HR Analytics Case Study

SUBMISSION

BY:

1. Nikhil Srivastava (DDA1720280)
2. Nandakishore Kulkarni (DDA1720263)
3. Manish Sharma (DDA1720248)
4. Ranju Ramesh (DDA1720293)

Abstract

Business Objective:

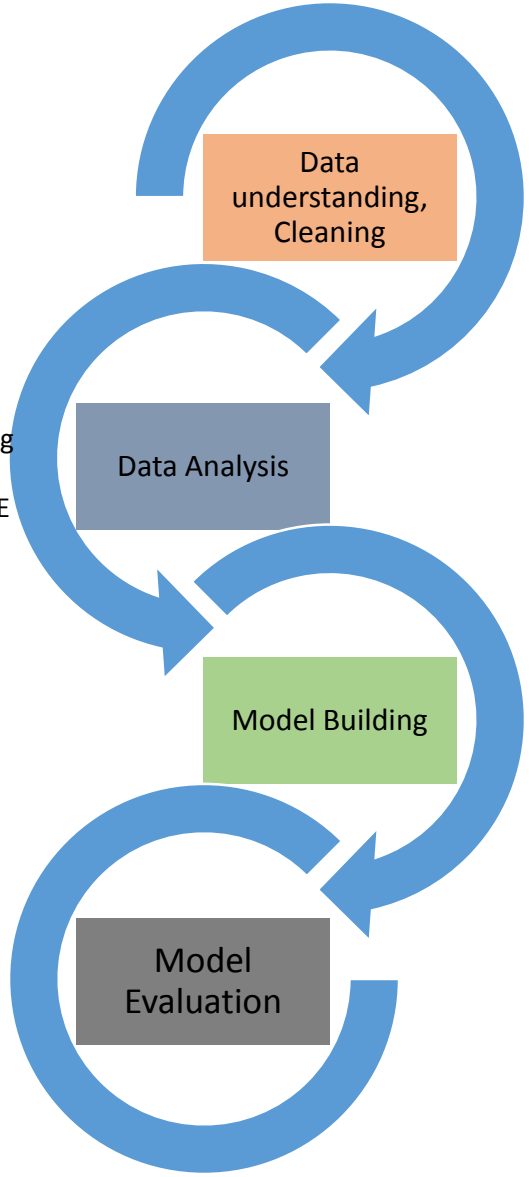
- Identify most important variables affecting Attrition
- What changes **XYZ** should make to their workplace, in order to get most of their employees to stay.

Data Source:

- Data set for 4410 employees which includes:
  - General details of every employee including his personal, education, and professional details
  - Attendance Data for 2015 for all employees
  - Manager survey & Employee survey feedback

Strategy: Predictive analysis using Logistic Regression to *identify key* variables affecting attrition

Problem solving methodology



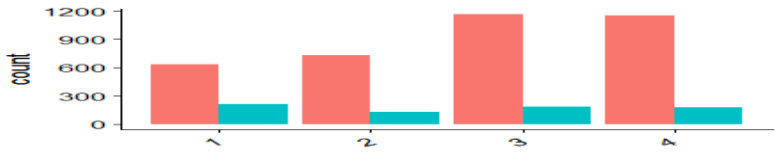
- AS-IS loading Loan in R
- Analysis of key columns of loan data.
- Remove unwanted columns having 0 and NA
- Check for Duplicate & null values
- Convert Date to standard format

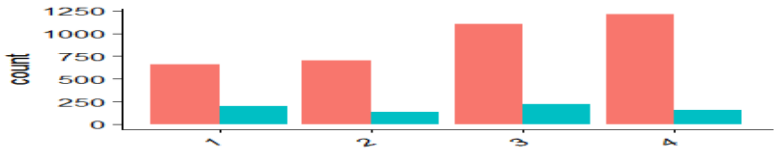
- Analysis of raw data using Univariate & Segmented analysis.
- Missing Value Imputation using WOE
- Binning of variables using WOE
- Dummy variable creation

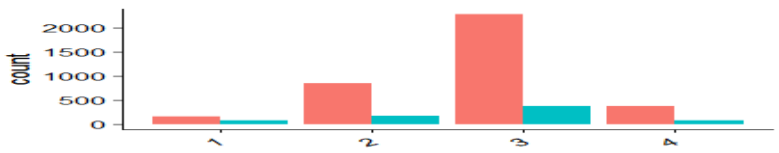
- Use StepAIC and glm to arrive at key factors influencing the attrition

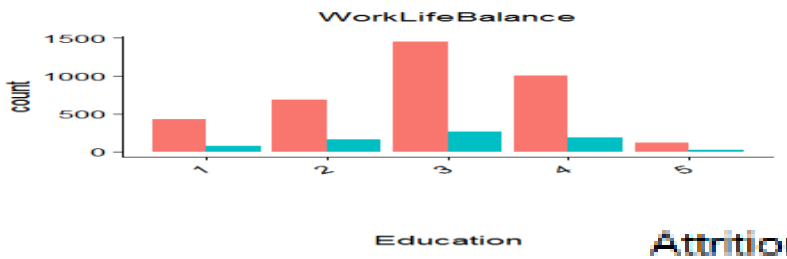
- Identify the model's
  - Accuracy
  - Sensitivity
  - Specificity
- Use Gain, Lift and KS Stats to predict best model

