HR Analytics Case Study

Group Members

- Kumar Abhilash
- Barsha Guha
- Jaya Rathina
- Rakhee Roy

Business Objectives

Business Objective

- XYZ is a large company having 4000+ employees.
- Every year around 15% of employees leave the company.
- The attrition of employees is bad for the company due to following reasons:
    - The former employees' projects get delayed, which makes it difficult to meet **timelines**, resulting in a reputation loss among consumers and partners.
    - A sizeable department has to be maintained, for the purposes of **recruiting** new talent.
    - The new employees have to be **trained** for the job and/or given time to acclimatize themselves to the company.
- To model the probability of attrition using a logistic regression. The results thus 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.

Approach

- Business understanding and objective.
- Data collection and understanding:
  - Employee survey data : Data collected from employees (contains 4 variables)
  - Manager survey data : Data collected from managers (contains 3 variables)
  - General data : General : Data about each employee (contains 24 variables)
  - In time and out time : In and out time of each employee for the year 2015
- Data cleaning and EDA.
- Identifying categorical and continuous variables.
- Scaling and creating dummy variables.
- Creating train and test data after splitting the master dataset.
- Model building to find out most significant variables and model evaluation to predict probability of attrition.

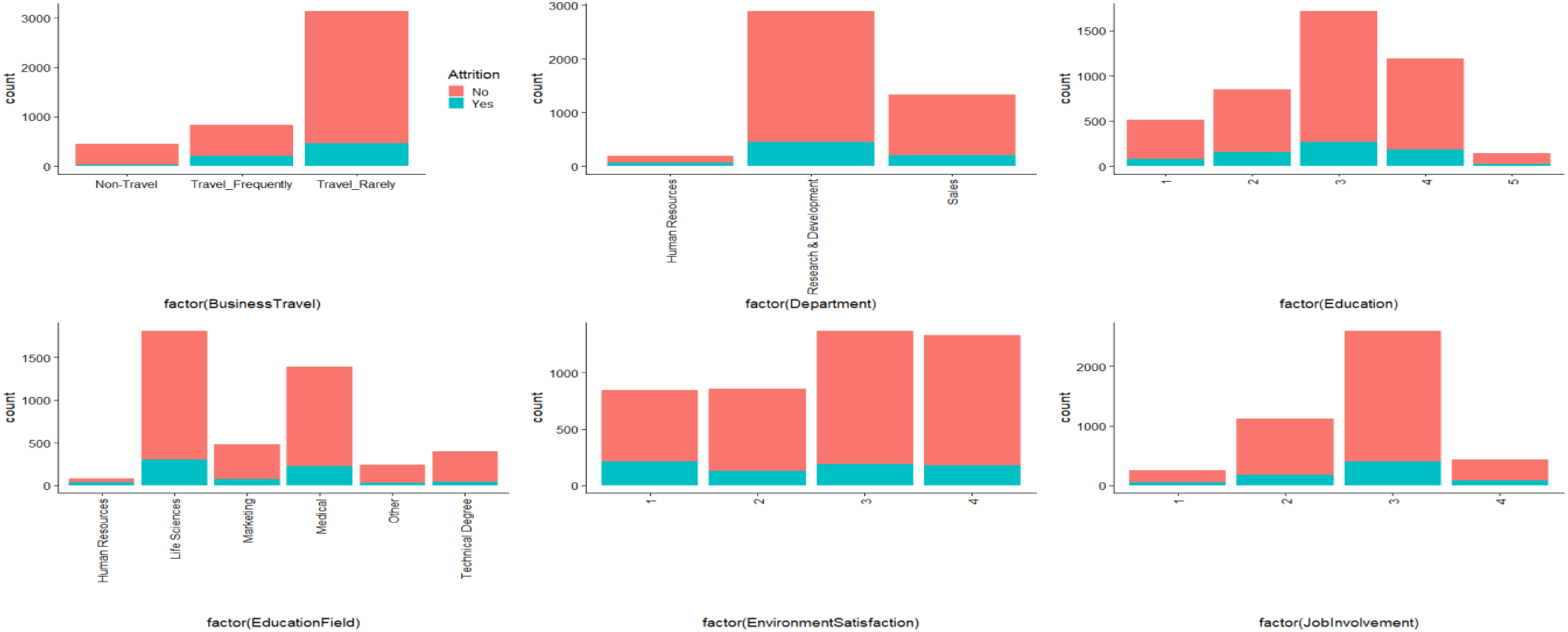
Data cleaning and derived metrics

- Converted in and out_time variables into standard date-time format
- Remove all variables that only contain one value or two values.
- Columns having NA values throughout the column for in and out_time have been removed
- Renamed the missing column name for the given dataset
- Calculated average time spent for each employee and derived a metric of overtime
- For Categorical variables having NA values, were replaced with median values.
- The two derived metrics
- Time_spent: Total time spent by an employee
- Average_Time: Average time spent by the employee for the given time period
- Over_Time: Time exceeding normal working hours. The threshold value considered=8.5 hours

Exploratory Data Analysis

- Converted below categorical variables with 2 level into 0 and 1:
  - Attrition and gender
- Created dummy variables for the following columns as they have more than two level:
  - "BusinessTravel","Department","Education","EducationField","EnvironmentSatisfaction","JobInvolvement", "JobLevel","JobRole","JobSatisfaction","MaritalStatus","PerformanceRating","WorkLifeBalance"
- Scaling of below continuous variables:
  - "Age",DistanceFromHome","MonthlyIncome","NumCompaniesWorked","PercentSalaryHike","StockOptionLevel","TotalWorkingYears", "TrainingTimesLastYear", "YearsAtCompany","YearsSinceLastPromotion","YearsWithCurrManager","Avg_TimeSpent"
- Relationship between attrition variable and independent variables with the help of univariate analysis.
- Check for outliers in continuous variables using boxplot and analyzed the quantile distribution.

