HR ANALYTICS CASE STUDY

PROBABILITY OF ATTRITION

Team Members

Praval Rastogi

Deepsi Shrivastava

Prakhar Shukla

Lakshmipathi Kakarla

Analysis Overview

- Business Objective
- Data Exploration
- Data Cleaning and Manipulation
- Data Analysis
- Model Building
- Model Evaluation
- Conclusion

Business Objective

Business Objective

- Minimize attrition rate by improving its retention strategies by developing a real time solution to target high risk employees and accordingly take better decision.

Data Exploration - 1

Data Exploration

- 4410 employees records with following information
  - General data about employees (Age, Gender, Income, Experience, Attrition etc.)
  - Employee survey about environment & job satisfaction, work life balance
  - Manager survey about employee's job involvement & performance rating.
  - Employees login and logout data for the year 2015

- Unordered Categorical Variables (Nominal)
  - BusinessTravel
  - Department
  - EducationField
  - Gender
  - JobRole
  - Over18

Data Exploration - II

Data Exploration

- Ordered Categorical Variables (Ordinal)
  - Education
  - JobLevel
  - MaritalStatus
  - StockOptionLevel
  - EnvironmentSatisfaction
  - JobSatisfaction
  - WorkLifeBalance
  - JobInvolvement
  - PerformanceRating

- Continuous Variables (Numerical)
  - Age
  - DistanceFromHome
  - EmployeeCount
  - MonthlyIncome
  - NumCompaniesWorked
  - PercentSalaryHike
  - StandardHours
  - TotalWorkingYears
  - TrainingTimesLastYear
  - YearsAtCompany,
  - YearsSinceLastPromotion
  - YearsWithCurrManager

Data Preparation and Processing

Data Preparation and Processing

- Missing Values:
  - Replaced missing values in below columns with mostly repeated value in the same column
    - NumCompaniesWorked
    - EnvironmentSatisfaction
    - JobSatisfaction
    - WorkLifeBalance
  - Replaced missing values in "TotalWorkingYears" with "YearsAtCompany"
- Removed outliers in below columns based on boxplot and quantile function
    - MonthlyIncome
    - NumCompaniesWorked
    - TotalWorkingYears
    - YearsAtCompany
    - YearsSinceLastPromotion
    - YearsWithCurrManager
    - AvgWorkHrs
- Removed columns that had identical data for all observations (E.g. EmployeeCount, Over18 & StandardHours)
- Removed columns that had unique values for all observations as analysis done on the collection (E.g. EmployeeID)
- Derived new metrics(Average Working Hours and Total Leaves) based on employee login and logout data
- Scaled continuous variables and created dummy variables for categorical variables for modelling