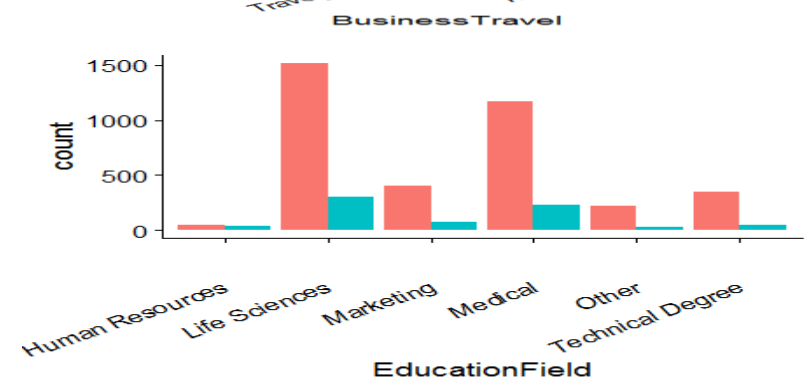
Univariate analysis on Key attributes

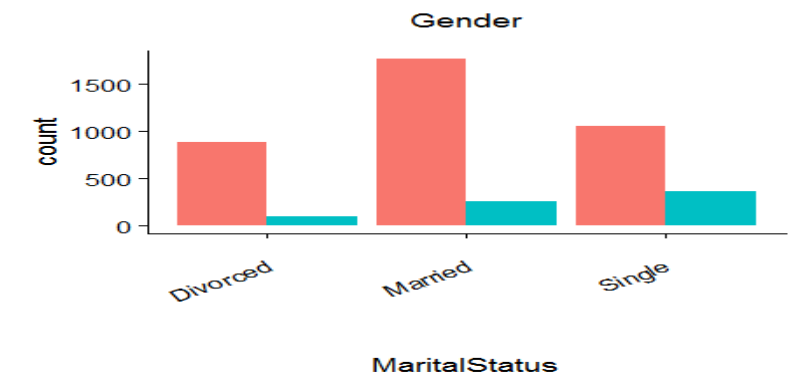


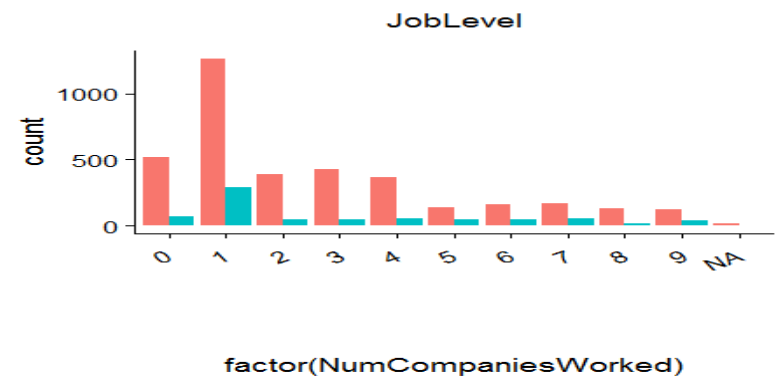




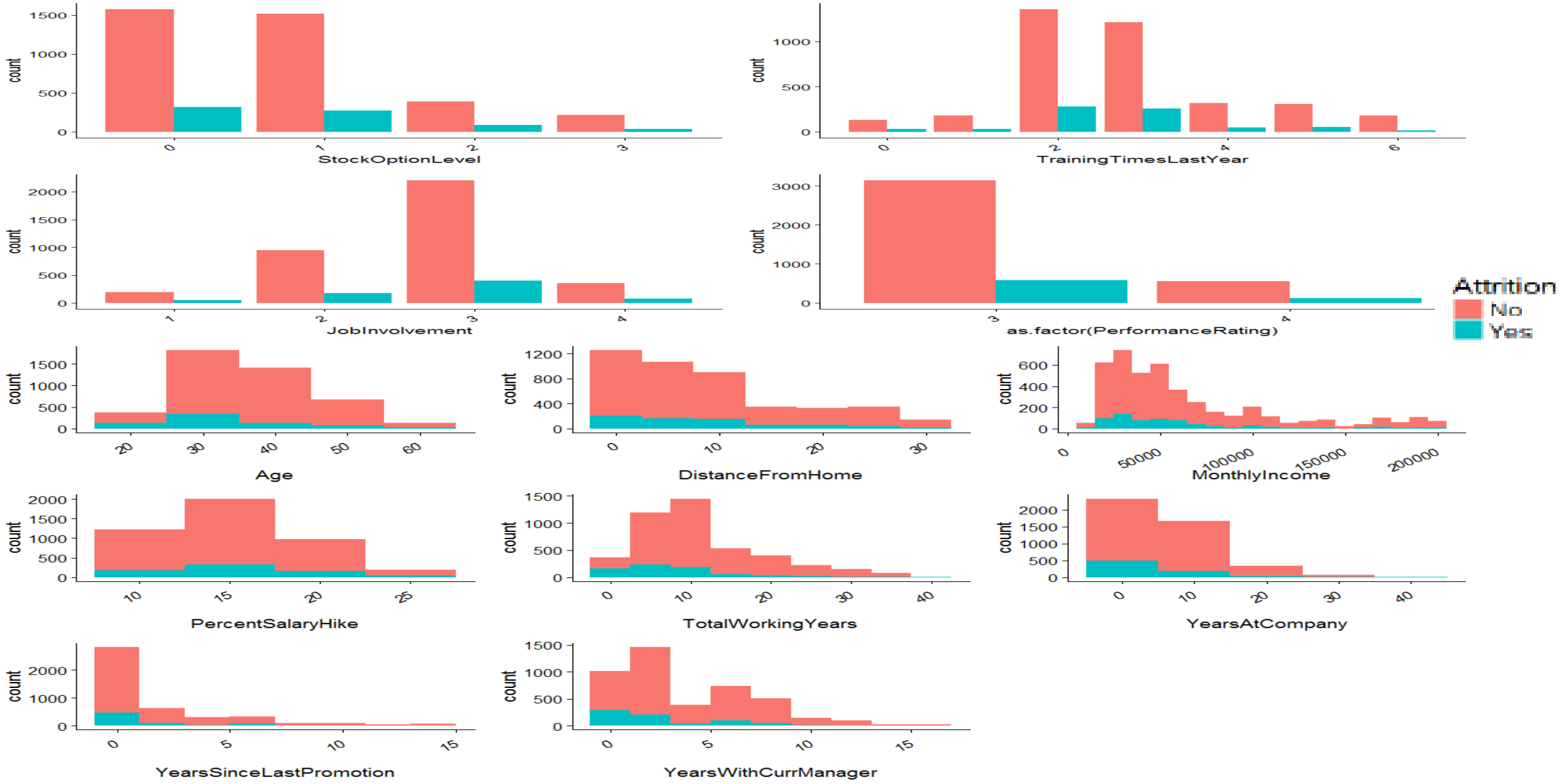




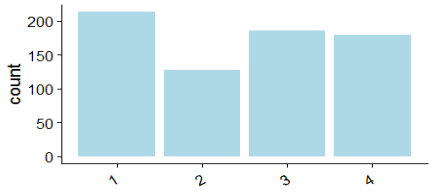