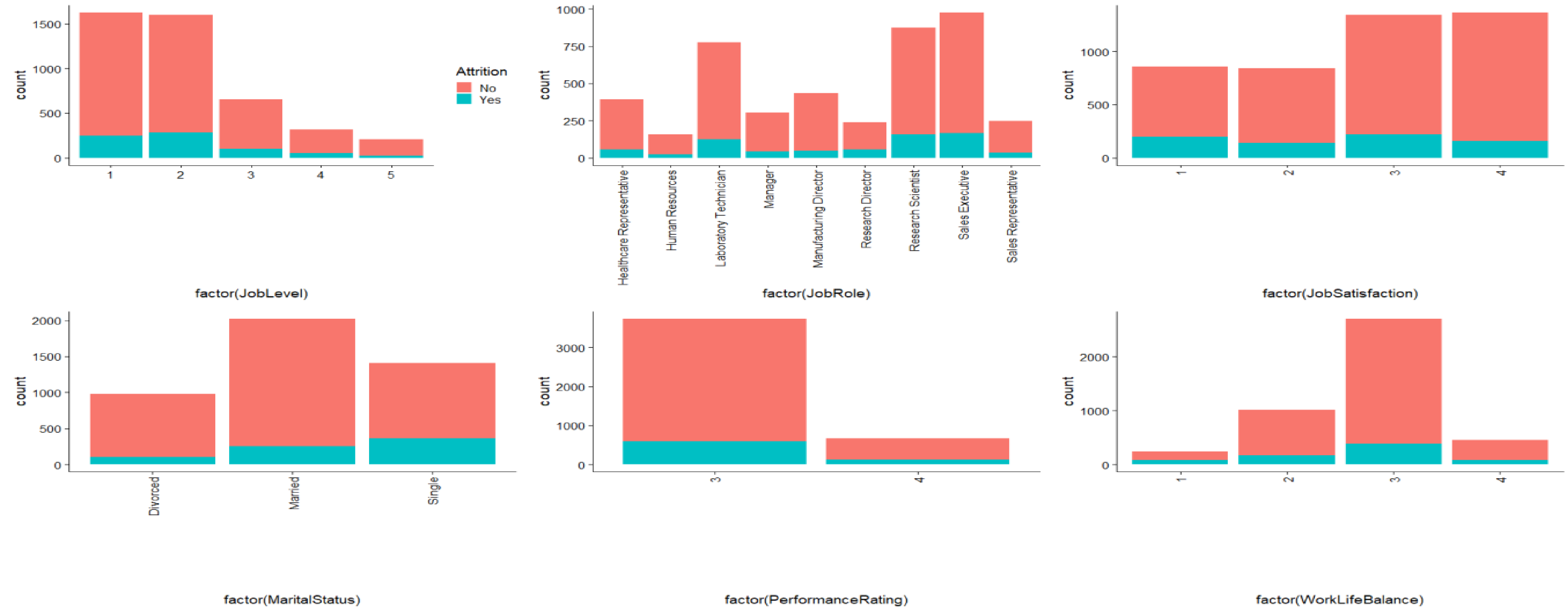
Uni-variate Analysis Plot-1 Categorical



High attrition can be seen for:

- Employees who travel rarely
- R&D department
- Life science and medical education field

Uni-variate AnalysisPlot-2 Categorical

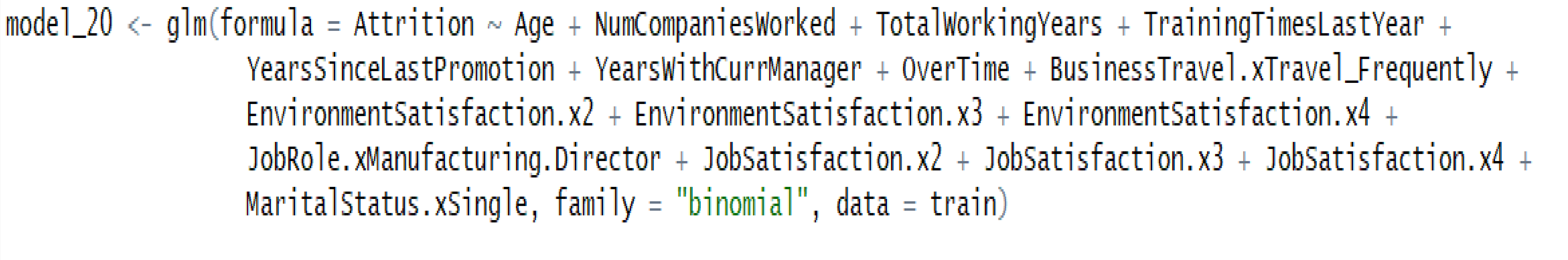


High attrition can be seen for:

- Employees with less experience
- Employees having job role as research scientist and sales executive

