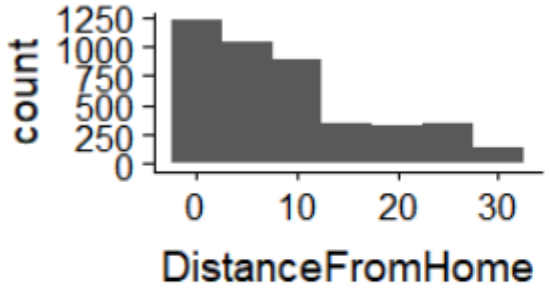
Exploratory Data Analysis -1

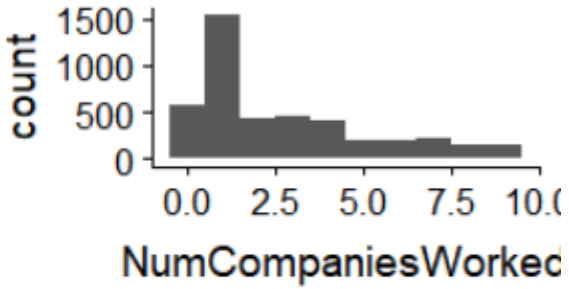
Key Insights

1. Maximum employees are in 30-40 age bracket.
2. MonthlyIncome has Outliers
3. Average Monthly Income of maximum employees is less than 50000.
4. Most of the employees stay in 10 Km vicinity of office.
5. NumCompaniesWorked has few Outliers
6. Most of the employees have worked in 1 company only.





