HR Analytics Case Study

Presented By:

- Avishek Sengupta
- Harkirat Dhillon
- Shubhra Karmahe
- Taranveer Singh Arora

Abstract

Business Objective:

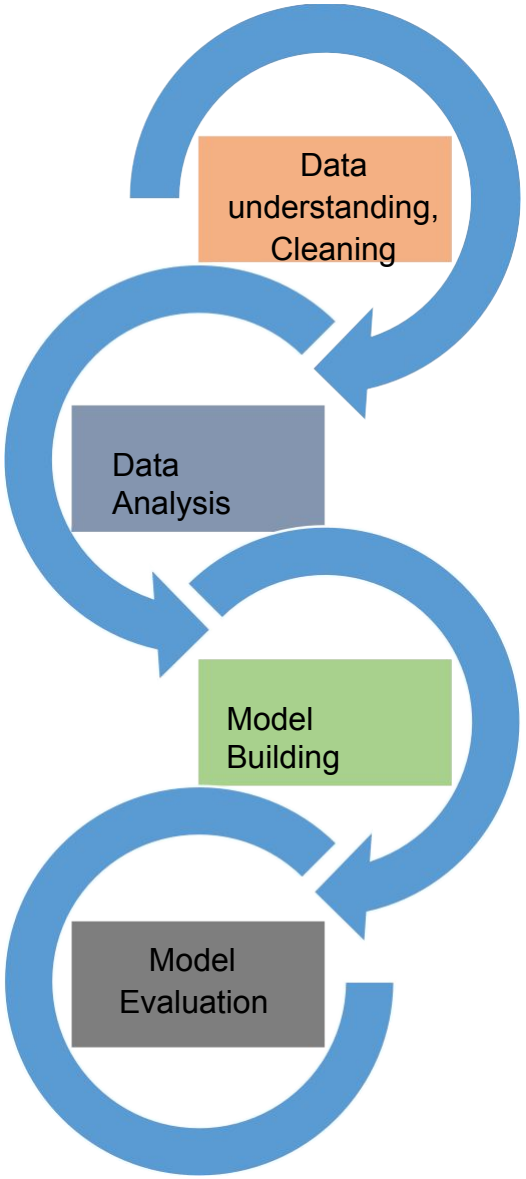
- Identify most important variables affecting Attrition
- What changes **XYZ** should make to their workplace, in order to get most of their employees to stay.