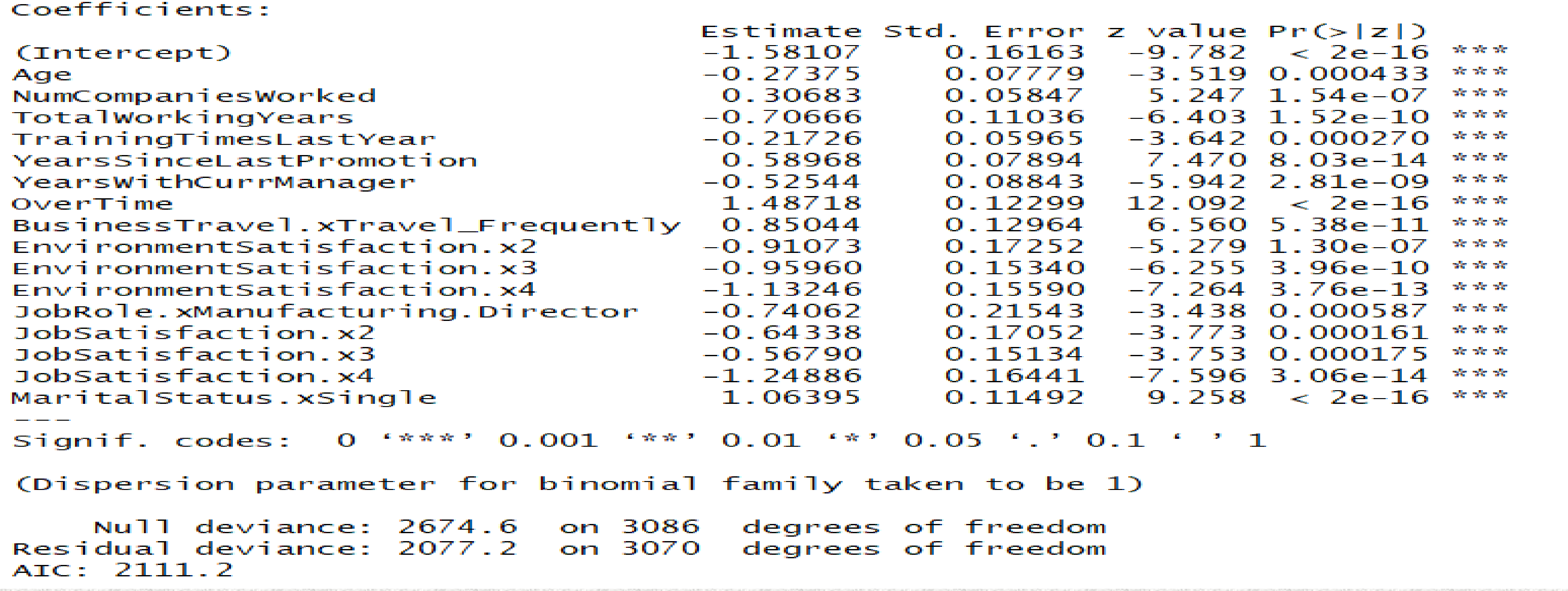
Model building and evaluation

- Split the data into train(70%) and test(30%) data set.
- Created generalized linear models and iterated based on significant p values and VIF.
- Evaluated the model to achieve the final model having all the significant variables.



```
model_20 <- glm(formula = Attrition ~ Age + NumCompaniesWorked + TotalWorkingYears + TrainingTimesLastYear +
                YearsSinceLastPromotion + YearsWithCurrManager + OverTime + BusinessTravel.xTravel_Frequently +
                EnvironmentSatisfaction.x2 + EnvironmentSatisfaction.x3 + EnvironmentSatisfaction.x4 +
                JobRole.xManufacturing.Director + JobSatisfaction.x2 + JobSatisfaction.x3 + JobSatisfaction.x4 +
                MaritalStatus.xSingle, family = "binomial", data = train)
```

Significant variables

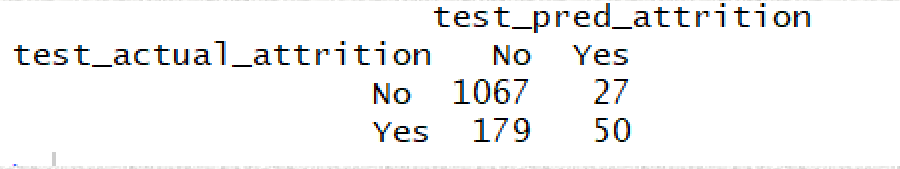


```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     -1.58107    0.16163  -9.782  < 2e-16 ***
Age                             -0.27375    0.07779  -3.519 0.000433 ***
NumCompaniesWorked               0.30683    0.05847   5.247 1.54e-07 ***
TotalWorkingYears               -0.70666    0.11036  -6.403 1.52e-10 ***
