HR ANALYTICS CASE STUDY

SUBMISSION

Presented By:
Ishita Aggarwal

Background of HR Analytics Case Study

Business Objective

- To identify the factors affecting attrition
- To Analyse and suggest what changes XYZ should make in their workplace to make their employee stay

Data and Strategy

We have different data sourced as below which are provided from XYZ firm.

Data:

employee_survey_data contains 4410 observations of employee information of firm

Manager_survey_data contains 4410 observations with job involvement and ratings details general_data contains 4410 observation wit employee personal & work details.

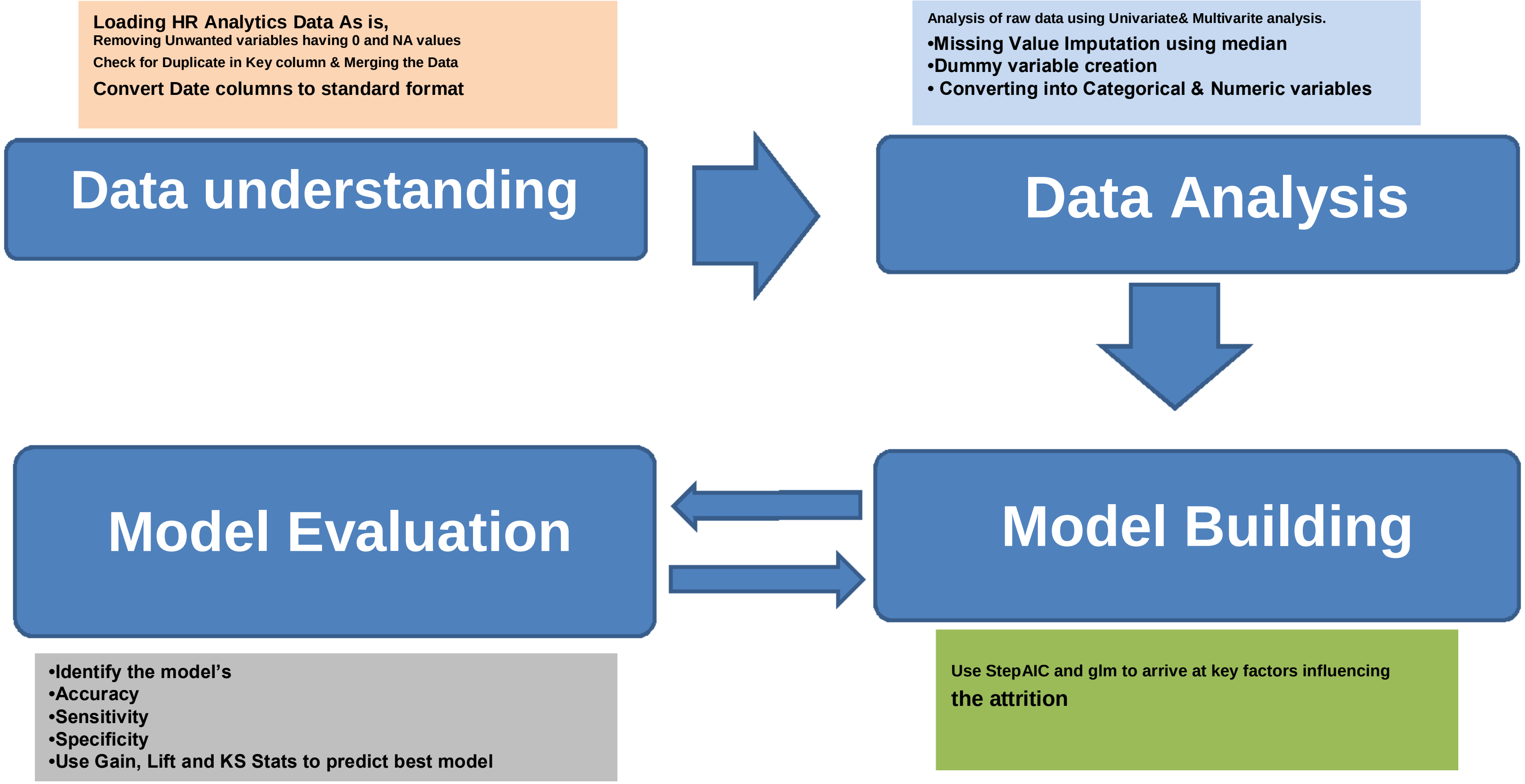
In_time contains in time information of an employee and holiday details of firm.

Out_time contains out time information of an employee and holiday details of firm

Strategy:

We are going to use predictive analysis of Logistic Regression to solve the business problem, i.e. factors affecting the attrition.

Model – Problem/ Solution



Loading HR Analytics Data As is,
Removing Unwanted variables having 0 and NA values
Check for Duplicate in Key column & Merging the Data
Convert Date columns to standard format

Data understanding

Analysis of raw data using Univariate& Multivarite analysis.

- Missing Value Imputation using median
- Dummy variable creation
- Converting into Categorical & Numeric variables

Data Analysis

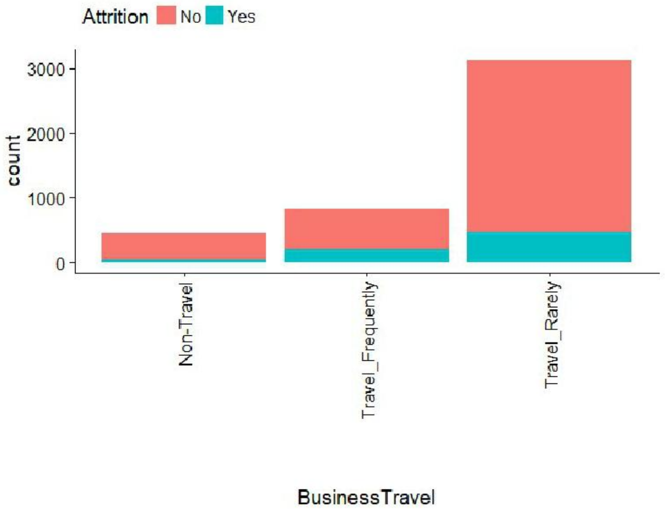
Model Building

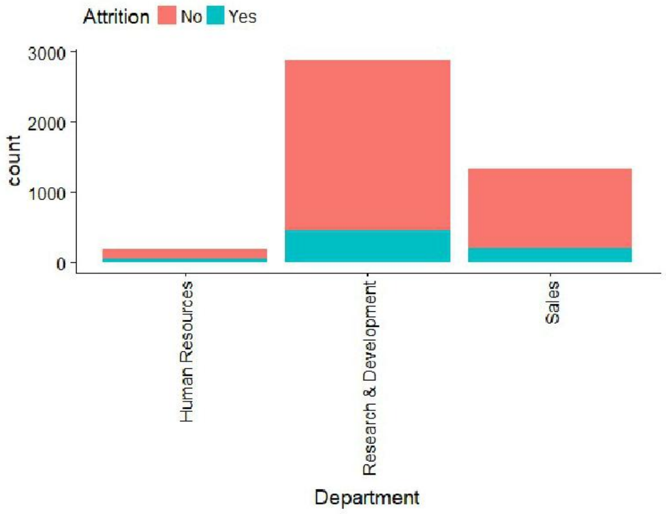
Use StepAIC and glm to arrive at key factors influencing the attrition