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



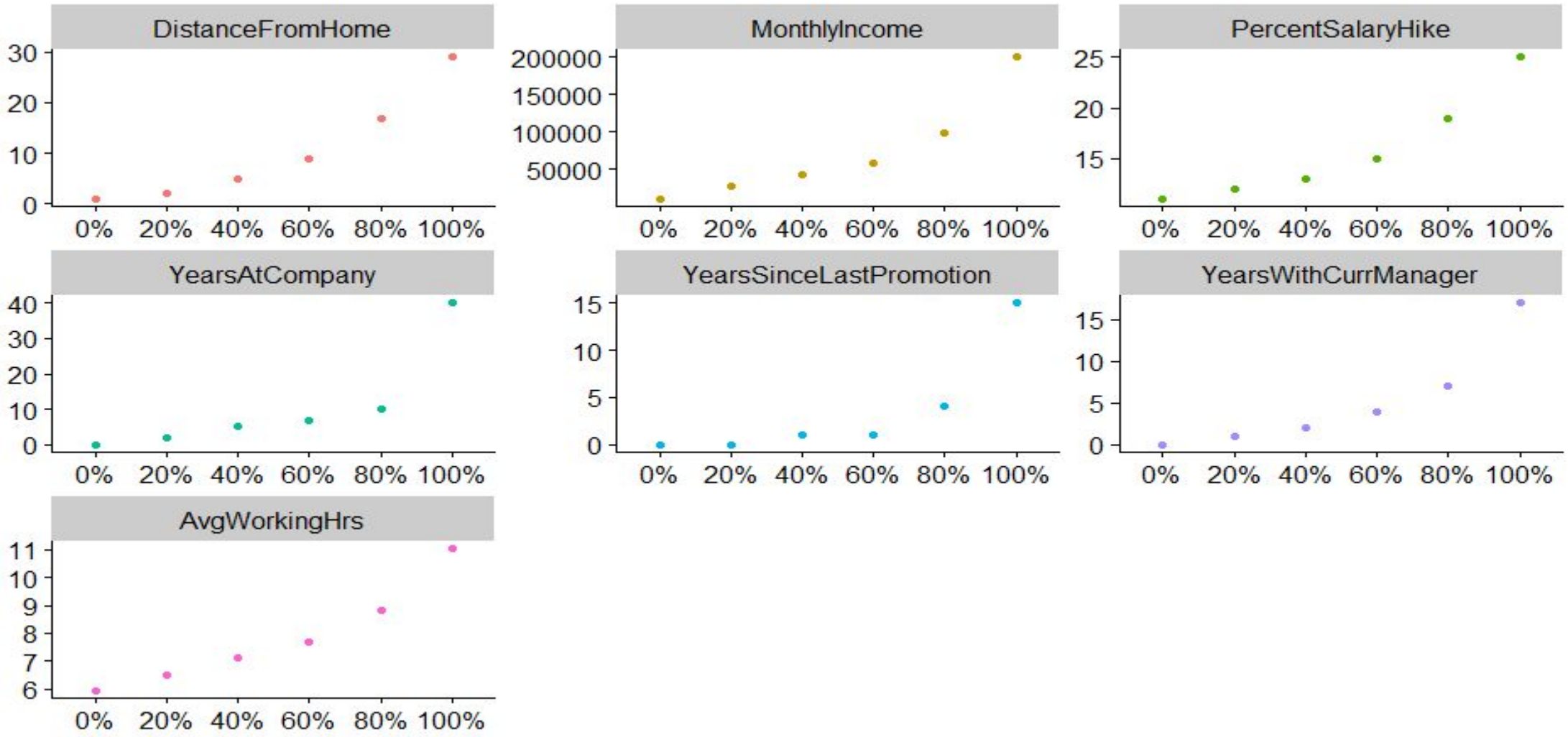
•Analysis of key columns of HR analytics data files..
•Remove unwanted columns having same values & NA
•Check for Duplicate & null values
•Convert Date to standard format

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

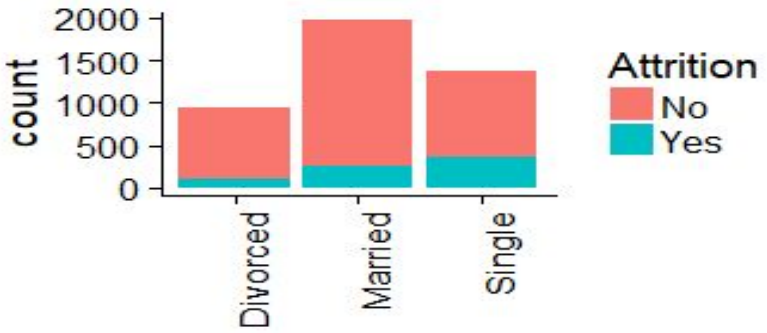
•Use StepAIC and glm() to arrive at key factors influencing the attrition

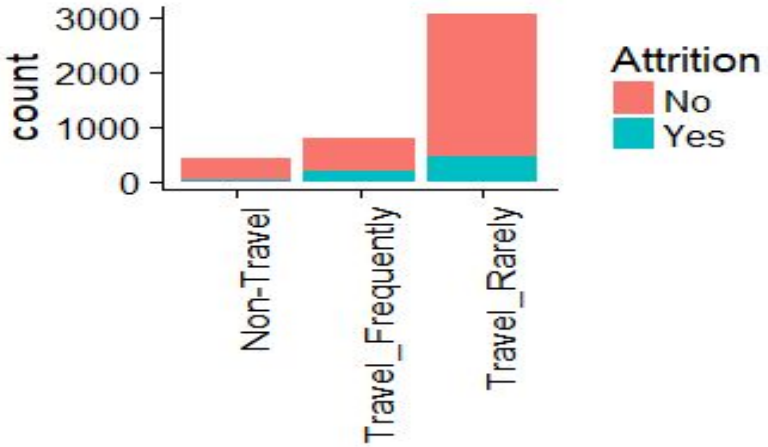
•Identify the model's
• Accuracy
• Sensitivity
• Specificity
•Use Gain, Lift and KS Stats to predict best model