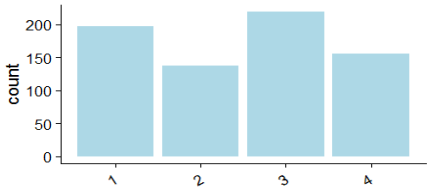


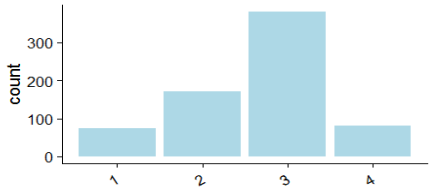
Univariate analysis on Key attributes (contd..)

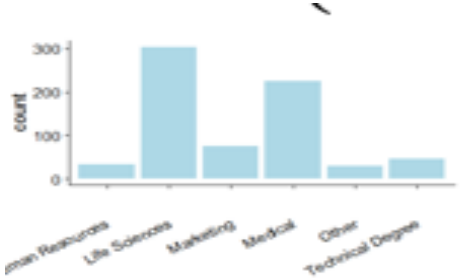


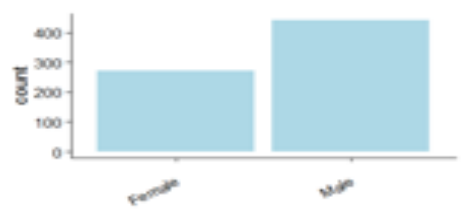
Segmented analysis on Key attributes (Attrition="Yes")