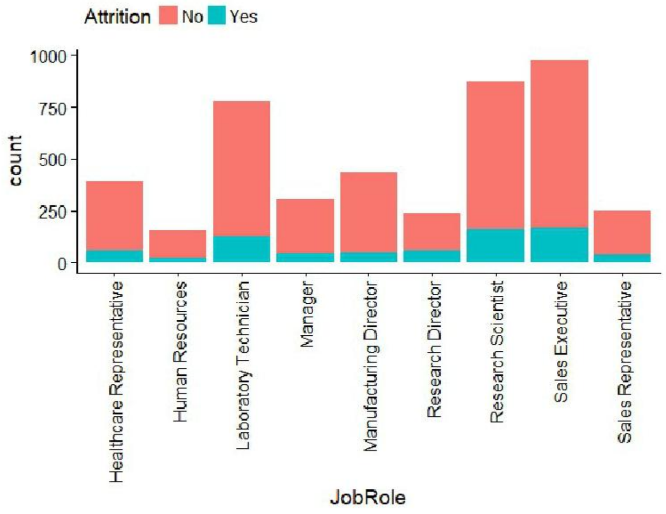
Model Evaluation

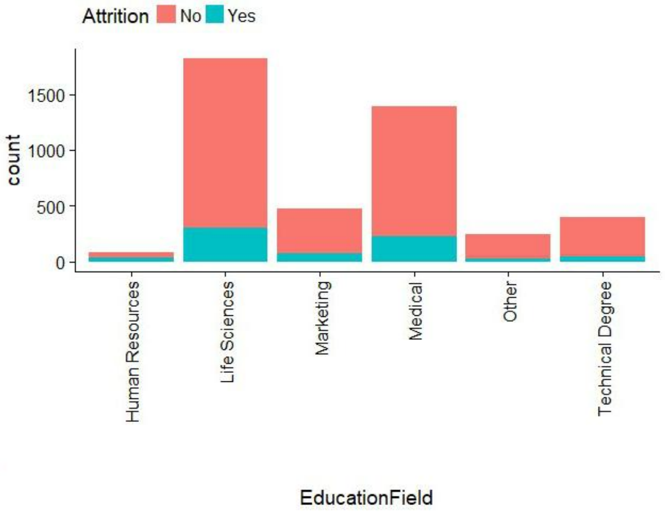
- Identify the model's
- Accuracy
- Sensitivity
- Specificity
- Use Gain, Lift and KS Stats to predict best model

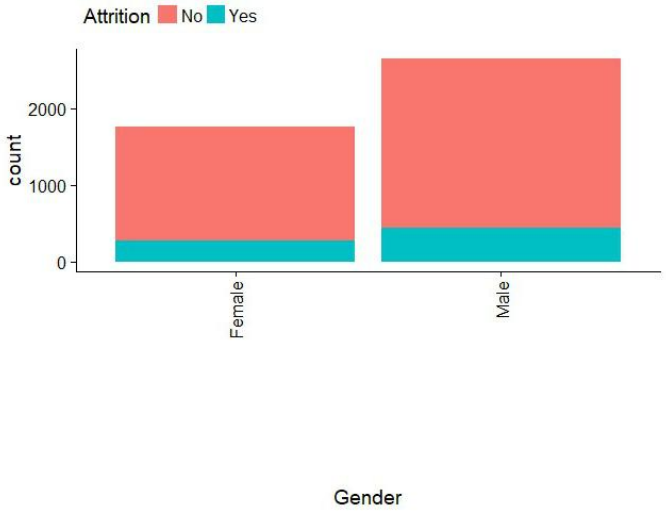
Univariate Analysis on Different Attributes – I (contd)