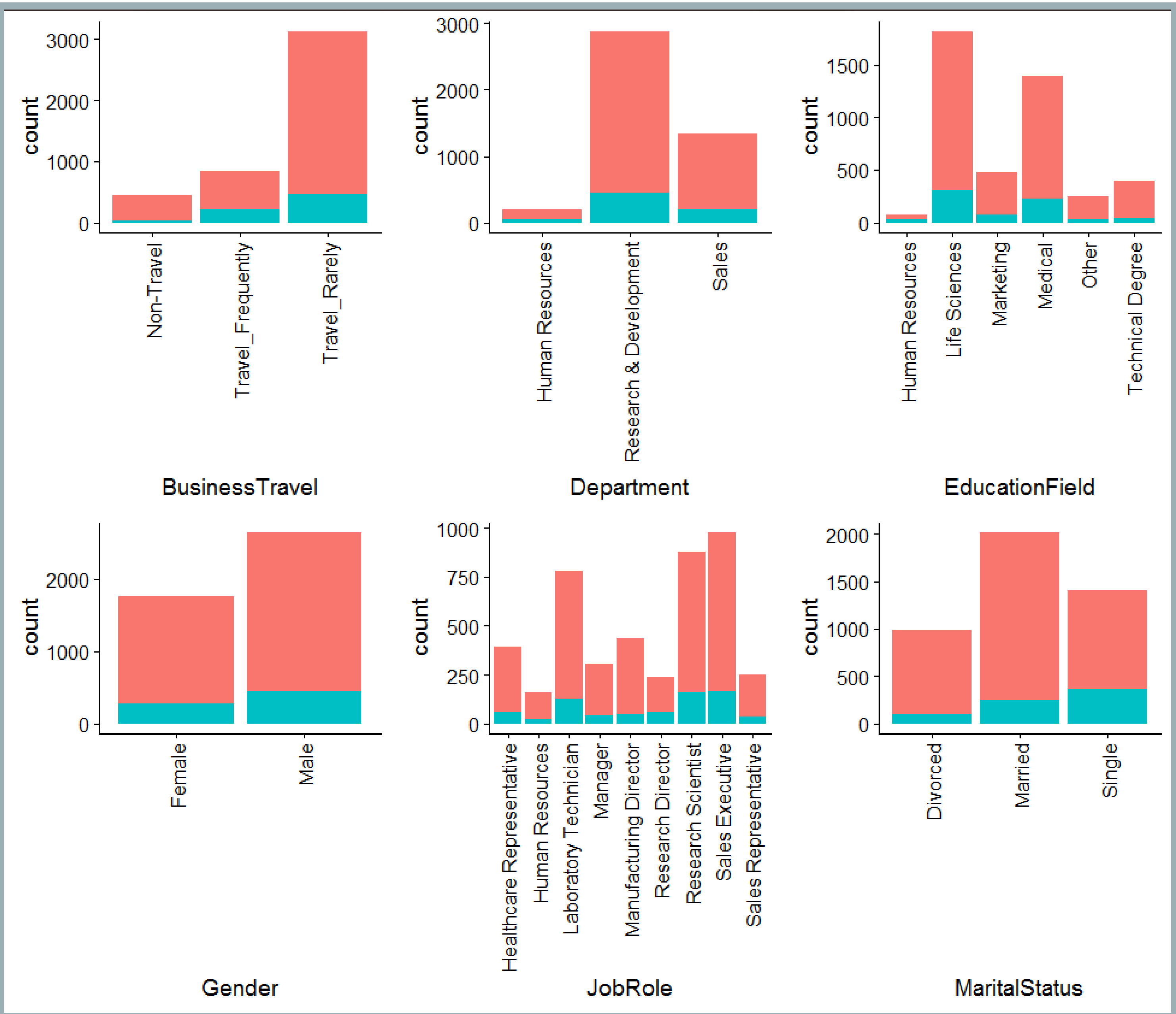
Assumptions

In-time and Out-time:

- Holiday: Day where no employees logged-in and logged-out and excluded those columns(variables).
- Working Hours: Considered as difference between logged-in and logged-out in a day
- Leave: Considered if logged-in data is not available(Exclude Public Holidays)