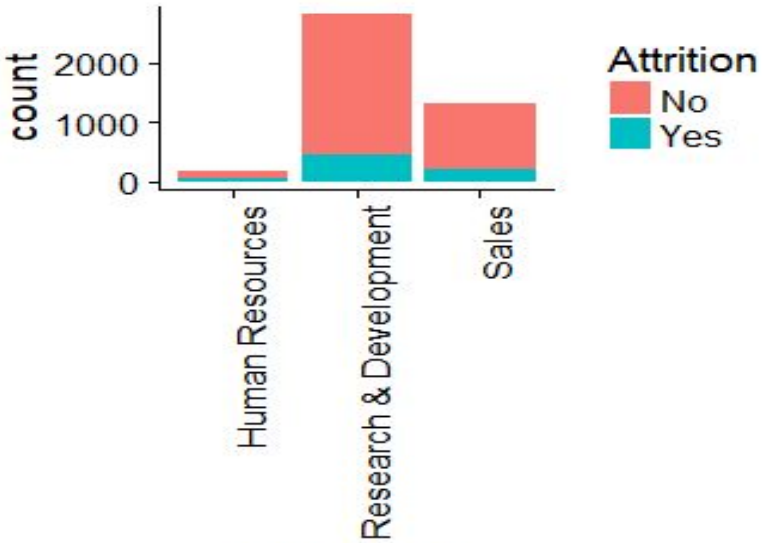
Visual Analysis



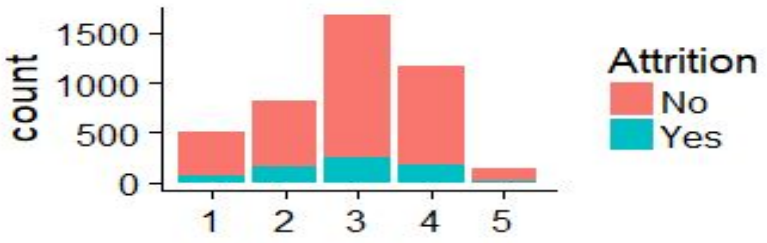
Visual Analysis (contd….)