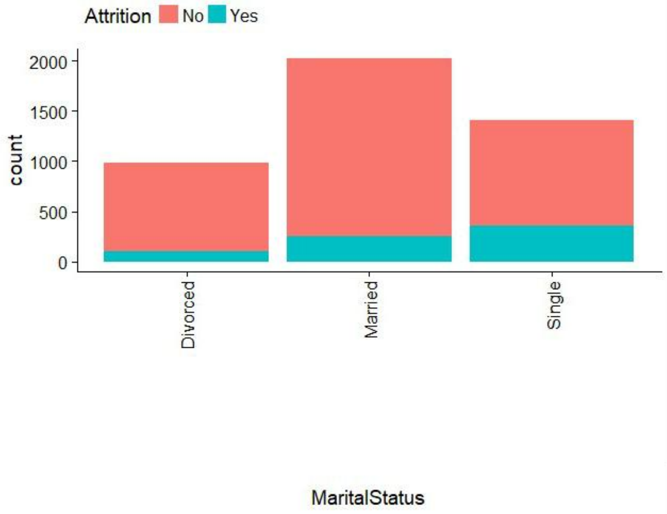




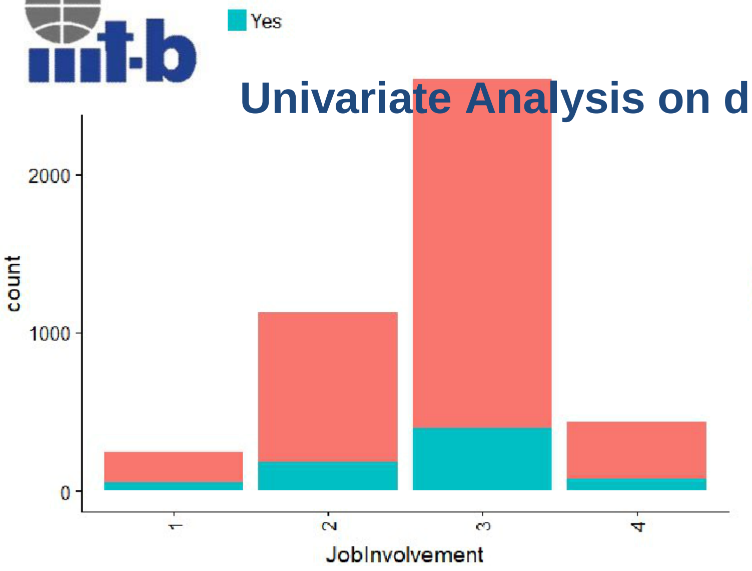


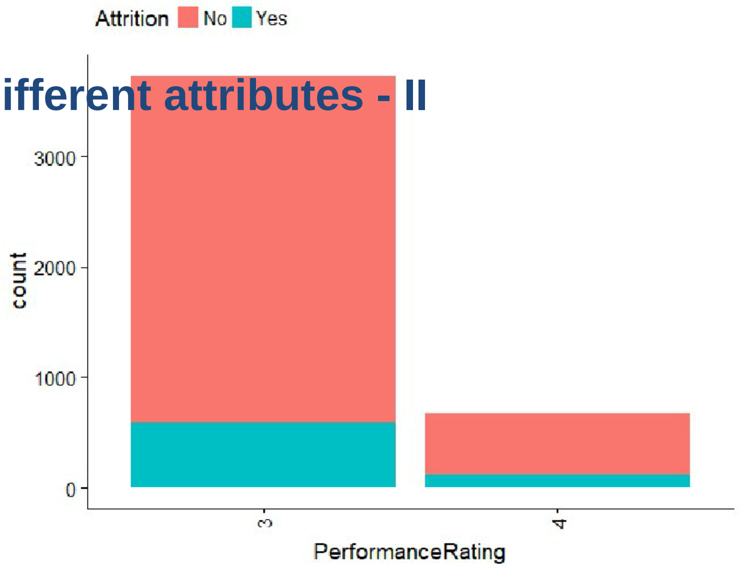


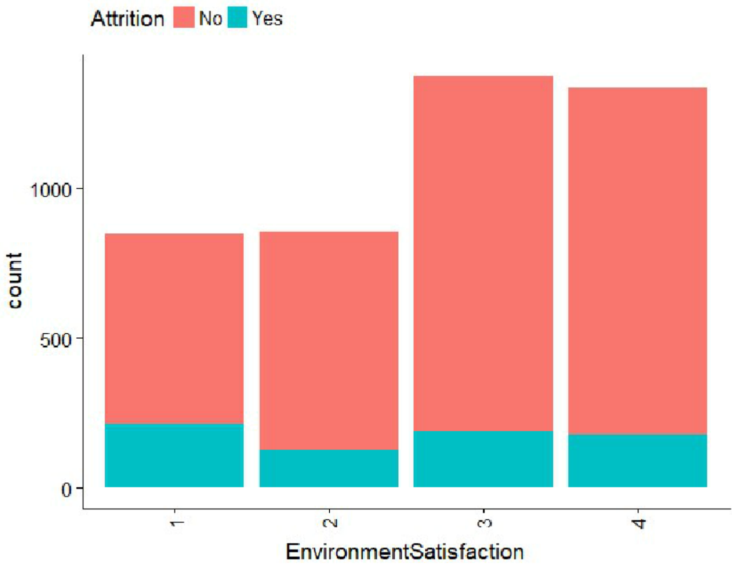


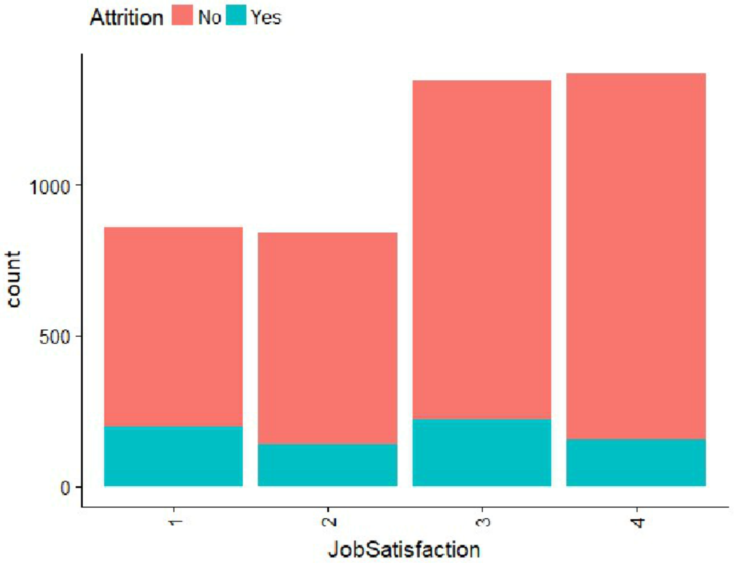


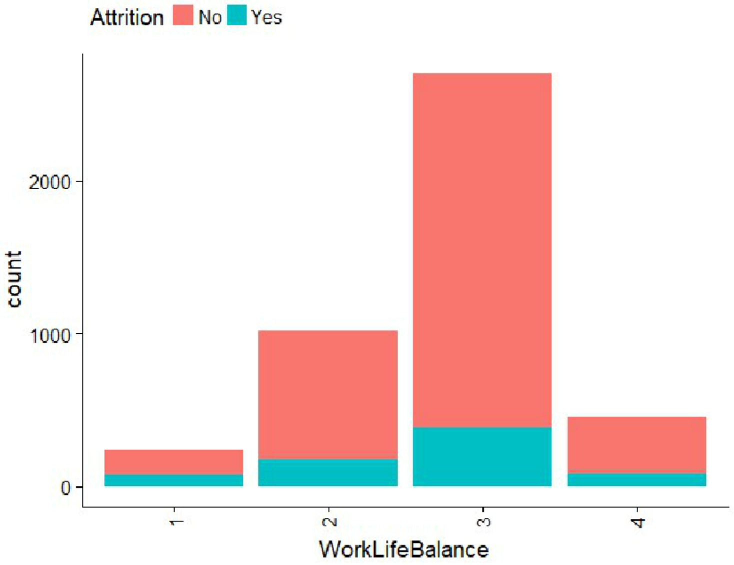
Univariate Analysis on different attributes - II