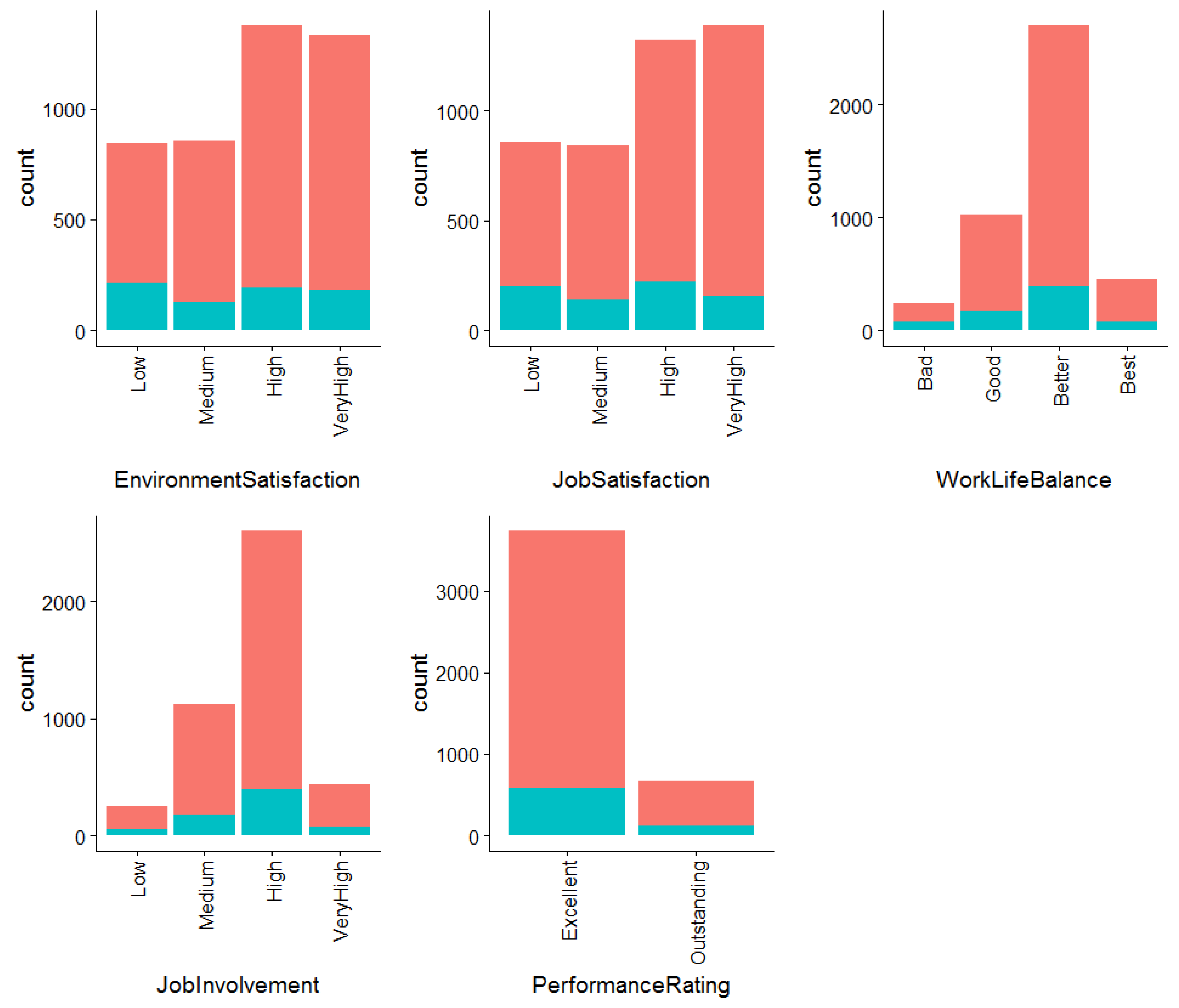
Data Analysis - Categorical Variables - 1

- Analysis on following set of categorical variables (BusinessTravel, Department, EducationField, Gender, JobRole, MartialStatus) indicates that,
  - unmarried employees have significant impact on attrition rate, followed by employees who travel frequently.