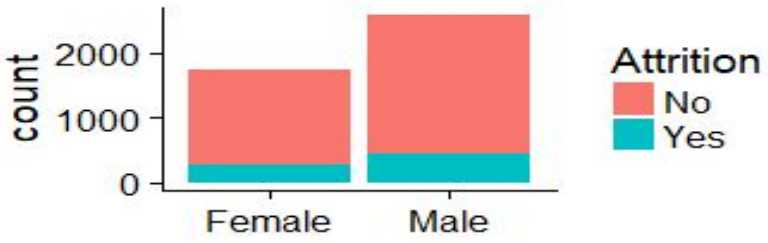




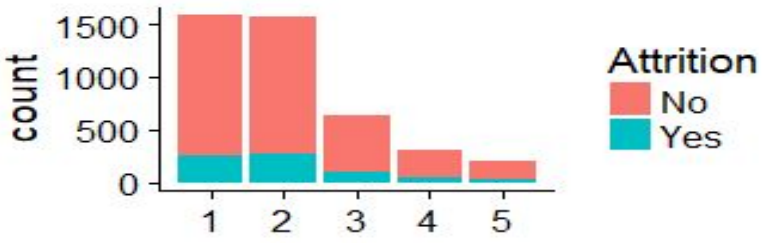
MaritalStatus



BusinessTravel



Department

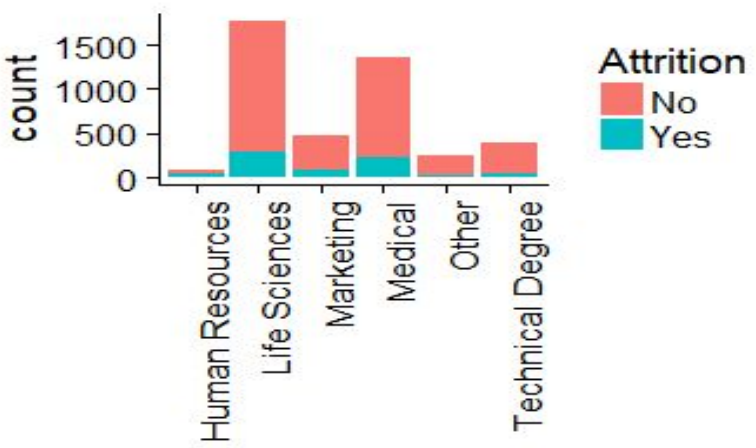


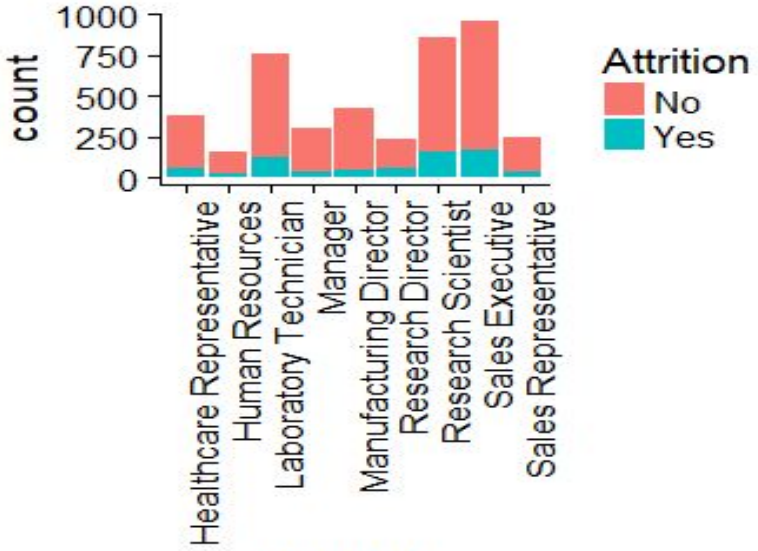
Education

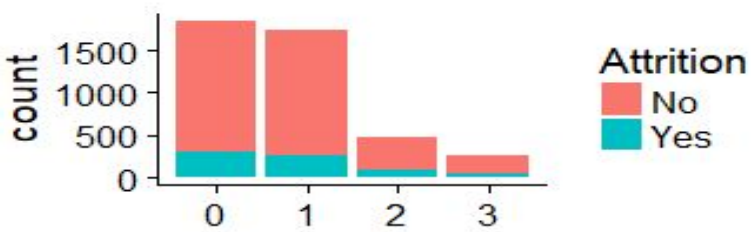
Gender

JobLevel

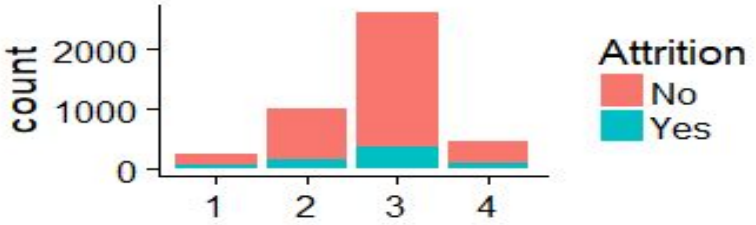
Visual Analysis (contd….)