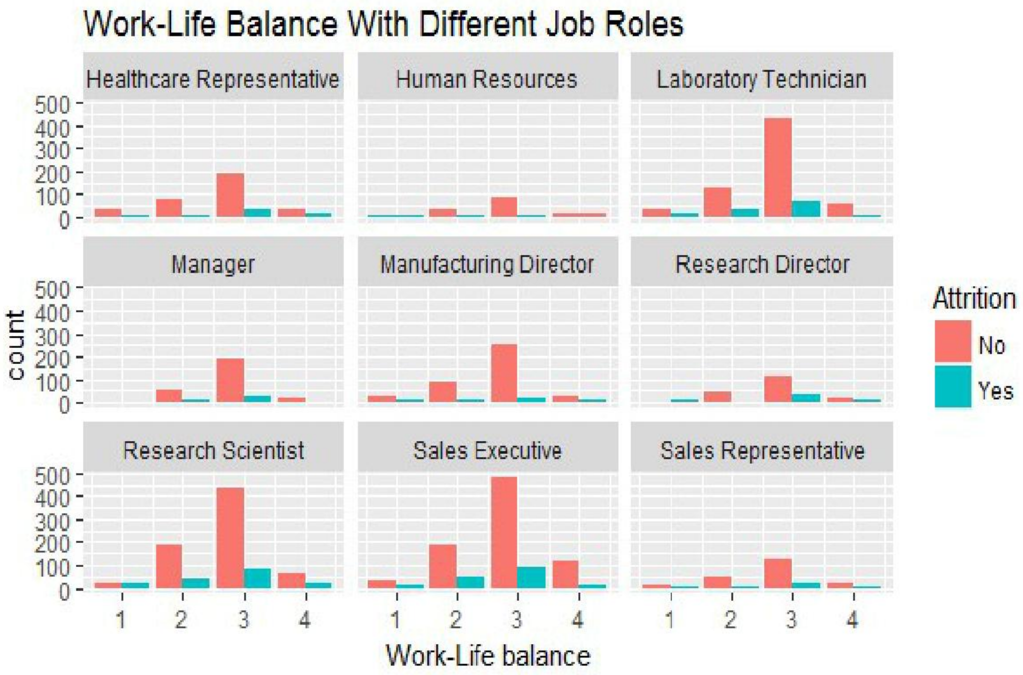


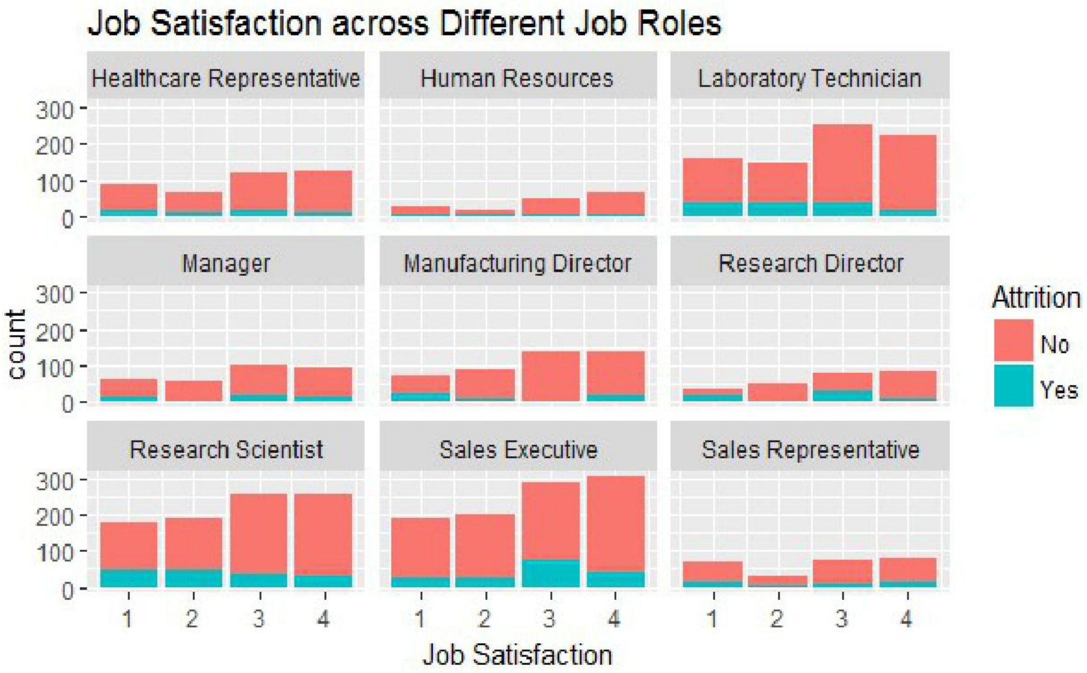


Summary of Univariate Analysis

Gender & Marital Status	Attrition ”.
Education Field	
Business Travel	
Job Involvement	
Job Role	
Work Life Balance	
Performance Rating	