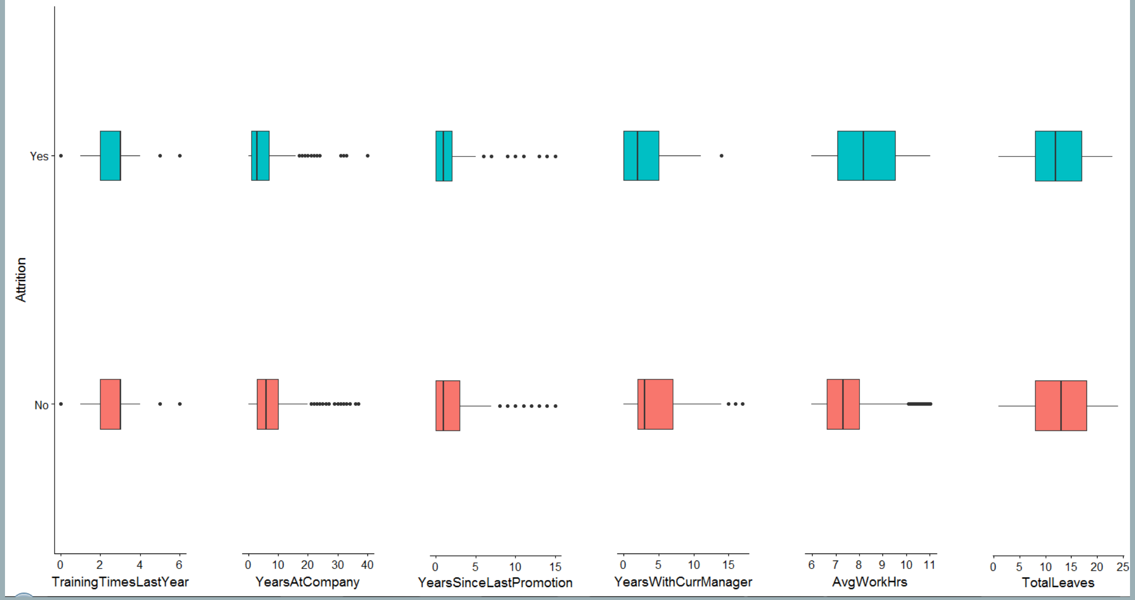
Data Analysis - Categorical Variables - II

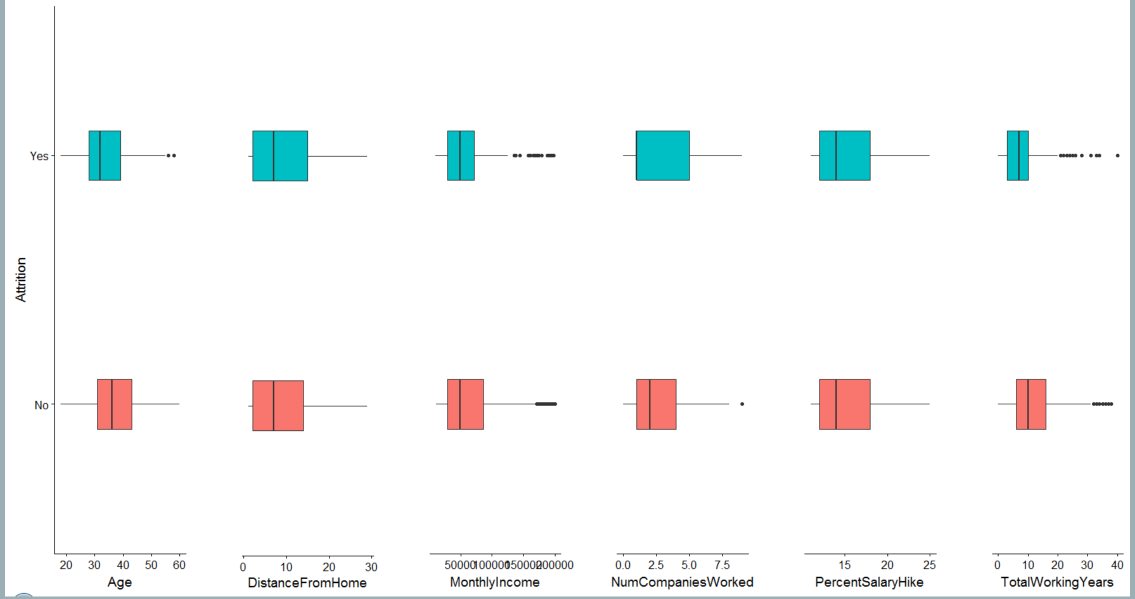
- Analysis on following set of categorical variables (EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance, JobInvolvement, PerformanceRating) indicates that,
  - Employees with Low EnvironmentSatisfaction have high attrition rate compare to other levels, followed by employees with low job satisfaction.