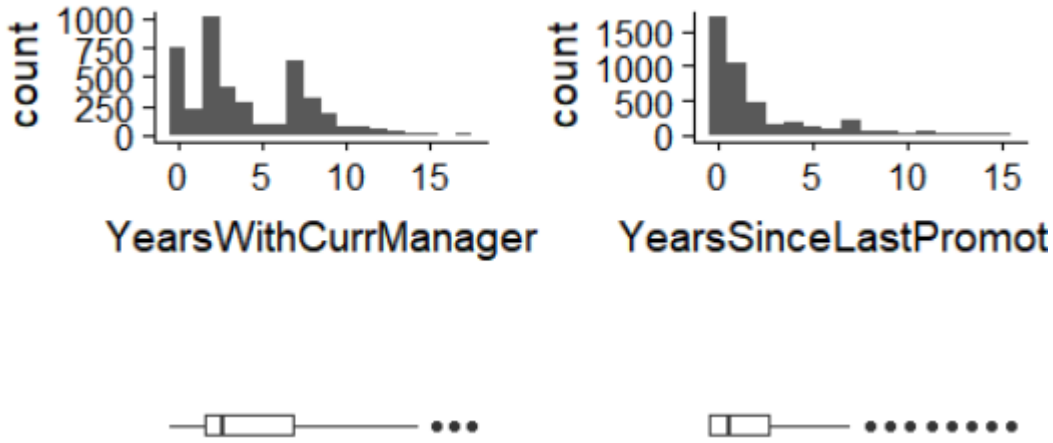


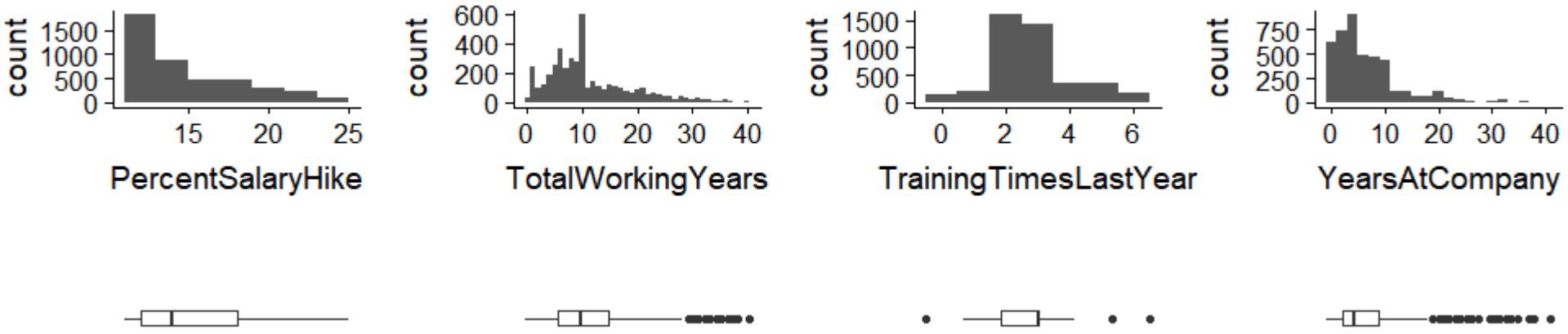


Exploratory Data Analysis -2

Key Insights

1. YearsWithCurrManager has Outliers
2. YearsSinceLastPromotion is skewed
3. TotalWorkingYears has Outliers
4. YearsAtCompany Has Outliers



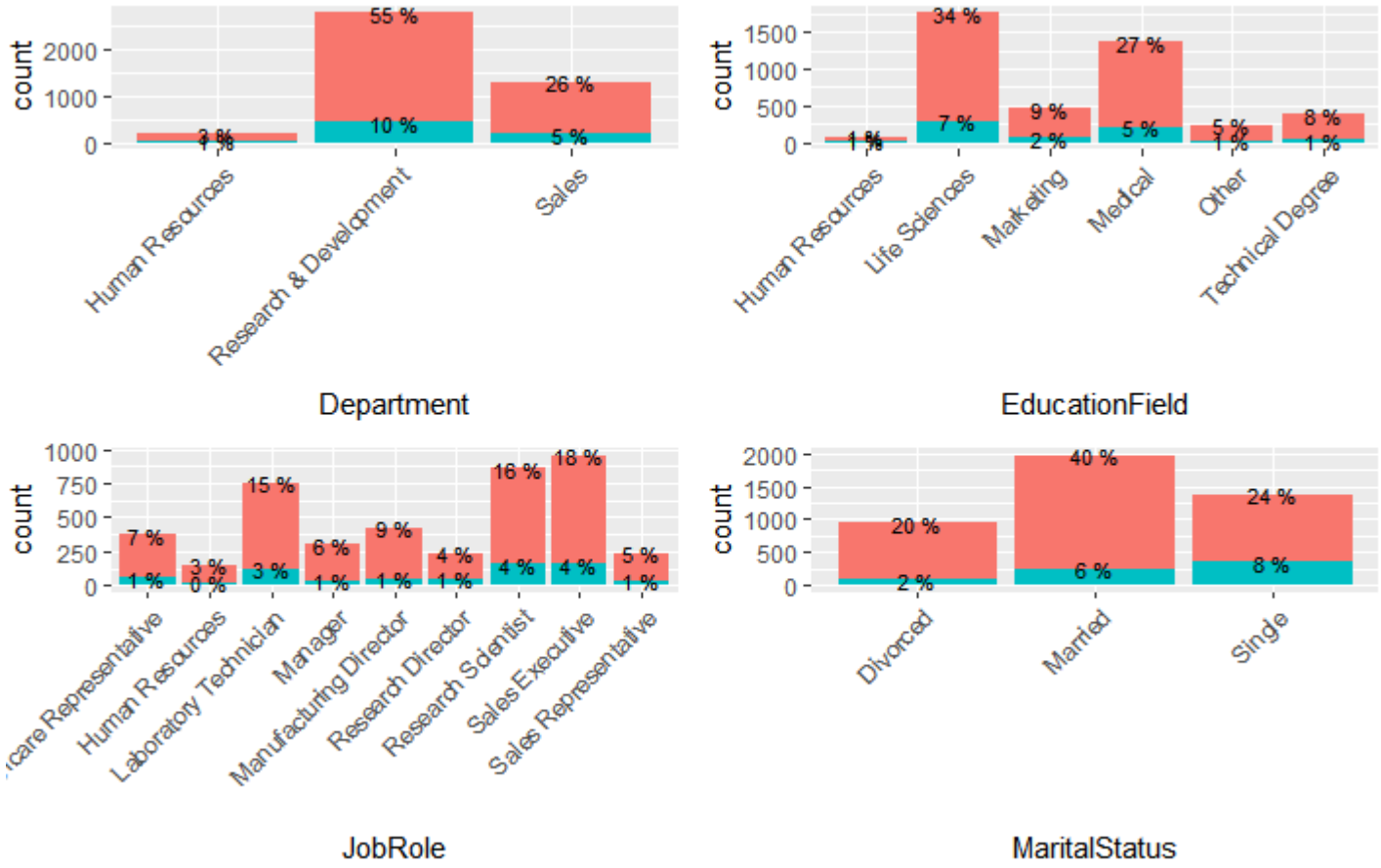


Exploratory Data Analysis -3

Key Insights on Attrition

Highest Attrition in:

1. Research and Development Department(10% Attrition)
2. Life Sciences & Medical Education fields (7% and 5%)
3. Married and Single people (6% and 8%)
4. Jobs of Sales Executives, Research Scientist and Lab Technicians (4%, 4% and 3%)

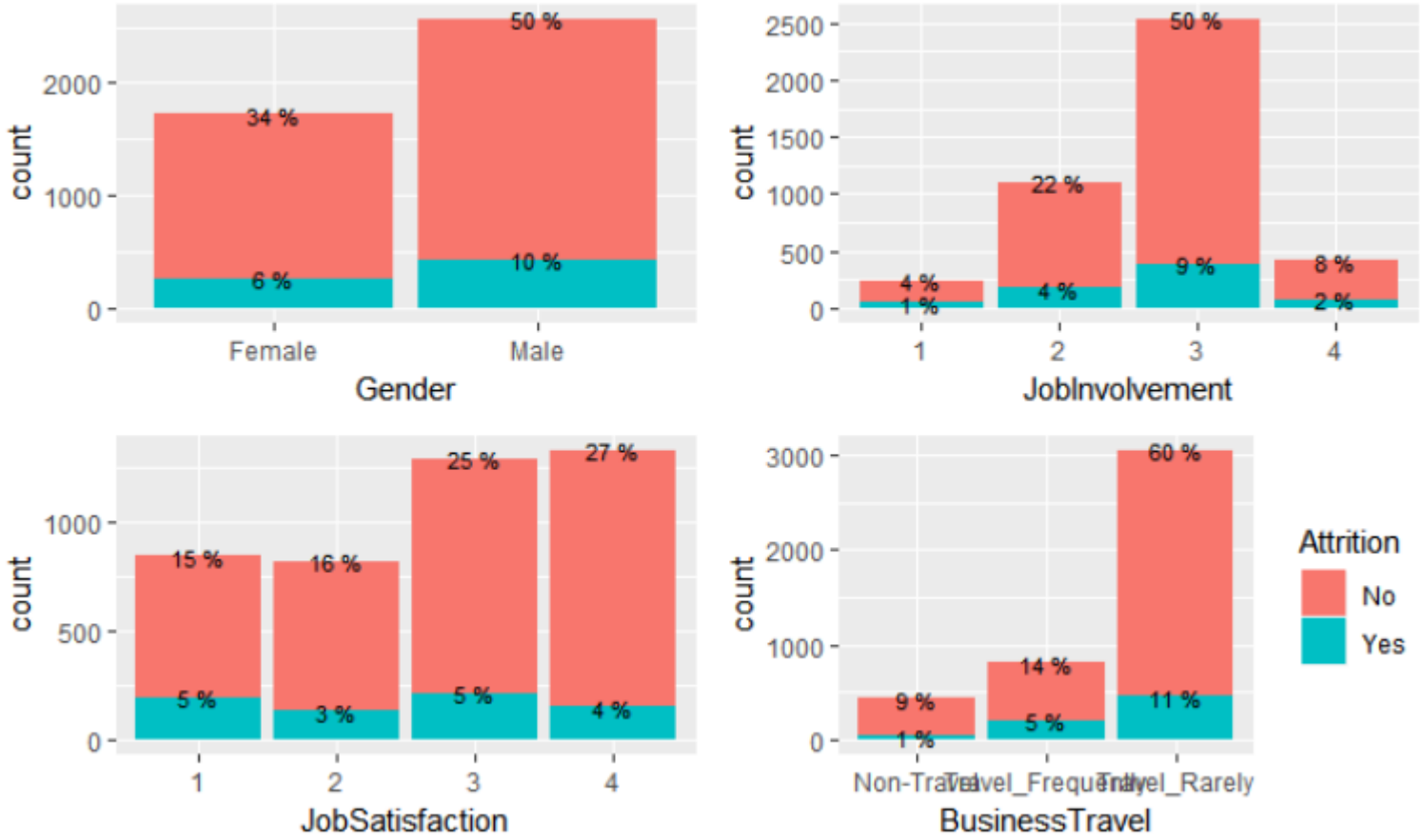


Exploratory Data Analysis -4

Key Insights on Attrition

Highest Attrition in:

1. Gender- Male (10%) and Female(6%)
2. Employees who travel rarely (11%)
3. Employees with high Job Satisfaction levels (Low-5% & High-5%)
4. Employees with Job Involvement Level-High-9%

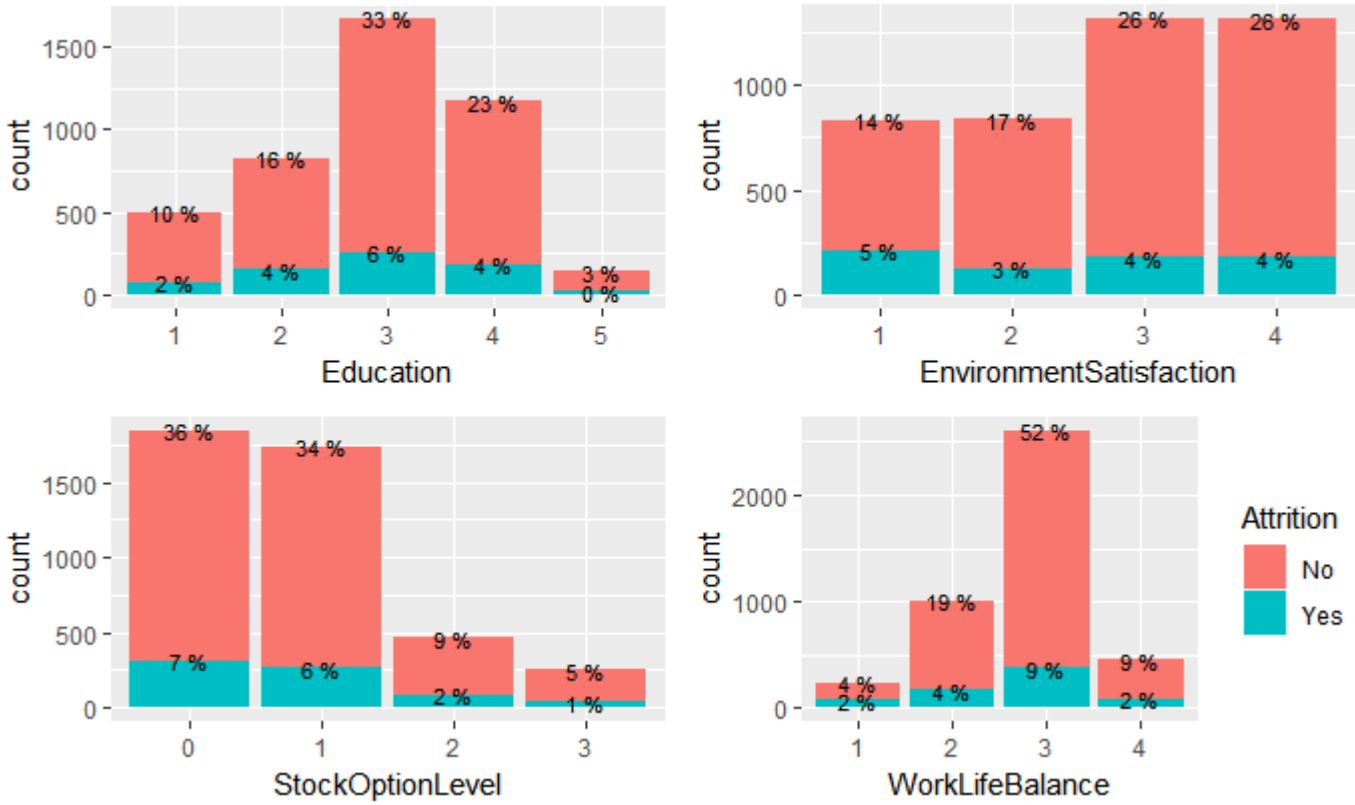


Exploratory Data Analysis -5

Key Insights on Attrition

Highest Attrition is observed in:

1. Employees with Bachelors degree -6%, Masters- 4%
2. Both non satisfied and satisfied employees in terms of Environment Satisfaction are leaving (Low-5%, High/Very High-4%)
3. Employees who are Better off in Work Life Balance are also leaving (Better – 9%)



Fitness Test of Observed Data

Chi-Square Goodness of Fit Test

- Perform Chi-Square test which tests the association of variables in two-way tables where the assumed model of independence is evaluated against the observed data.
- Calculate X-Squared values for combinations of two variables at a time
- Based on p-value reject/accept the hypothesis whether two variables are independent or not.

Results of Chi-Square Test:

- EducationField : 6 levels -- related with Education
- JobRole : 9 levels -- related with StockOptionLevel & Education
- JobInvolvement : 4 levels -- related with JobRole
- JobLevel: 5 levels -- related with JobRole
- Education: 5 levels - related with EducationField
- StockOptionLevel:4 levels -- related with JobRole

Model Creation

- Employee Attrition is the binomial variable that is to be predicted
- hrdb.final data-frame has 4300 observations with 57 variables
- set.seed of 100
- Split the final data-frame ‘hrdb.final’ into **Train** and **Test** datasets. SplitRatio used 0.7
- Create Logistic Regression model ‘model_1’ using **glm()** function in R which is used to fit generalized linear models (GLMs). AIC - 2121
- Apply stepAIC in both directions (forward and backward) to reduce the insignificant variables in model_1. stepAIC selects the model based on Akaike Information Criteria, not p-values. The goal is to find the model with the smallest AIC by removing or adding variables in your scope. AIC-2094, No of Variables – 36.
- Train the model - Iteratively remove variables with the objective of finding the most significant variables which don’t have multi-collinearity. Use VIF (Variance Inflation Factor) and p-value collectively to take decision. p-value is expected to be < 0.05 and VIF’s acceptable maximum range is 3-4.
- Final Model:
  - No of Iterations – 22 (based on low p-value and high VIF)
  - Actual model resulted in 13 variables. 11 variables in the model and all are significant (Business point of view).
  - AIC- 2200