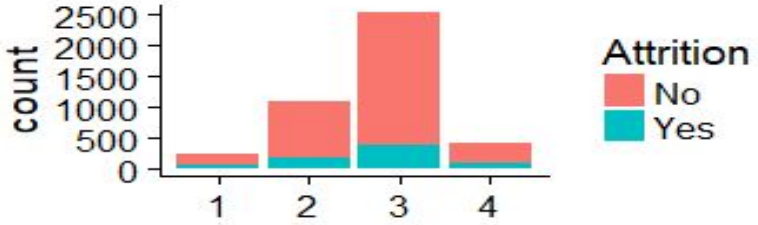




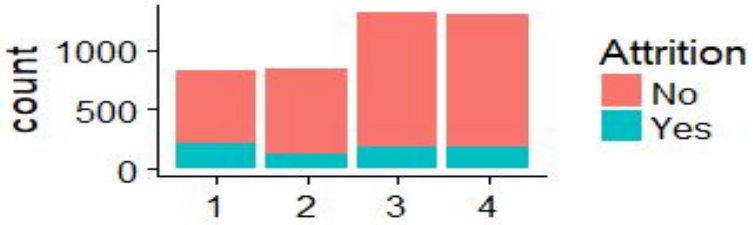
EducationField



JobRole



StockOptionLevel

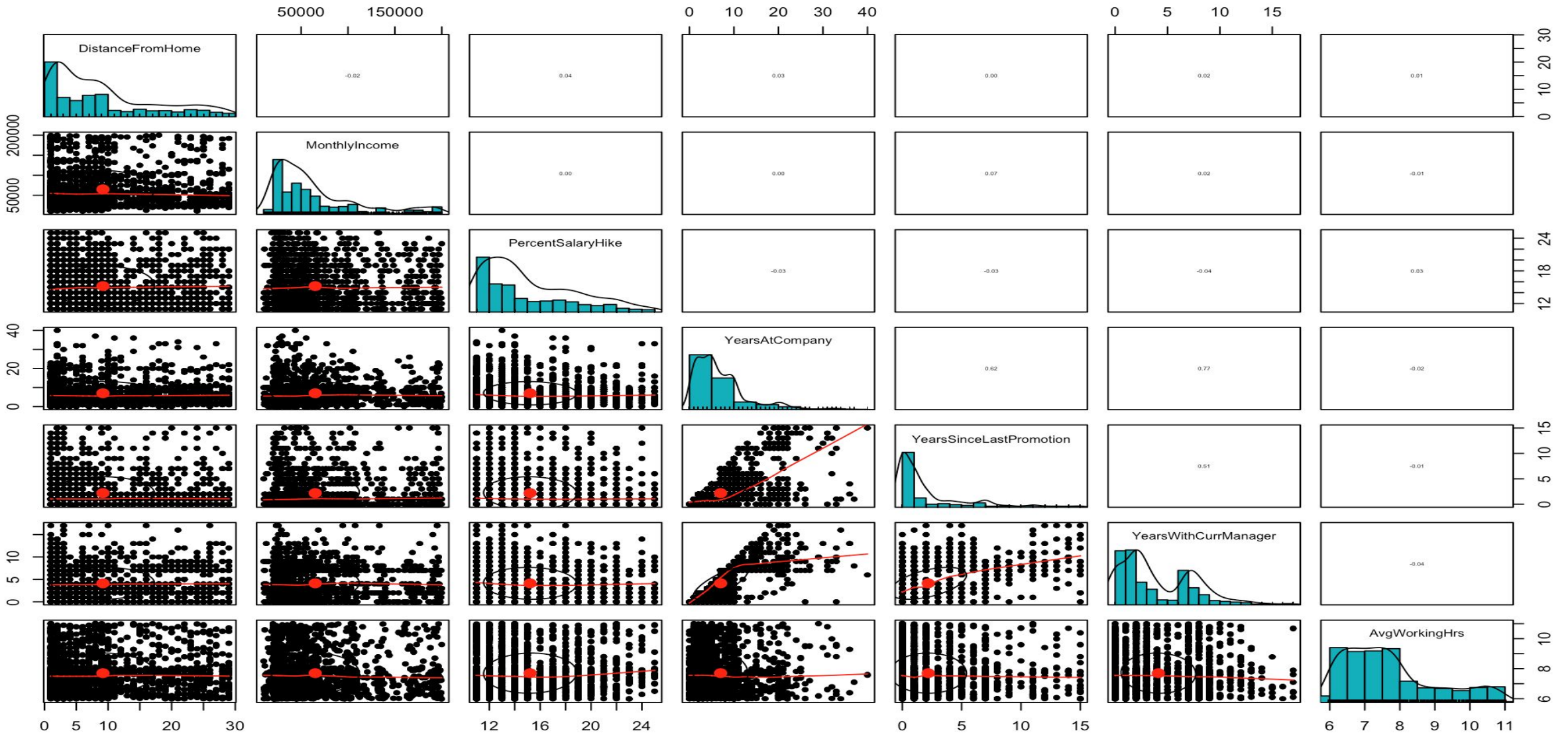


WorkLifeBalance

JobInvolvement

EnvironmentSatisfaction

Correlation Matrix



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	• Employees who gave "**low"** Environment Satisfaction have 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 & 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

1. **Missing Value Imputation:**
   - In order to avoid bias, WOE analysis was used to replace NA values as shown in **Table 1.**
   - Variable "TotalWorkingYears" has only 0.2% values as NA and hence were 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 variables.