TrainingTimesLastYear           -0.21726    0.05965  -3.642 0.000270 ***
YearsSinceLastPromotion          0.58968    0.07894   7.470 8.03e-14 ***
YearsWithCurrManager            -0.52544    0.08843  -5.942 2.81e-09 ***
OverTime                         1.48718    0.12299  12.092  < 2e-16 ***
BusinessTravel.xTravel_Frequently  0.85044    0.12964   6.560 5.38e-11 ***
EnvironmentSatisfaction.x2      -0.91073    0.17252  -5.279 1.30e-07 ***
EnvironmentSatisfaction.x3      -0.95960    0.15340  -6.255 3.96e-10 ***
EnvironmentSatisfaction.x4      -1.13246    0.15590  -7.264 3.76e-13 ***
JobRole.xManufacturing.Director -0.74062    0.21543  -3.438 0.000587 ***
JobSatisfaction.x2              -0.64338    0.17052  -3.773 0.000161 ***
JobSatisfaction.x3              -0.56790    0.15134  -3.753 0.000175 ***
JobSatisfaction.x4              -1.24886    0.16441  -7.596 3.06e-14 ***
MaritalStatus.xSingle            1.06395    0.11492   9.258  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2674.6  on 3086  degrees of freedom
Residual deviance: 2077.2  on 3070  degrees of freedom
AIC: 2111.2
```

These are the contributing variables affecting the attrition rate of the employees obtained from final model evaluation.