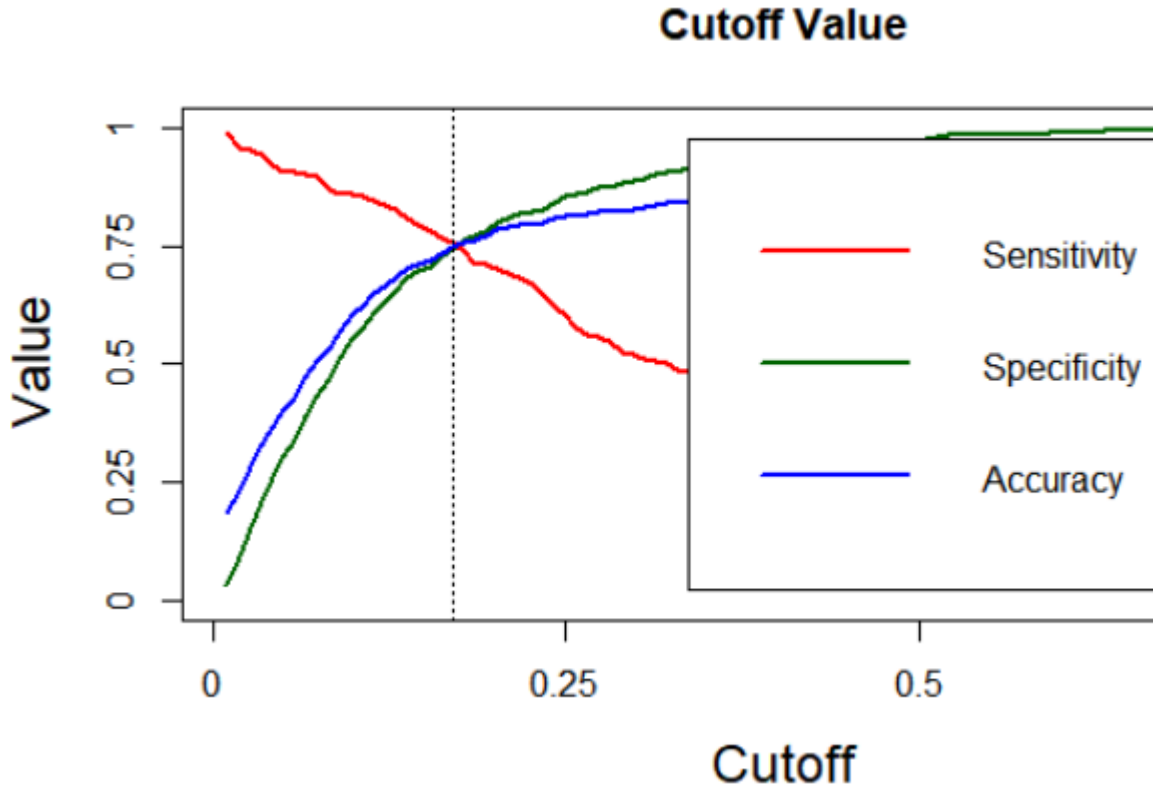
Model Output- Top Driver Variables

- HR Analytics recommends company 'XYZ' to take decisions on the following parameters/variables which are the key to control Attrition:

S.No	Variable	Variable Description (specific categories)	Positive/Negative Impact- Beta0
1	EnvironmentSatisfaction.xVery.High	Work Environment Satisfaction Level - Very High	-0.97603
2	JobRole.xManufacturing.Director	Job Role - Manufacturing Director	-0.82613
3	EnvironmentSatisfaction.xHigh	Work Environment Satisfaction Level - High	-0.73866
4	JobSatisfaction.xVery.High	Job Satisfaction Level- Very High	-0.69041
5	YearsWithCurrManager	Number of years under current manager	-0.66873
6	EnvironmentSatisfaction.xMedium	Work Environment Satisfaction Level -Medium	-0.62835
7	Age	Age of the employee	-0.5592
8	TrainingTimesLastYear	Number of times training was conducted for this employee last year	-0.19367
9	NumCompaniesWorked	Total number of companies the employee has worked for	0.25228
10	YearsSinceLastPromotion	Number of years since last promotion	0.53711
11	BusinessTravel.xTravel_Frequently	How frequently the employees travelled for business purposes in the last year- Travel Frequently	0.6285
12	AvgWorkHours	Average number of working Hours by an Employee	0.65604
13	MaritalStatus.xSingle	Marital status of the employee – Single (Unmarried employees)	0.8826

Model Evaluation & Summary

- Apply the model on the Test dataset and predict attrition for each observation by creating a new variable.
- Use probability of 17% as the cut-off to identify Attrition in the newly created variable.
- Create metrics to gauge the discriminative power of logistic regression model:
  - Build Confusion Matrix
  - Accuracy – 0.75
  - Sensitivity- 0.76
  - Specificity- 0.74
  - KS Statistic- 0.50



Model Evaluation & Summary- 2

- Create deciles and bucket employees based on the probabilities of Attrition.

Decile	Observations	Attrition	Cumulative Attrition	Gain	Lift	Cumulative Non Attrition	Gain Non Attrition	KS
1	129	74	74	35.41	3.54	55	5.09	30.32
2	129	43	117	55.98	2.80	141	13.04	42.94
3	129	32	149	71.29	2.38	238	22.02	49.28
4	129	20	169	80.86	2.02	347	32.10	48.76
5	129	10	179	85.65	1.71	466	43.11	42.54
6	129	6	185	88.52	1.48	589	54.49	34.03
7	129	4	189	90.43	1.29	714	66.05	24.38
8	129	7	196	93.78	1.17	836	77.34	16.44
9	129	3	199	95.22	1.06	962	88.99	6.22
10	129	10	209	100.00	1.00	1081	100.00	0.00

Model Evaluation & Summary- 3

- Model looks good as KS statistic > 40% is achieved in 2nd decile itself.
- Safest Model- Cut-Off of 0.1695 equalizes Accuracy, Sensitivity and Specificity.
- Gain of 80% by 4rth decile- Attrition of 80% can be addressed in top 4 buckets of Employees.
- Lift of 2 by 4rth decile. The model performs 2 times better than random model.
- KS Statistic of 49% in 3rd decile. Employees likely to leave the company are present in the top 3 deciles.