   - 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	11-20	21-30					
Age	18-25	26-33	34-37	38-60			
DistanceFromHome	1-2	3-10	11-29				
MonthlyIncome	10090-23130	23140-68770	68770+				
YearsAtCompany	>10	<= 10					
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 4400 observations across -- 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 -- models were created to arrive at final model
- Key Variables:
  - The final model has -- variables which together impact the attrition rate

Model Evaluation

- Use of predict function to come up with predicted probabilities
- Probability range: 0.2% to 87%

Confusion Matrix on 50% Probability:

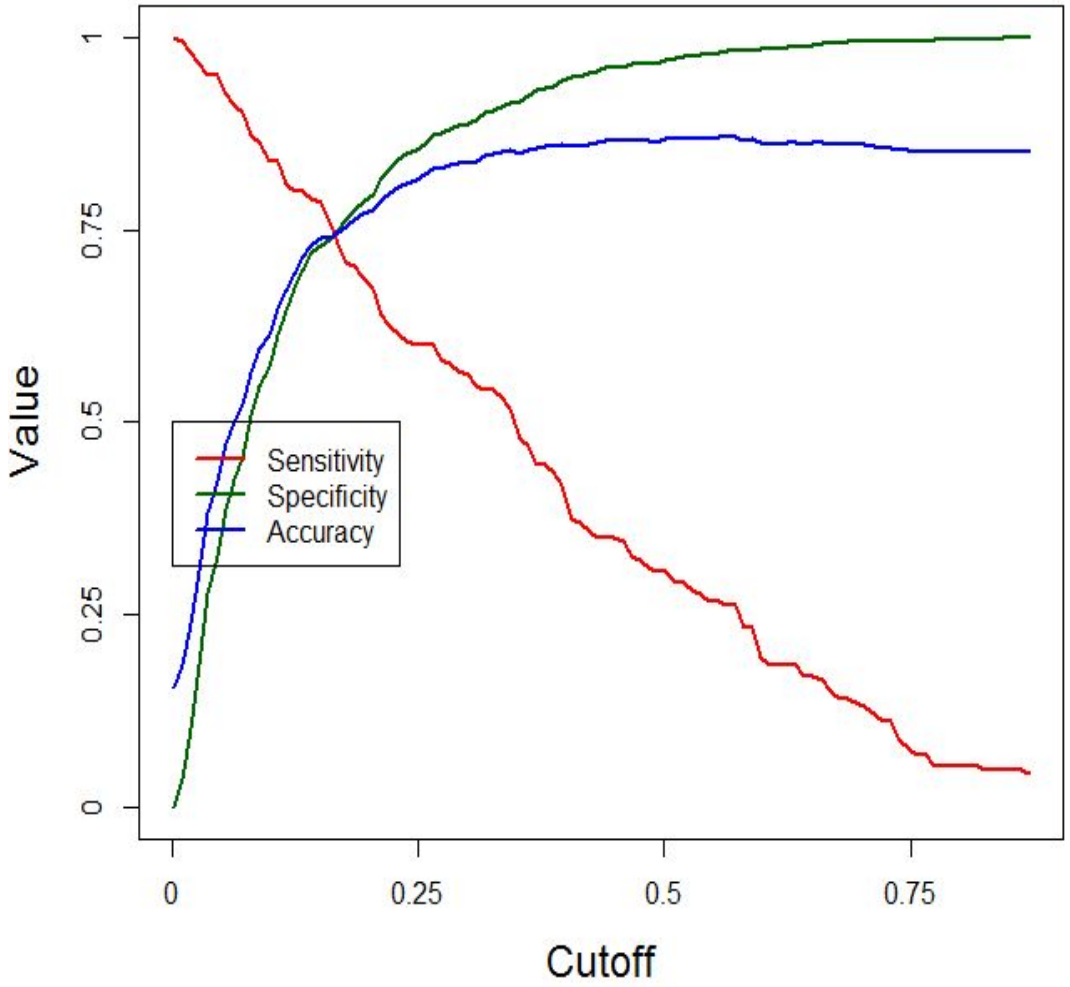
- **Accuracy** : **0.862**
- **Sensitivity** : **0.287**
- **Specificity** : **0.967**