Data Analysis - Continuous Variables

- Analysis on continuous variables, indicates that,
  - NumCompaniesWorked and AvgWorkHrs variables have more effect on attrition rate.
  - Outliers noticed in following variables
    - MonthlyIncome
    - NumCompaniesWorked
    - TotalWorkingYears
    - YearsAtCompany
    - YearsSinceLastPromotion
    - YearsWithCurrManager
    - AvgWorkHrs

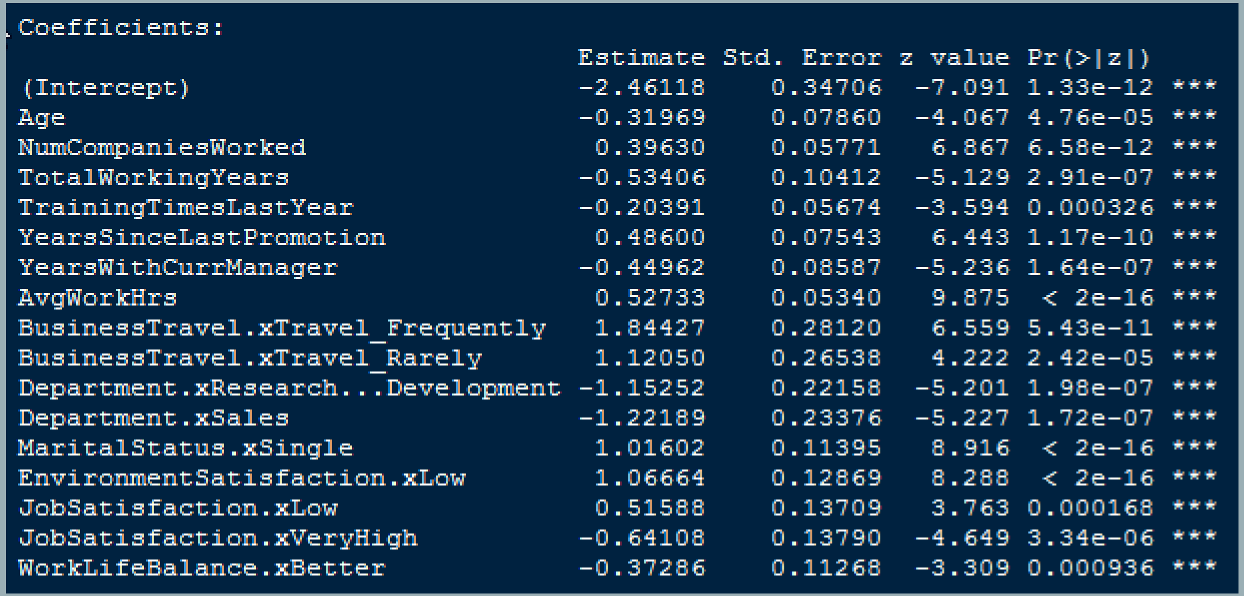




Model Building - Logistic Regression

Based on logistic regression model analysis, following variables identified as significant predictors to calculate the probability of the attrition rate:

- Age
- Number of Companies Worked
- Total Working Years
- Training Times Last Year
- Years Since Last Promotion
- Years With Current Manager
- Average Working Hours
- Business Travel (Travel Frequently, Travel Rarely)
- Department (Research & Development, Sales)
- Marital Status (Single)
- Environment Satisfaction (Low)
- Job Satisfaction (Low, Very High)
- Work Life Balance (Better)



```
Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                        -2.46118    0.34706  -7.091 1.33e-12 ***
Age                                -0.31969    0.07860  -4.067 4.76e-05 ***
NumCompaniesWorked                  0.39630    0.05771   6.867 6.58e-12 ***
TotalWorkingYears                  -0.53406    0.10412  -5.129 2.91e-07 ***
TrainingTimesLastYear              -0.20391    0.05674  -3.594 0.000326 ***
YearsSinceLastPromotion             0.48600    0.07543   6.443 1.17e-10 ***
YearsWithCurrManager               -0.44962    0.08587  -5.236 1.64e-07 ***
AvgWorkHrs                          0.52733    0.05340   9.875  < 2e-16 ***
BusinessTravel.xTravel_Frequently   1.84427    0.28120   6.559 5.43e-11 ***
BusinessTravel.xTravel_Rarely       1.12050    0.26538   4.222 2.42e-05 ***
Department.xResearch...Development -1.15252    0.22158  -5.201 1.98e-07 ***
Department.xSales                  -1.22189    0.23376  -5.227 1.72e-07 ***
MaritalStatus.xSingle               1.01602    0.11395   8.916  < 2e-16 ***
EnvironmentSatisfaction.xLow        1.06664    0.12869   8.288  < 2e-16 ***
JobSatisfaction.xLow                0.51588    0.13709   3.763 0.000168 ***
JobSatisfaction.xVeryHigh          -0.64108    0.13790  -4.649 3.34e-06 ***
WorkLifeBalance.xBetter            -0.37286    0.11268  -3.309 0.000936 ***
```

Model Evaluation - Cut Off 50%

Probability cut off at 50%:

- Accuracy of the model - 86%
- Sensitivity (True Positive Rate) - 25%
- Specificity (True Negative Rate) - 98%

Analysis:

- Even though the accuracy of the model is high, the sensitivity of the model is very low. Since management wants to identify attrition rate, we need to maximize the sensitivity of the model.

		Predicted Attrition	
		No	Yes
Actual Attrition	No	1080	30
	Yes	159	54

Model Evaluation - Cut Off 40%

Probability cut off at 40%:

- Accuracy of the model - 85%
- Sensitivity (True Positive Rate) - 33%
- Specificity (True Negative Rate) - 95%

Analysis:

- Even though the sensitivity (TPR) value is increased with cut off value at 40%, the sensitivity of the model is still low.