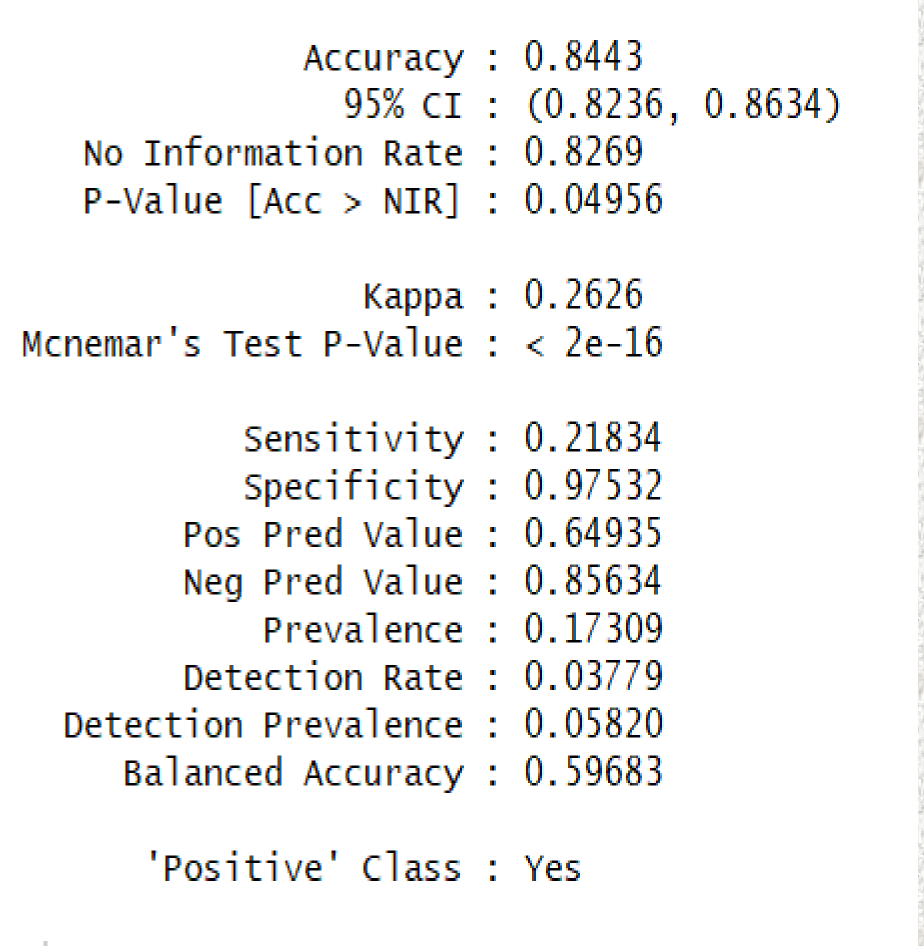
Model evaluation

- Predicted the model for test data with probability of attrition of cutoff value as 0.5.



```
                     test_pred_attrition
test_actual_attrition   No  Yes
                  No  1067   27
                  Yes  179   50
```

- Confusion matrix generated for cutoff value 0.5.
- Observations:
  - Specificity : 0.97
  - Sensitivity : 0.22
  - Accuracy : 0.84
  - The specificity is considerably higher compared to sensitivity.