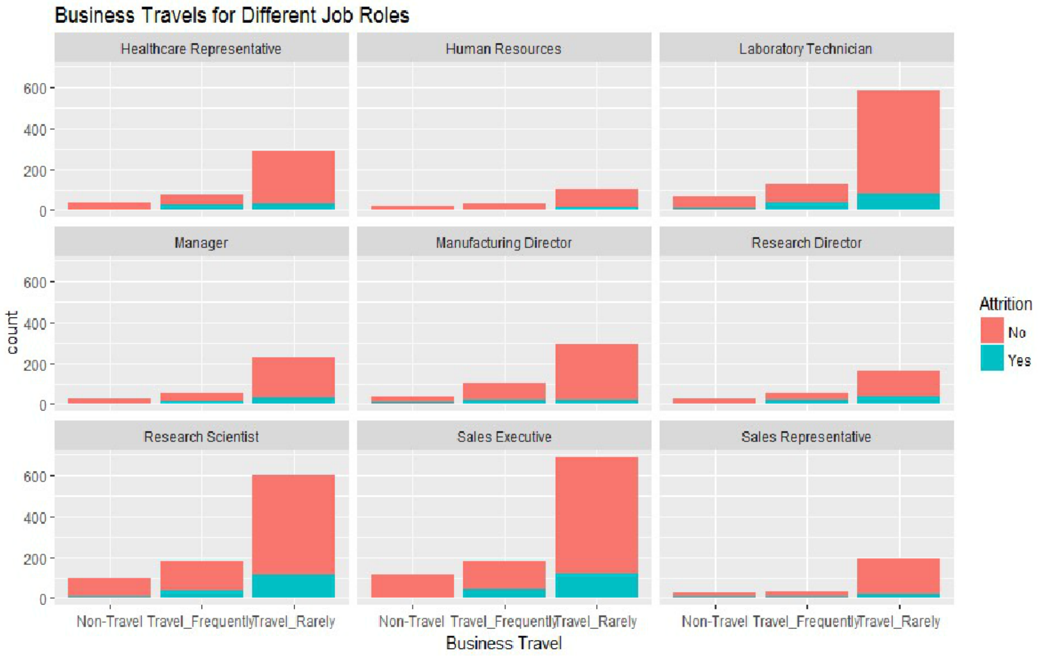
Multivariate Analysis

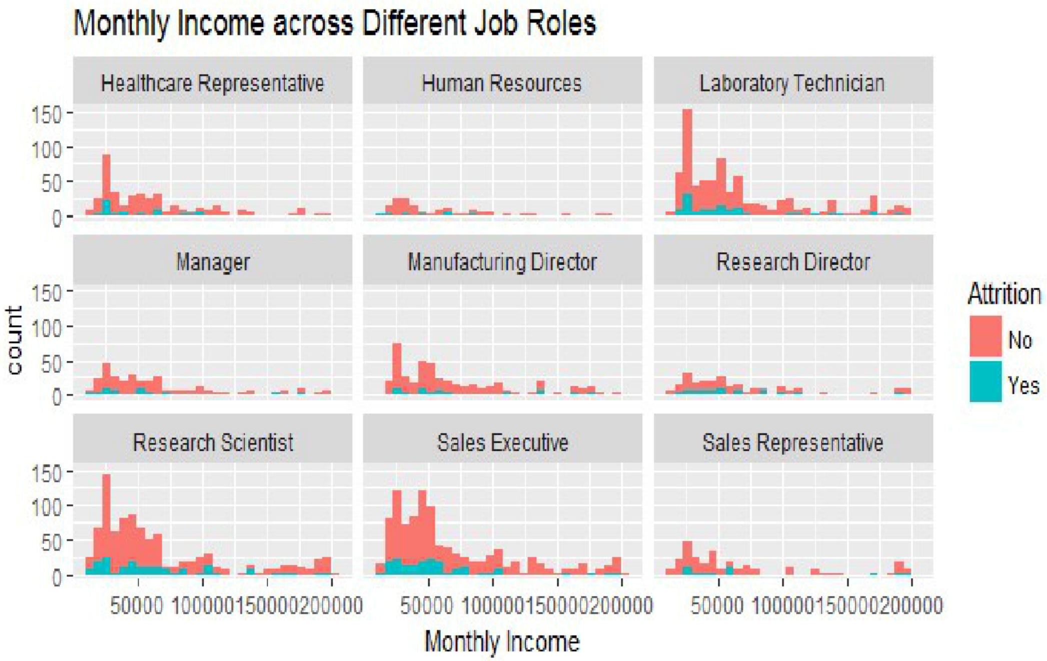




Attrition Rate is comparatively high for Laboratory Technician, Research Scientist and Sales executive. Despite of Employer providing a balance between personal and professional life, employees with the mentioned job roles are switching to different companies. Interestingly, Employees satisfied with their job are also adding to the attrition and is evident primarily in Roles as Research Director and Sales Executives followed by research Scientists and Lab technician.

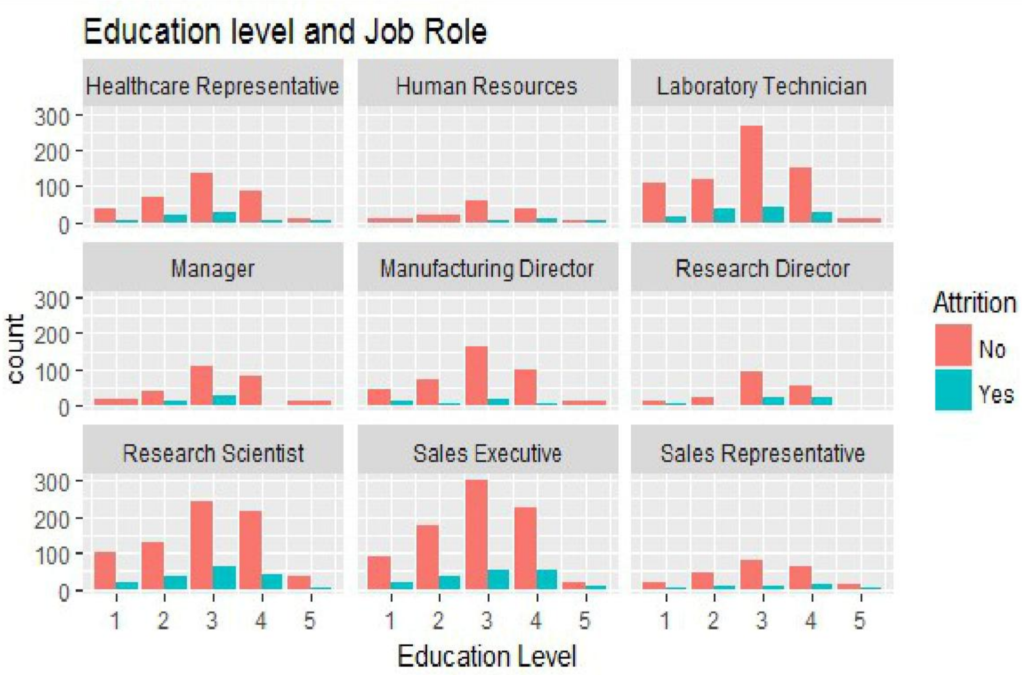
Multi Variate Analysis – II (contd)