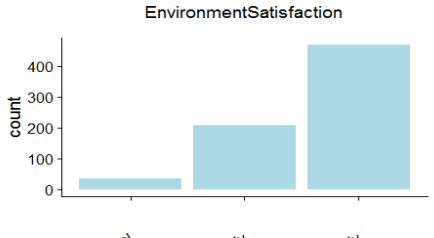


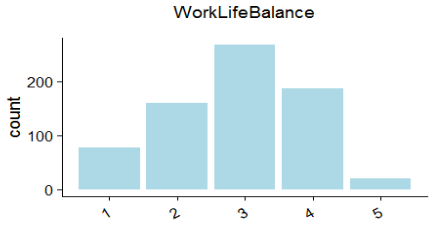


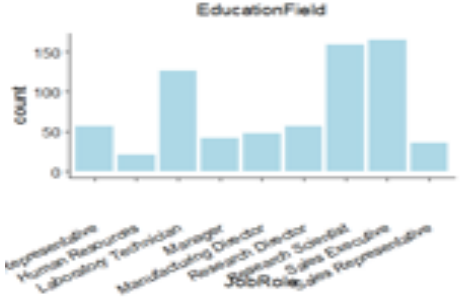


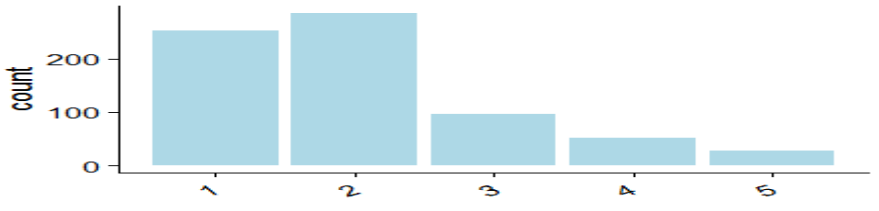


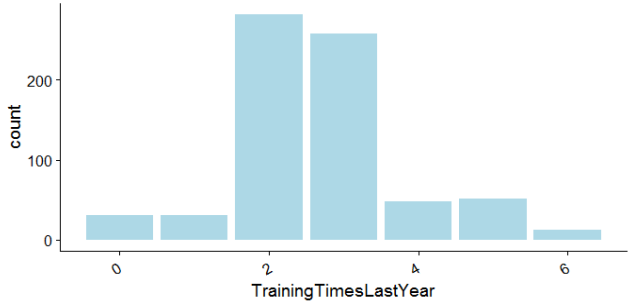


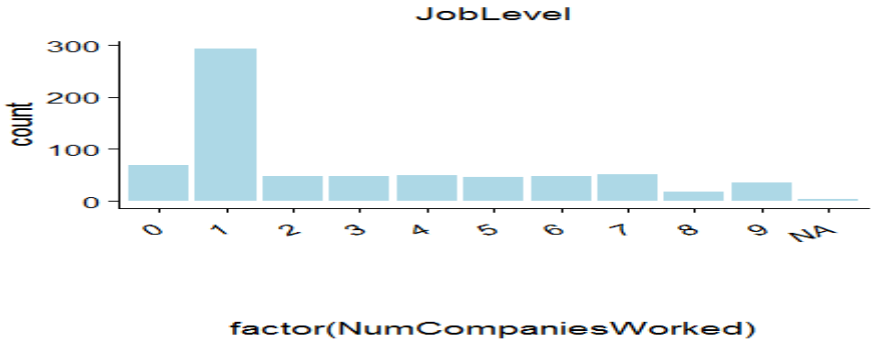


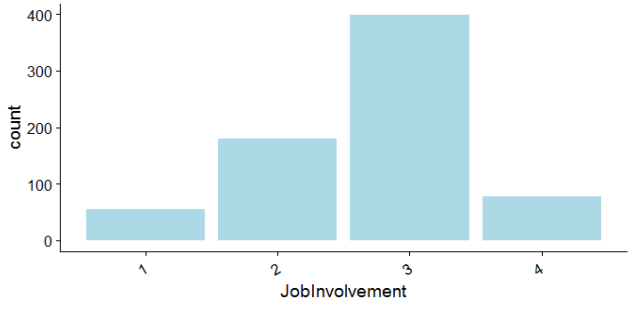