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

To overcome low Sensitivity, the optimal probability threshold is calculated as **0.16**

Confusion Matrix at cutoff:

- **Accuracy** : **0.734**
- **Sensitivity** : **0.741**
- **Specificity** : **0.732**



Model Evaluation (contd.)

Model Evaluation using the below techniques:

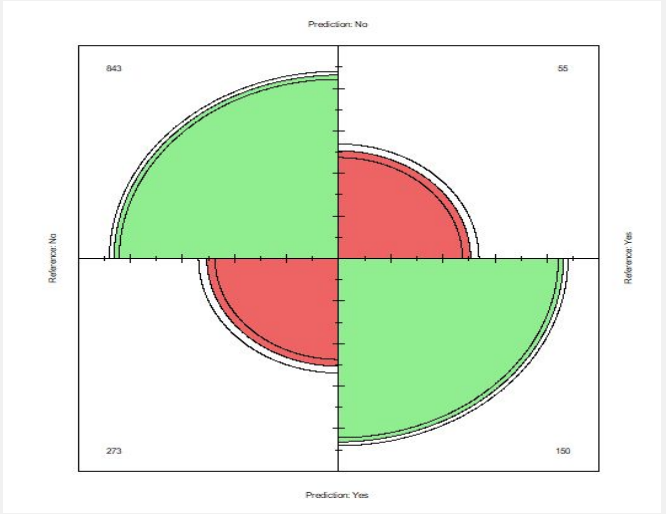
- KS-Statistic
- Lift
- ROC

KS-Stats

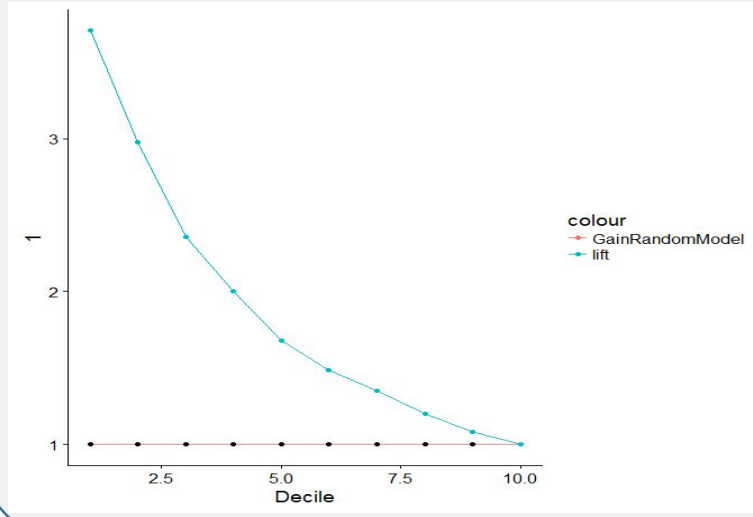
Decile	AttritionCount	CumulativeAttrition	GainPercent	NonAttritionCount	CumulativeNonAttrition	GainPercentNonAttrition	KS_stat	GainRandomModel	lift
1	76	76	37.07317	56.1	56.1	5.026882	32.046289	10	3.707317
2	43	119	58.04878	89,1	142.2	13.010753	45.038028	20	2.902439
3	28	147	71.70732	104.1	249.3	222.33871	49.368607	30	2.390244
4	17	164	80	115.1	364.4	32.65233	47.34767	40	2
5	6	170	82.92683	126.1	490.5	43.951613	38.975216	50	1.658537
6	6	176	85.85366	126.1	616.6	55.250896	30.602762	60	1.430894
7	11	187	91.21951	121.1	737.7	66.102151	25.117362	70	1.303136
8	6	193	94.14634	126.1	863.8	77.401434	16.744908	80	1.176829
9	8	201	98.04878	124.1	987.9	88.521505	9.527275	90	1.089431
10	4	205	100	128.1	1116	100	0	100	1

Model Evaluation (contd.)

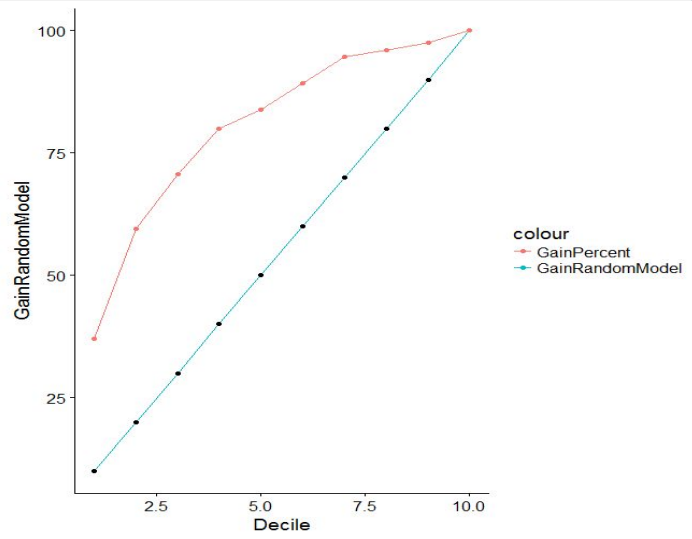
Confusion Matrix



Lift Chart



Gain Chart



Inference:

- KS-Statistic table show max value of 49% 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.

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, 2 and 3 signifying Low/Average Satisfaction
Work Life Balance	Level 1 signifying Bad Work Life Balance
Business Travel	Employees who travel Frequently/Rarely
Department	1. Research & Development 2. Sales
Years Since Last Promotion	Employees who haven't been promoted in last 4 years or more
Marital Status	Single
Number of Companies Worked	Employees who have worked in more than 5 companies
Years With Current Manager	Employees working with the same manager for more than 1-3 years.
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

Recommendations

1. Company should ensure conducive work atmosphere.
2. There should be a healthy relationship between manager & employee.Manager should be considerate on employees career goal.Managers should not be changed very often.
3. Before hiring, they should check whether an individual’s aspirations are inline with companies aspiration or not.
4. There should be a travel policy which should promote travel on rotation basis.
5. Work Life balance of the employee can be improved by organizing team building activities like competitions within departments for cricket, indoor games.
6. Work in team should be divided uniformly to avoid extended working hours.
7. As per our analysis, Salary is not contributing towards Attrition Rate.