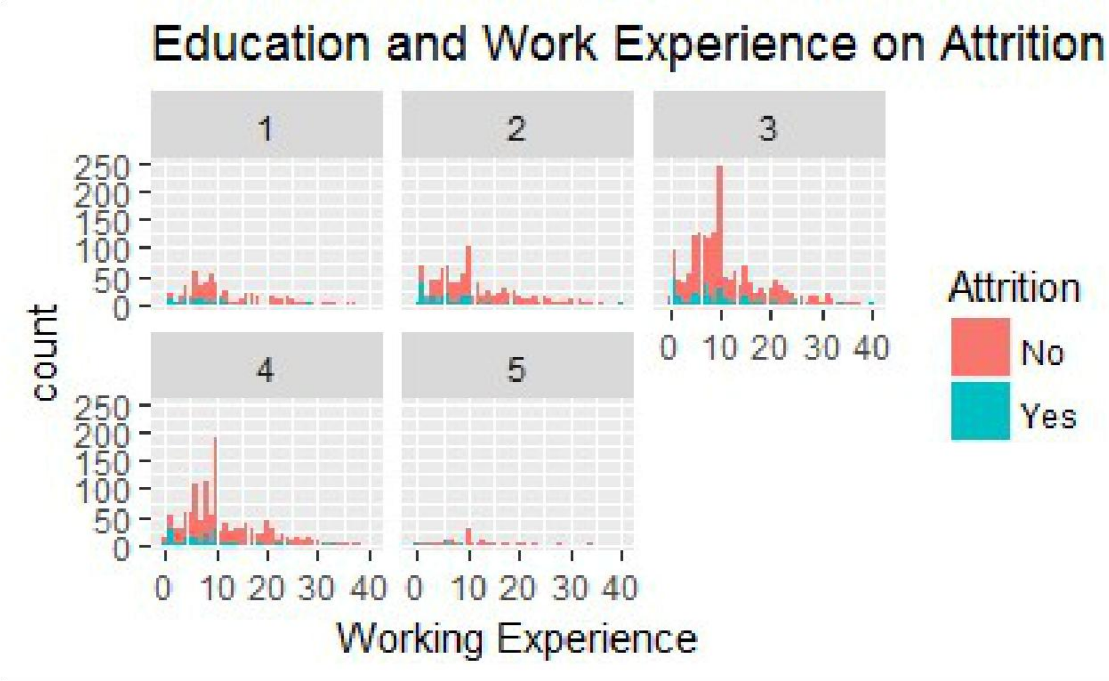




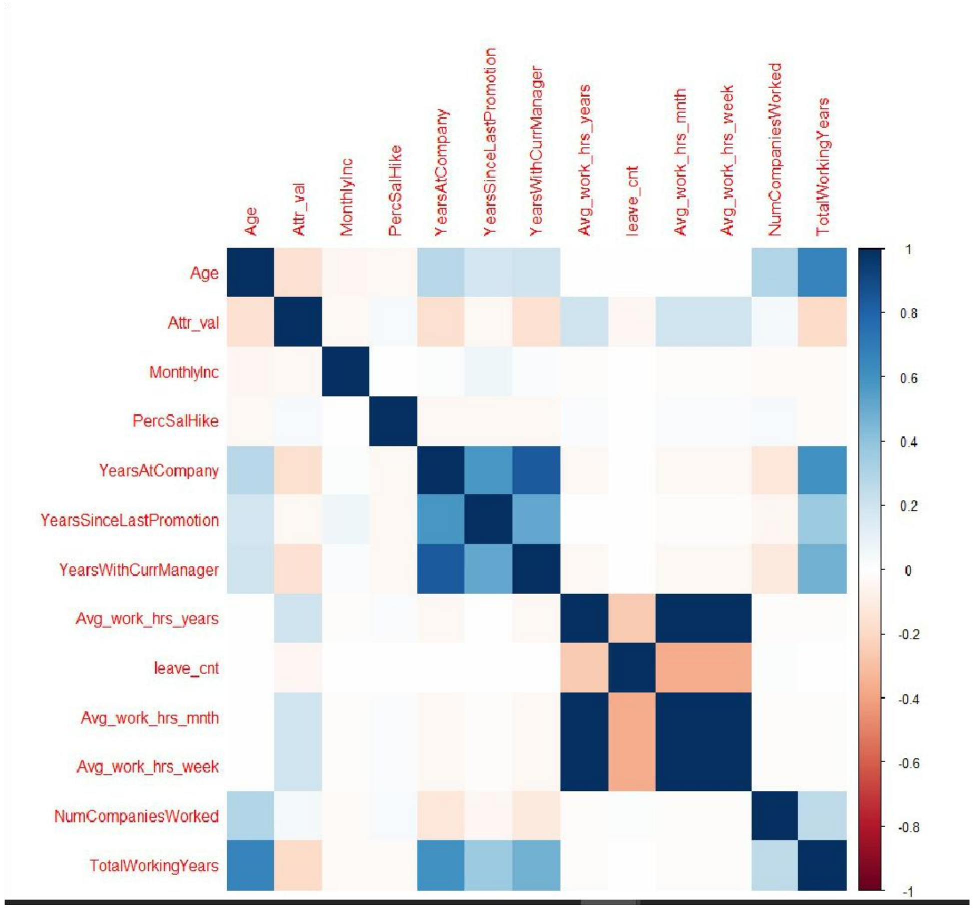
- It is clearly evident that Business Travel plays a role in retaining employees. Attrition Rate is high across different Job roles as there are rarely any Business Travel.
- **Employees with Monthly Income of $75000 or less prefer changing the organisation more frequently.**

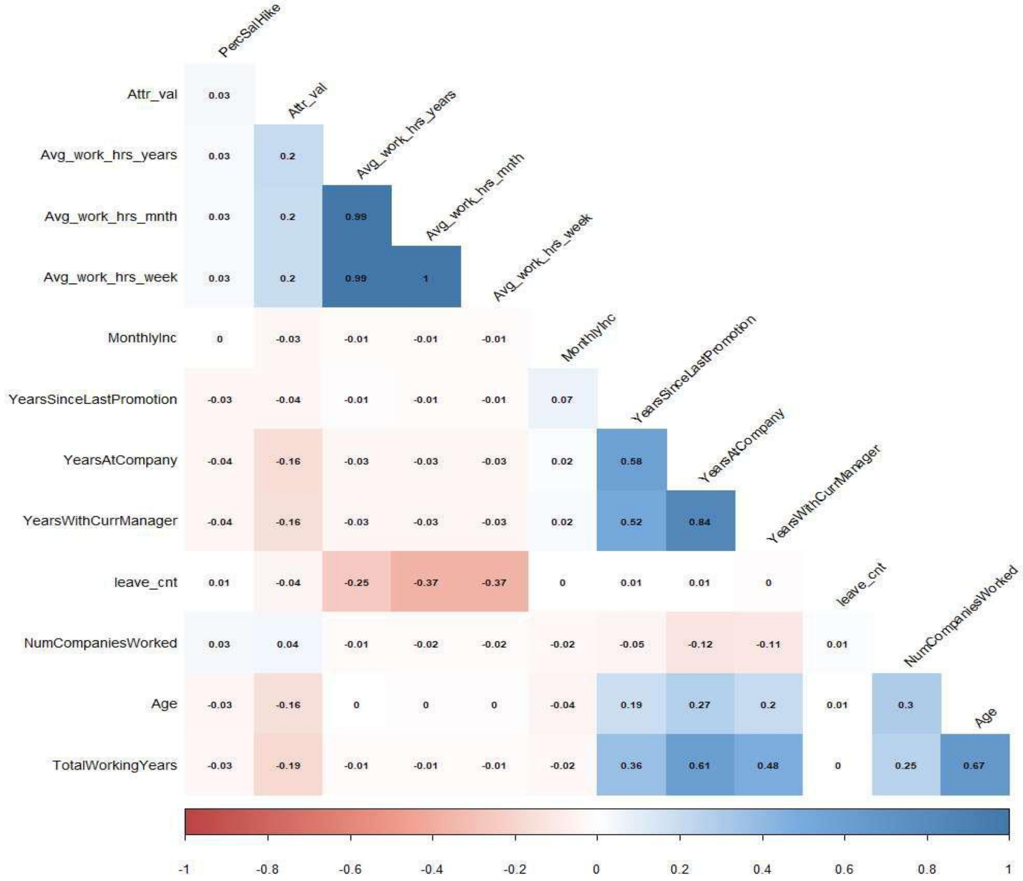
Multi Variate Analysis (Educataion Level/ Work Experience) -- Contd