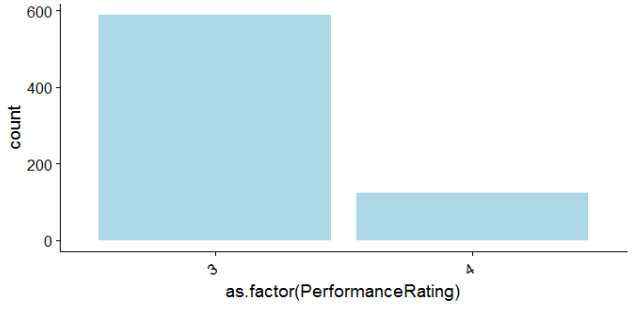




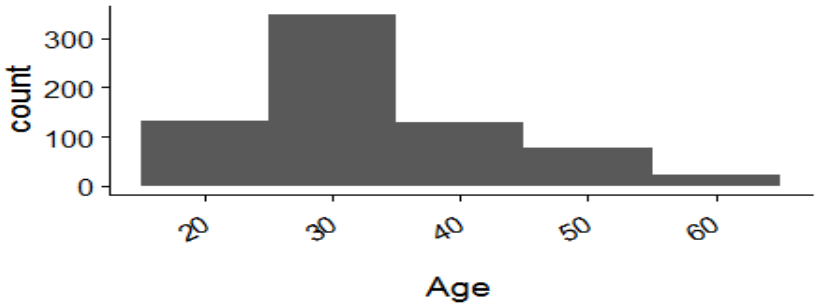


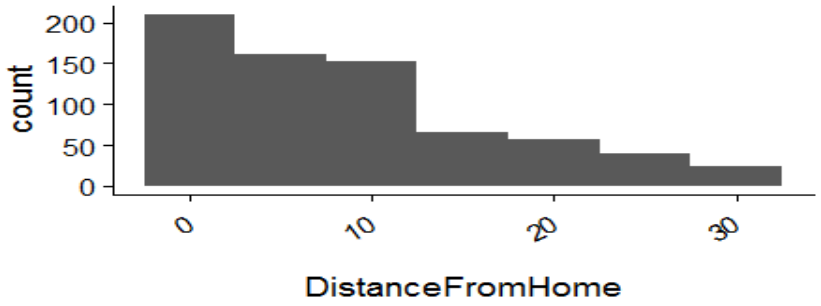


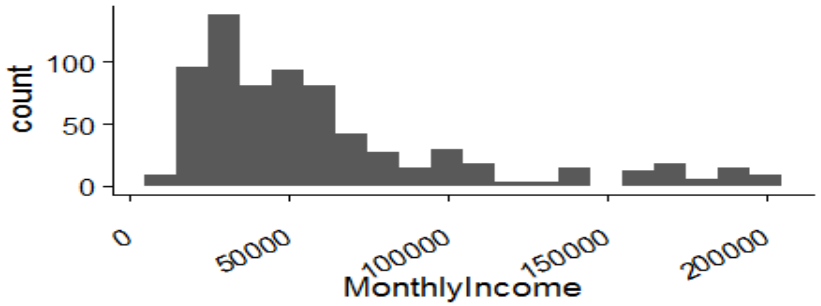


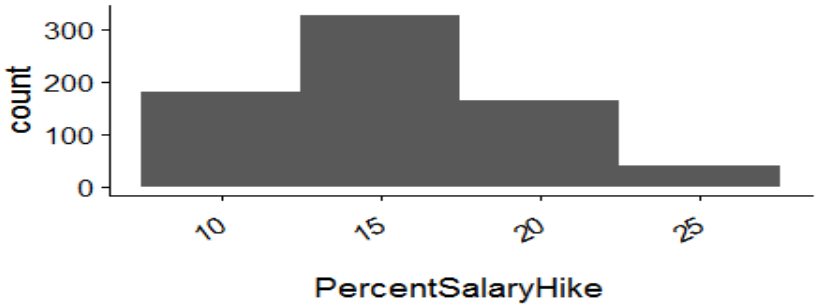


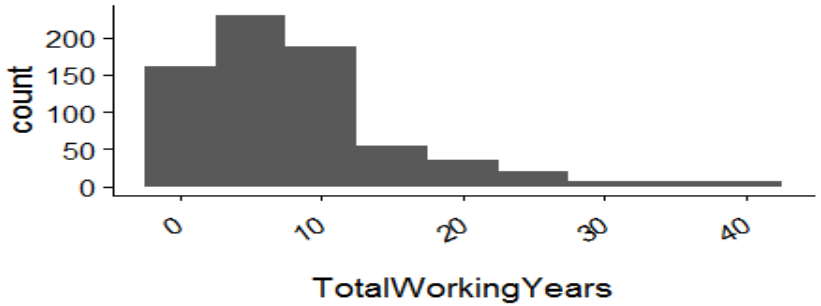
Segmented analysis on Key attributes
(Attrition="Yes")