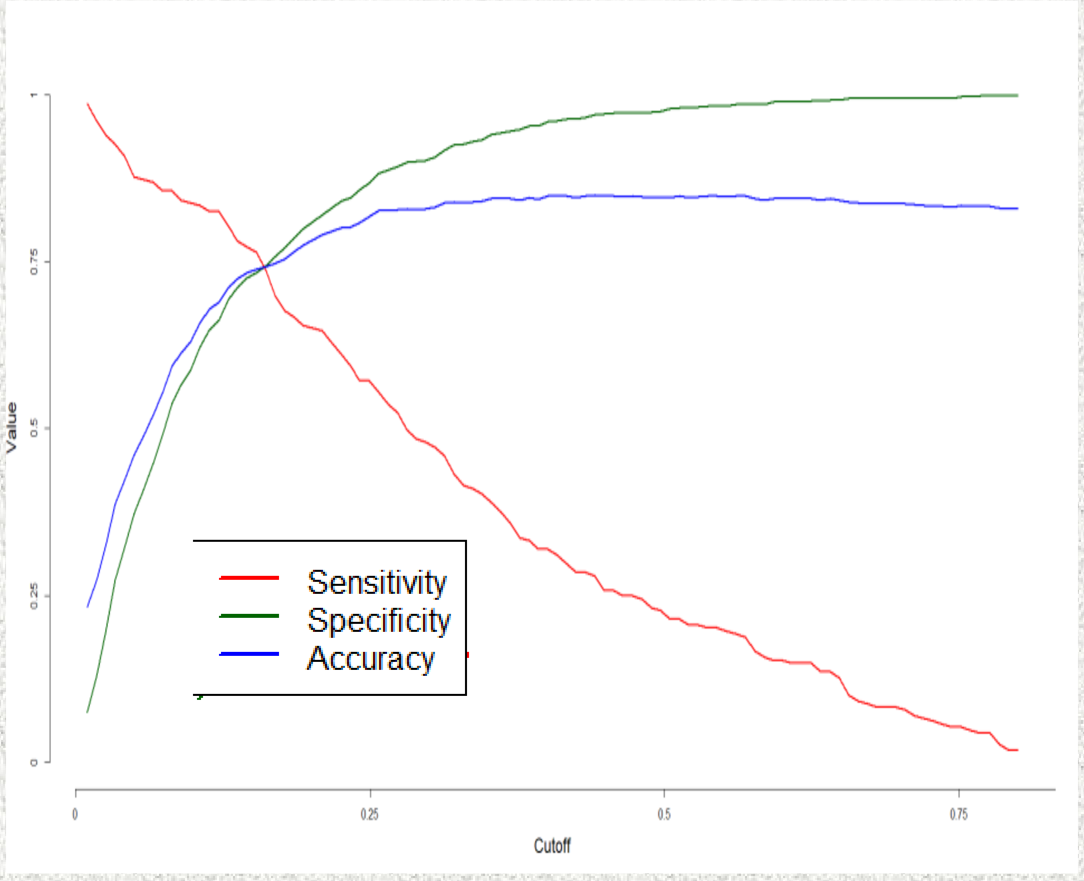
```
               Accuracy : 0.8443
                 95% CI : (0.8236, 0.8634)
    No Information Rate : 0.8269
    P-Value [Acc > NIR] : 0.04956

                  Kappa : 0.2626
 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.21834
            Specificity : 0.97532
         Pos Pred Value : 0.64935
         Neg Pred Value : 0.85634
             Prevalence : 0.17309
         Detection Rate : 0.03779
   Detection Prevalence : 0.05820
      Balanced Accuracy : 0.59683

       'Positive' Class : Yes
```

Model evaluation contd..

- To obtain the optimal cutoff value, we plotted the accuracy, sensitivity and specificity vs cutoff.
- The intersection of the three given attributes gives us the optimal cutoff value.

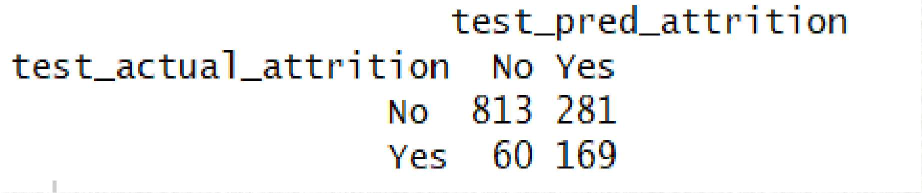
Observations:

- As observed from the graph the cutoff values is less than 0.25.
- After few iterations, 0.16 was considered as the optimal cutoff value.