Correlation Matrix





	PercSalHike	Attr_val	Avg_work_hrs_years	Avg_work_hrs_mnth	Avg_work_hrs_week	MonthlyInc	YearsSinceLastPromotion	YearsAtCompany	YearsWithCurrManager	leave_cnt	NumCompaniesWorked	Age
Attr_val	0.03											
Avg_work_hrs_years	0.03	0.2										
Avg_work_hrs_mnth	0.03	0.2	0.99									
Avg_work_hrs_week	0.03	0.2	0.99	1								
MonthlyInc	0	-0.03	-0.01	-0.01	-0.01							
YearsSinceLastPromotion	-0.03	-0.04	-0.01	-0.01	-0.01	0.07						
YearsAtCompany	-0.04	-0.16	-0.03	-0.03	-0.03	0.02	0.58					
YearsWithCurrManager	-0.04	-0.16	-0.03	-0.03	-0.03	0.02	0.52	0.84				
leave_cnt	0.01	-0.04	-0.25	-0.37	-0.37	0	0.01	0.01	0			
NumCompaniesWorked	0.03	0.04	-0.01	-0.02	-0.02	-0.02	-0.05	-0.12	-0.11	0.01		
Age	-0.03	-0.16	0	0	0	-0.04	0.19	0.27	0.2	0.01	0.3	
TotalWorkingYears	-0.03	-0.19	-0.01	-0.01	-0.01	-0.02	0.36	0.61	0.48	0	0.25	0.67

Data Manipulation

Variable Name	NA count	Missing Value Treatment
NumCompaniesWorked	19	Replaced NA value with Median
TotalWorkingYears	9	Replacing NA with 1 or 0 if no of companies worked is 1 or 0, other than these value we are replacing NA's this with 11
WorkLifeBalance	38	Replaced NA value with Median
JobSatisfaction	20	Replaced NA value with Median
EnvironmentSatisfaction	25	Replaced NA value with Median

1. Scaling

Performed Scaling different continuous variables

2. Dummy Variables

Introduced dummy variables for all categorical variables

3. Outlier treatment

Performed outlier treatment for below variables

- YearsAtCompany
- YearswithCurrentManager
- YearsSinceLastPromotion
- TotalWorkingYears
- TrainingTimesLastYear

4. Derived Variables Introduced

new variables like

- Leave_cnt,
- Avg_work_hrs_year, avg_work_hr_mnth
- Avg_work_hrs_week

Model Building

Model Building

- Using glm model for logistic regression a final dataset of 4410 obs and 59 variables is used for building model.

Training

- We used 70% of observations as train and 30% of data as test
- StepAIC is used to improve performance of model by eliminating insignificant variables
- VIF is used to eliminate variable with high p-value > 0.05

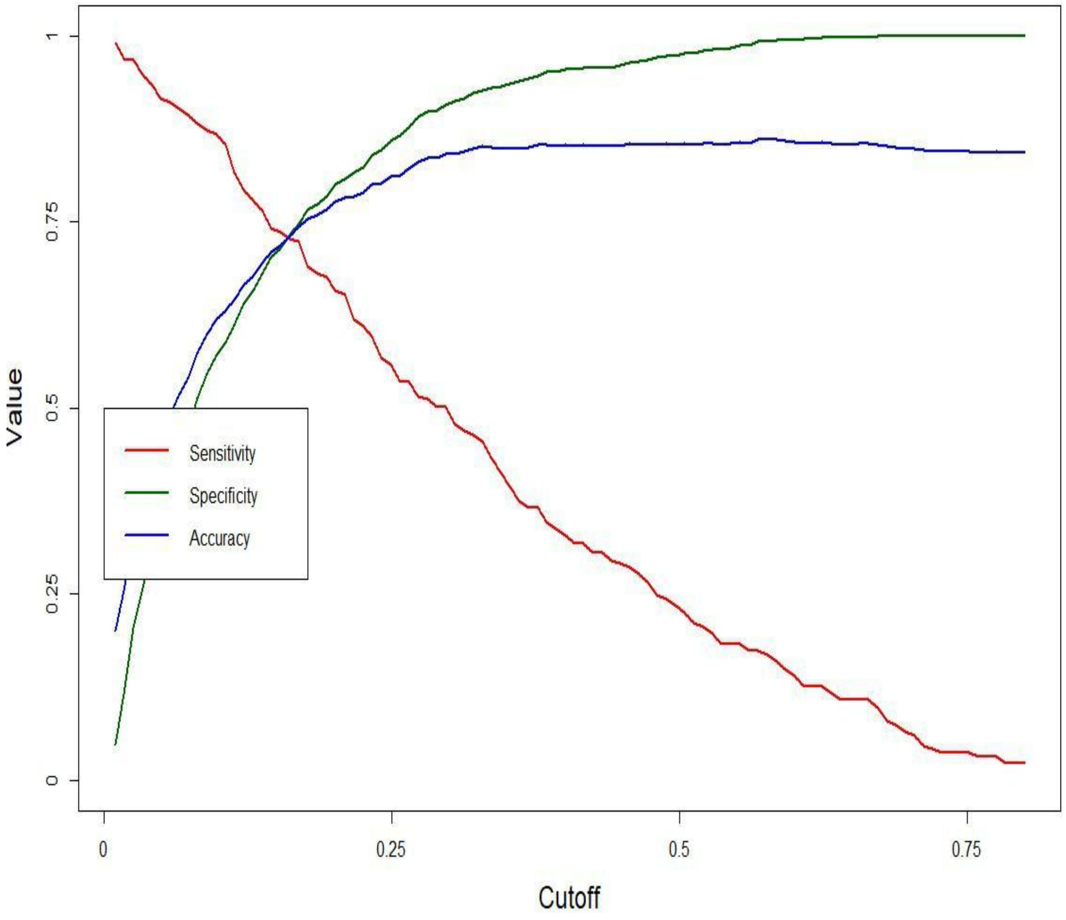
Results

- Total of 19 models were created to arrive at final model
- Key Variables:
- The final model has 16 variables which together impact the attrition rate