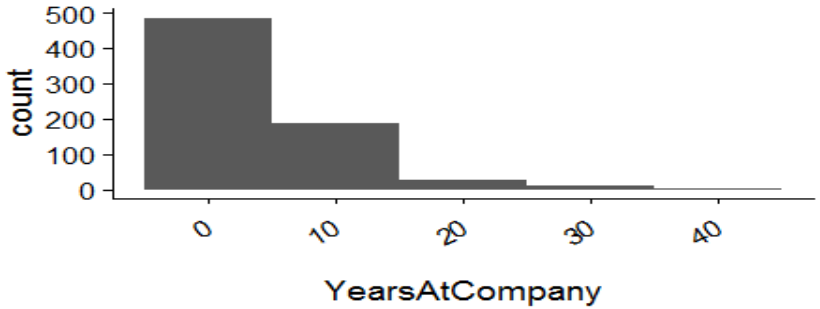


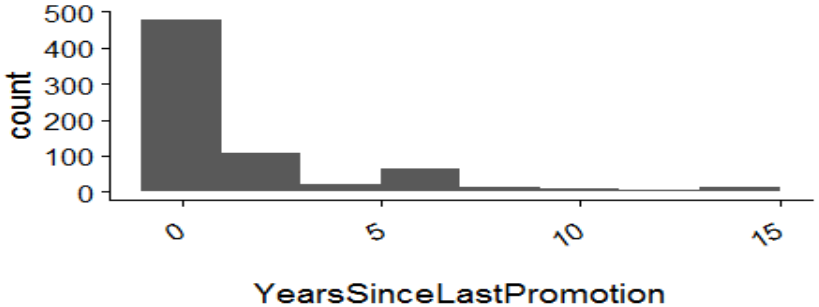


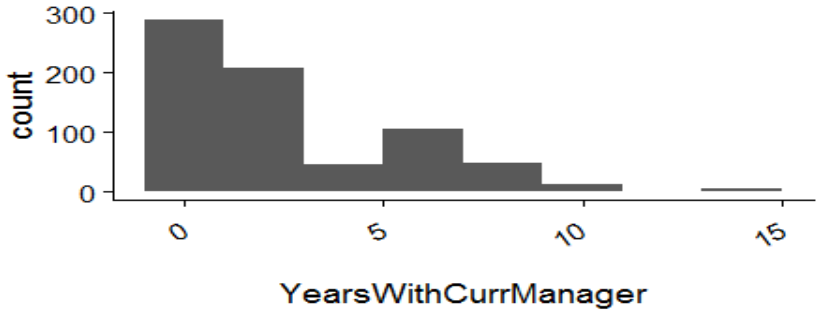












Summary: Univariate & Segmented Univariate analysis

Key observations: No variable is single handedly causing the attrition in XYZ company.
Below are a few variables which are having significant impact

Gender & Marital Status	• **Males have higher attrition rate than females in the company** • Attrition rate is higher in Employees who are **"Single".**
Environment Satisfaction	• Employee who gave "**low**" Environment Satisfaction has high attrition rate

Education Field	• Higher Attrition rate in employees belonging to below education field:
	  • Life Sciences
	  • Medical

Business Travel	• Employees who "Travel_Rarely" have high attrition rate than others
Department	• Department – "Research and Development" has higher attrition rate
Job Level	• lower job levels have higher attrition rate
Age	• Most attrition happening in age group of 20-35

Total Work Exp and No. of years in Company	• Attrition level is higher in employees who:
	  • Have a total work experience of less than 10 years
	  • Have been working in the company for 10 years or less

Data Manipulation

Data Manipuation

1. **Missing Value Imputation:**
   - In order to avoid bais, WOE analysis was used to replace NA values as shown in **Table 1.**
   - Variable “TotalWorkingYears” has only 0.2% values as NA and hence such removed
2. **Binning:**
   - Binning as shown in **Table 2** is done to convert all continuous variables to categorical variables
3. **Scaling:**
   - Variable “Monthly Income” was scaled
4. **Outlier treatment:**
   - Monthly Income: Outlier treatment done at 90% and 91%
5. **Dummy Variables:**
   - Dummy variables were created for all categorical and continuous variable.
   - Variable EmployeeID was not considered for model creation

Table 1: Missing Value Imputation

Variable	count of NA's	Method used
EnvironmentSatisfaction	25	Replace NA’s with WOE value
JobSatisfaction	20	Replace NA’s with WOE value
WorkLifeBalance	38	Replace NA’s with WOE value
NumCompaniesWorked	19	Replace NA’s with WOE value
TotalWorkingYears	9	NA values were removed

Table 2: Binning

Variables	Bins						
TotalWorkingYears	0-2	3-4	5-7	8-12	13-16	17-22	23-40
PercentSalaryHike	11-12	13-14	15-18	19-20	21-25		
Age	18-25	26-33	34-37	38-60			
DistanceFromHome	1-2	3-10	11-29				
MonthlyIncome	10090-23130	23140-68770	68770+				
YearsAtCompany	0-2	3-8	9-14	15-40			
YearsSinceLastPromotion	0	1-3	4-15				
YearsWithCurrManager	0	1-3	4-8	9-15			

Model Building

Problem Statement

- Use Logistic Regression technique to predict the classification of attrition rate

Approach

- Final dataset has 4401 observations across 80 variables.
- The data set was divided such that 70% of the data is used as a training data-set while 30% is used as test data set

Steps

- First model is created by applying glm function for all variables.
- StepAIC is performed to remove non-significant variables
- Variables with p-value>0.05 and high VIF were removed to arrive at final model.

Result

- Total of 19 models were created to arrive at final model
- Key Variables:
  - The final model has 23 variables which together impact the attrition rate

Overall Analysis: Factors impacting attrition of employees

- Using the model building approach, we could identify the below variables that were affecting the attrition rate of XYZ company:

Variable	Value(s) affecting Attrition
Environment Satisfaction	Level 1 signifying Low Satisfaction
Job Satisfaction	Level 1 signifying Low Satisfaction
Work Life Balance	Level 1 signifying Bad Work Life Balance
Business Travel	Employees who travel Frequently
Department	1. Research & Development 2. Sales
Job Role	Research Director
Marital Status	Single
Number of Companies Worked	Employees who have worked in more than 5 companies
Training Times Last Year	Employees who have attended 6 trainings in last year
Average Working Hours	Employees who spent more than the average of 8.5 hrs in office in the year
Total Work Experience	Employees who have more than 3 years of work experience
Age	Employee who are 34 years or more
Years Since Last Promotion	Employees who haven't been promoted in last 4 years or more
Years With Current Manager	Employees working with the same manager for more than 1 years