		Predicted Attrition	
		No	Yes
Actual Attrition	No	1052	58
	Yes	143	70

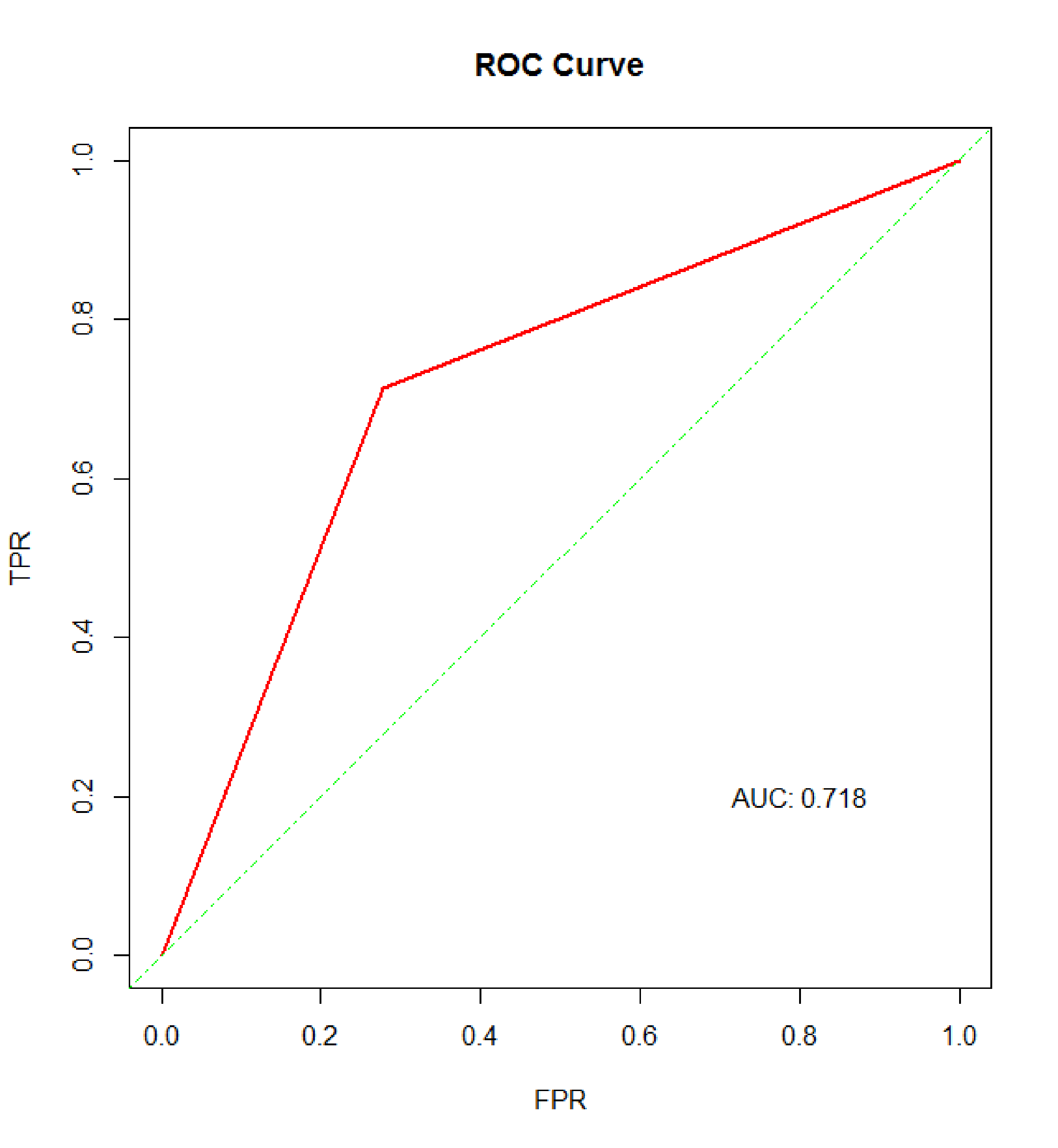
Model Evaluation - Optimal Cut Off

- Based on analysis the optimal cut off value calculated as 16.16%.
- At optimal cut off value:
  - Accuracy of the model - 72%
  - Sensitivity (True Positive Rate) - 71%
  - Specificity (True Negative Rate) - 72%
- Analysis:
  - With slight decline in accuracy, we were able to achieve high sensitivity value of 71%.