Factors affecting Attrition

NumCompaniesWorked
TotalWorkingYears
TrainingTimesLastYear
YearsSinceLastPromotion
YearsWithCurrManager   Being with Same manager for more than 1 years has to be focused
Avg_work_hrs_year   If the average work hours for an employee is more than 9 it has to be focused
BusinessTravel.xTravel_Frequently
BusinessTravel.xTravel_Rarely
Department.xResearch...Development   Employees belong to research department has to be focused
Department.xSales
MaritalStatus.xSingle   Marital Status as Single turns out to be cause for attrition
EnvironmentSatisfaction.x2
EnvironmentSatisfaction.x3
EnvironmentSatisfaction.x4
JobSatisfaction.x4   Level 4 signifying poor job satisfaction which turns to be cause
WorkLifeBalance.x3   Level 3 signifying poor work life balance

Model Evaluation



Confusion matrix on Probability with 40%

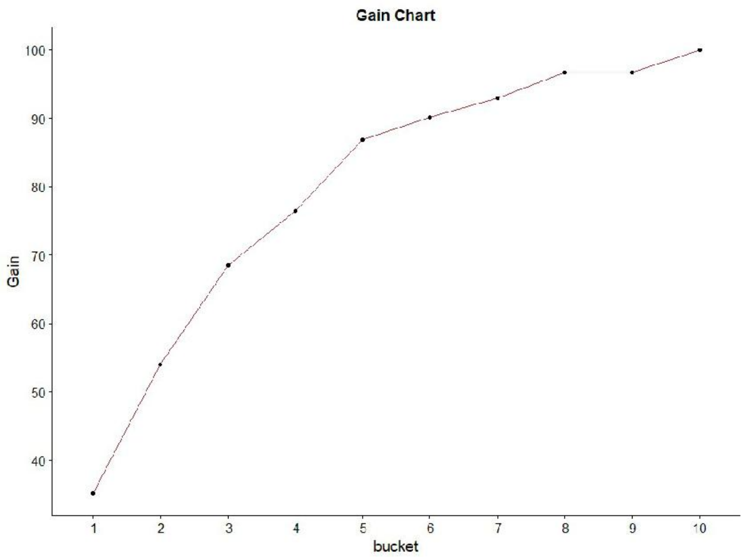
Accuracy -> 0.85
Sensitivity -> 0.32
specificity-> 0.95
It clearly shows Sensitivity is very poor

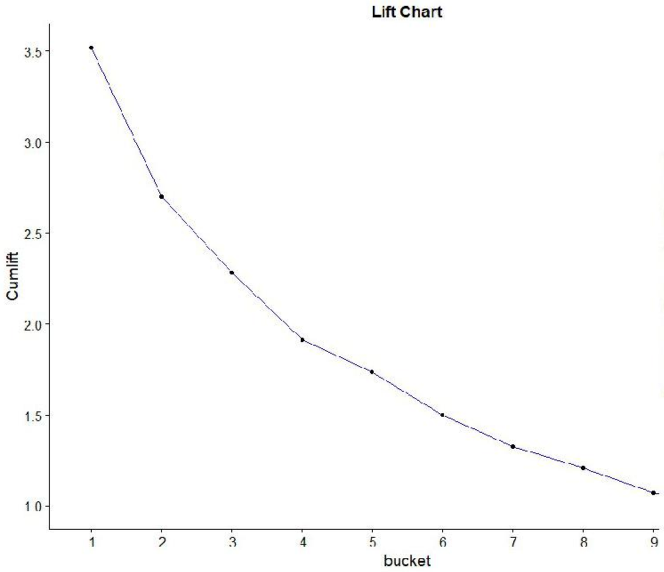
To overcome low Sensitivity, user defined function created to identify cutoff value •Optimal probability threshold for best prediction: 0.161

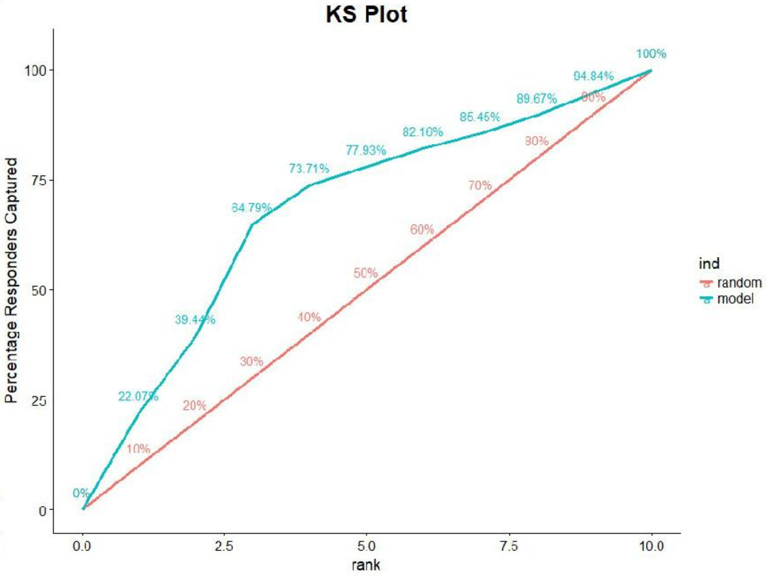
Confusion Matrix at cutoff level of 0.161

•Accuracy : 0.7316
•Sensitivity : 0.7276
•Specificity : 0.7324

Lift Chart/Gain Chart /KS -Plot







1. **The Gain chart infers that the model covers 73% in 4th decile .**
2. **KS static for the model is 0.46 (46%) and it is calculated by KS_table.**
3. **KS plot infers the model prediction is good compared to random model.**

Thank You