Model Evaluation

- Use of predict function to come up with predicted probabilities
- Probability range: 0.2% to 92%

Confusion Matrix on 50% Probability:

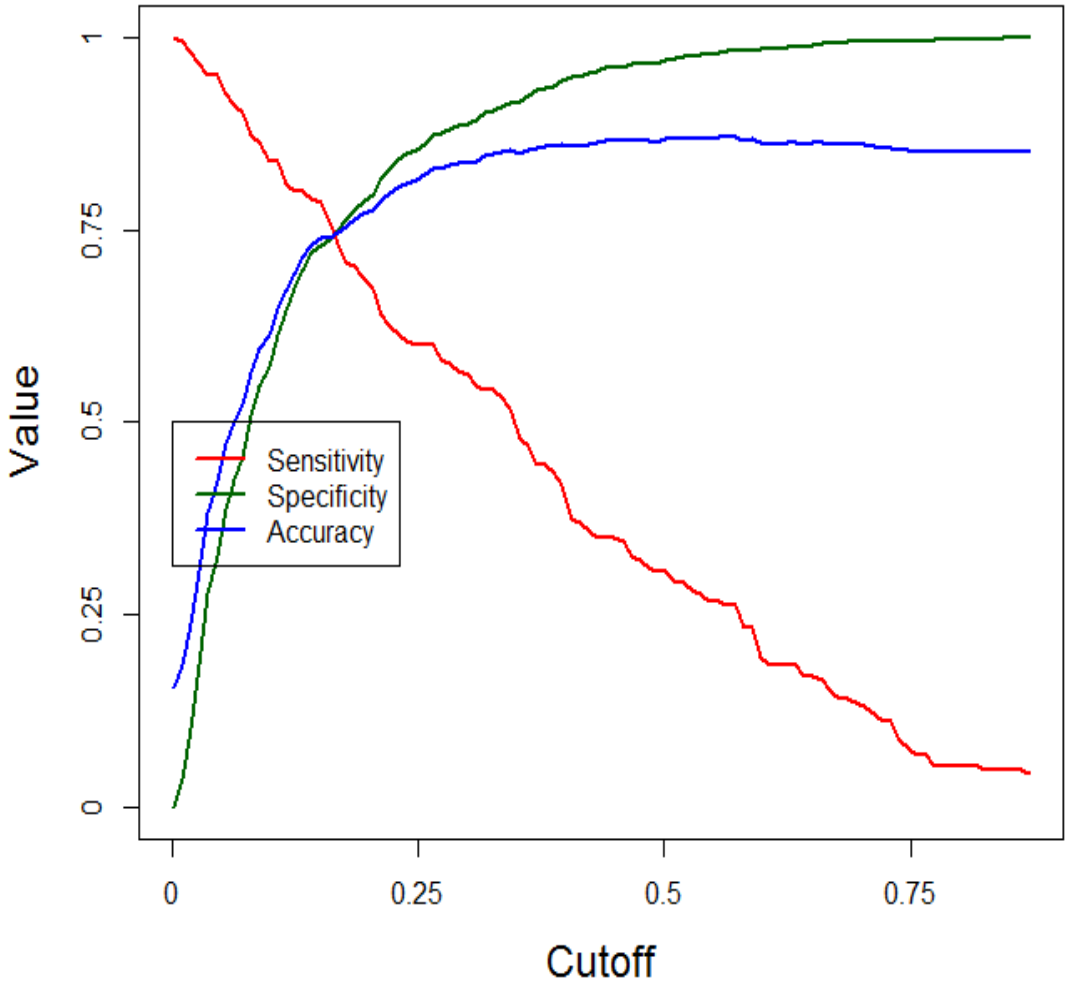
- Accuracy : 0.8675
- Sensitivity : 0.30732
- Specificity : 0.97043

Though Accuracy and Specificity is good, sensitivity is very low.

- To overcome low Sensitivity, user defined function created to identify cutoff value
- Optimal probability threshold for best prediction**: 0.1686**

Confusion Matrix at cutoff:

- Accuracy : 0.7456
- Sensitivity : 0.7317
- Specificity : 0.7482



Model Evaluation (contd.)