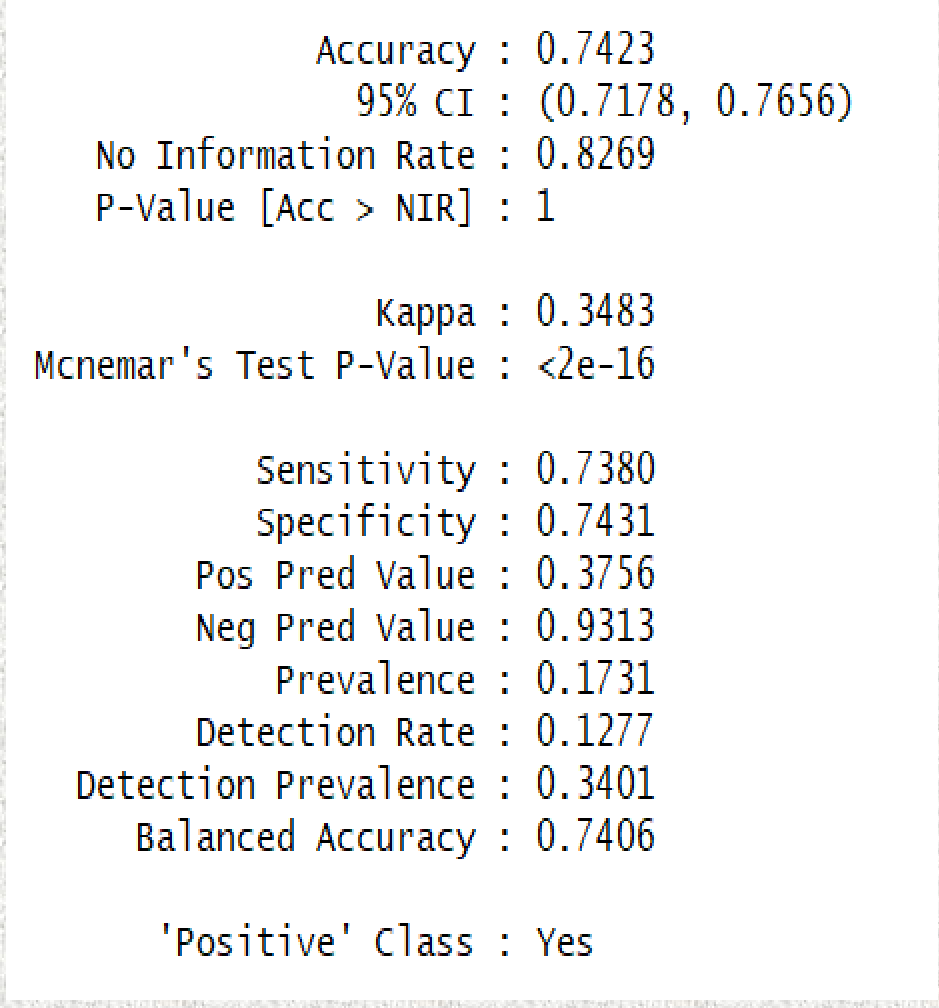
Model evaluation contd..

- Predicted the model for test data with probability of attrition of cutoff value as 0.16.



```
                     test_pred_attrition
test_actual_attrition  No Yes
                  No  813 281
                  Yes  60 169
```