		Predicted Attrition	
Actual Attrition		No	Yes
	No	801	309
	Yes	61	152

Model Evaluation - KS Statistic

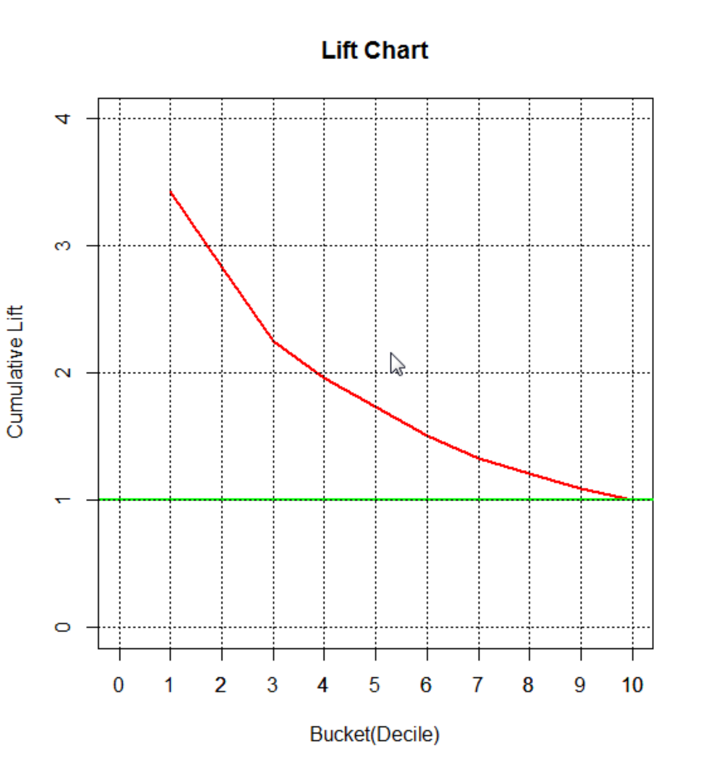
- KS Test measures to check whether model is able to separate events and non-events. In probability of our Attrition model, it checks whether the our model is able to distinguish between employees who will leave and employee who will not leave.
- Ideally, the KS score lies between 40 and 70. In this case, KS score > 40 (i.e. 43%), which is good model.

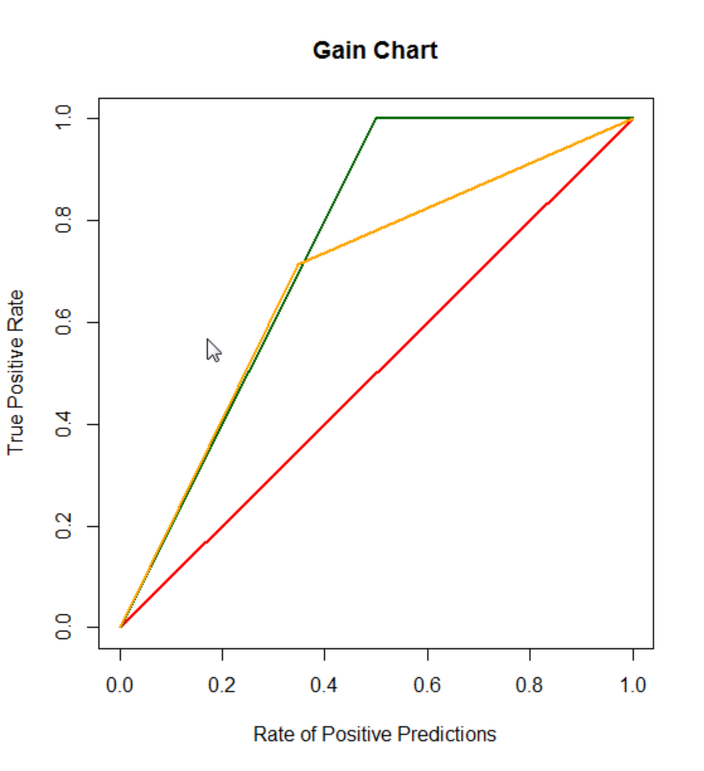


Model Evaluation - Lift and Gain Charts

- Gain Chart: Based on gain chart, we identified our model is near to the perfect model, which is good.

- Lift Chart: Based on lift chart, we identified our model is outperforming a random model. We can predict true positive rate more efficiently using our model compared to a random model.





Conclusion

- Logistic regression model analysis identified following critical factors related to employees, which will help the management to make changes in their workplace to reduce the attrition rate.
  - Age
  - Number of Companies Worked
  - Total Working Years
  - Training Times Last Year
  - Years Since Last Promotion
  - Years With Current Manager
  - Average Working Hours
  - Business Travel
  - Department
  - Marital Status
  - Environment Satisfaction
  - Job Satisfaction
  - Work Life Balance