Model Evaluation using the below techniques:

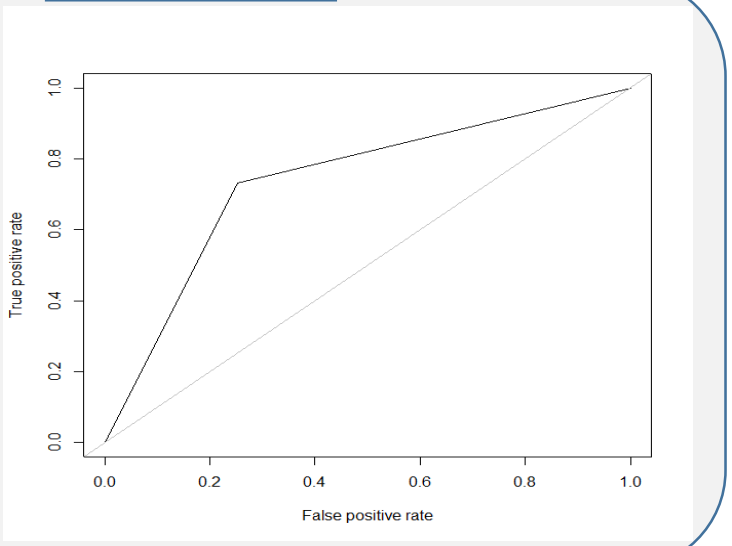
- KS-Statistic
- Lift
- ROC

KS-Stats

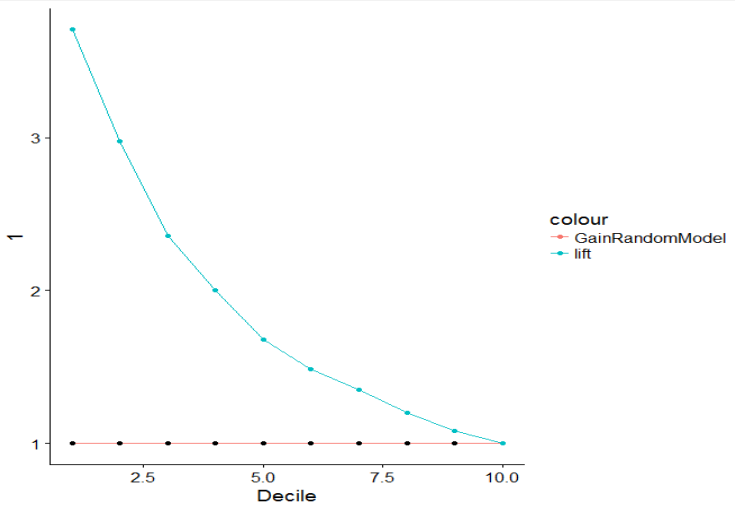
Decile	AttritionCount	CumulativeAttrition	GainPercent	NonAttritionCount	CumulativeNonAttrition	GainPercentNonAttrition	KS_stat	GainRandomModel	lift
1	76	76	37.07317073	56.1	56.1	5.02688172	32.04628901	10	3.707317073
2	46	122	59.51219512	86.1	142.2	12.74193548	46.77025964	20	2.975609756
3	23	145	70.73170732	109.1	251.3	22.51792115	48.21378617	30	2.357723577
4	19	164	80	113.1	364.4	32.65232975	47.34767025	40	2
5	8	172	83.90243902	124.1	488.5	43.77240143	40.13003759	50	1.67804878
6	11	183	89.26829268	121.1	609.6	54.62365591	34.64463677	60	1.487804878
7	11	194	94.63414634	121.1	730.7	65.47491039	29.15923595	70	1.351916376
8	3	197	96.09756098	129.1	859.8	77.04301075	19.05455022	80	1.201219512
9	3	200	97.56097561	129.1	988.9	88.61111111	8.949864499	90	1.08401084
10	5	205	100	127.1	1116	100	0	100	1

Model Evaluation (contd.)

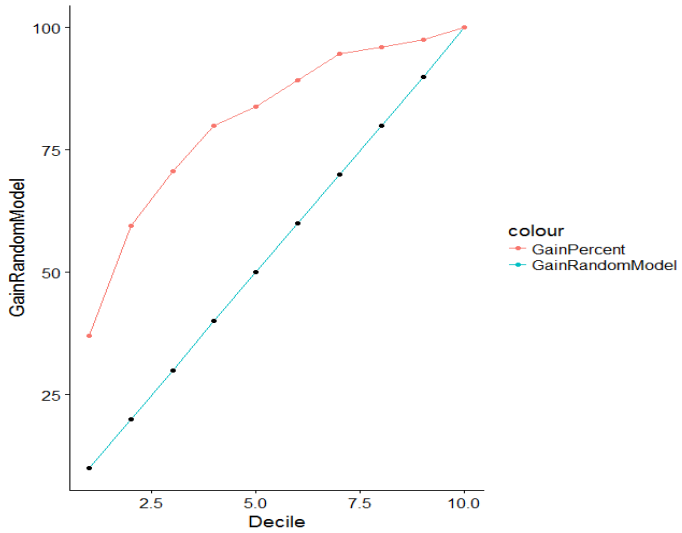
ROC Curve



Lift Chart



Gain Chart



Inference:

- KS-Statistic table show max value of 48% at 3rd Decile.
- This suggest that our model would help in identifying 71% of employees by targeting 30% of workforce.
- Lift chart and ROC shows that our model is able to distinguish between which employees would leave and which would not.


Thank You