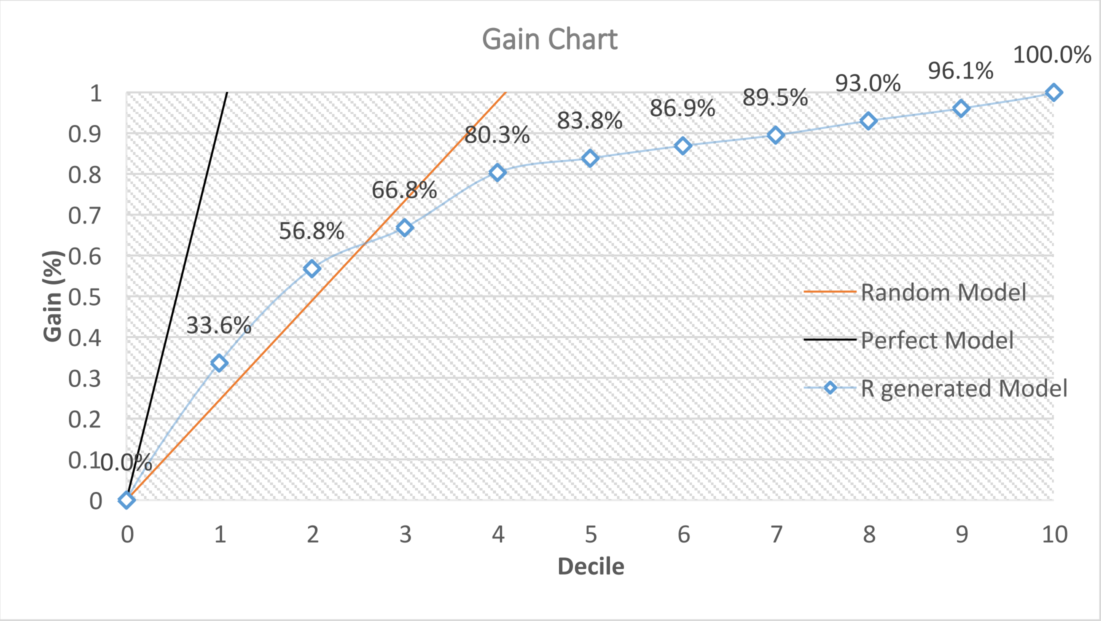
- Confusion matrix generated for cutoff value 0.16.
- Observations:
  - Specificity : 0.74
  - Sensitivity : 0.73
  - Accuracy : 0.74
  - The specificity is now comparable to sensitivity.



```
               Accuracy : 0.7423
                 95% CI : (0.7178, 0.7656)
    No Information Rate : 0.8269
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3483
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7380
            Specificity : 0.7431
         Pos Pred Value : 0.3756
         Neg Pred Value : 0.9313
             Prevalence : 0.1731
         Detection Rate : 0.1277
   Detection Prevalence : 0.3401
      Balanced Accuracy : 0.7406

       'Positive' Class : Yes
```

Model Assessment-Gain Chart

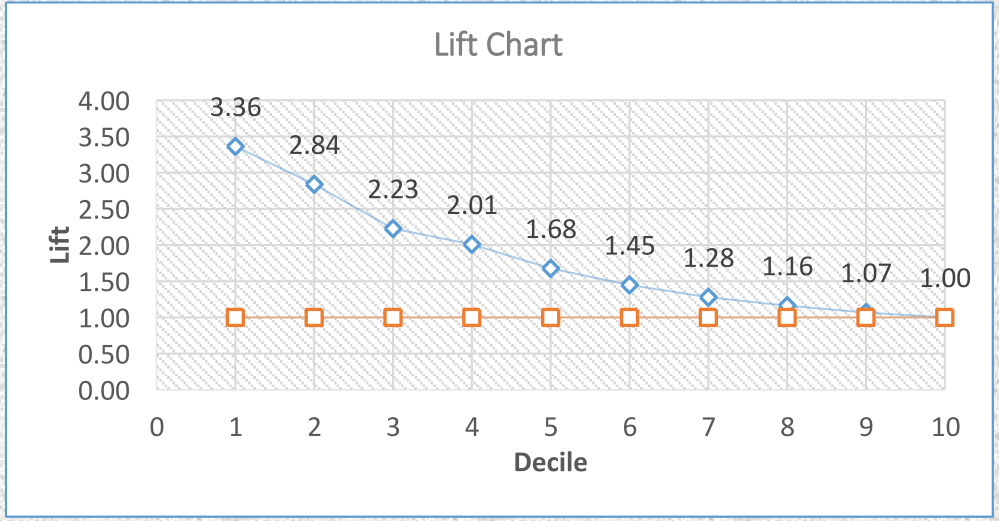
Gain Chart				
Decile	Observations	Attrition	Cum- Attrition	Gain(%Cum-Attrition)
0	0	0	0	0
1	133	77	77	33.6%
2	132	53	130	56.8%
3	132	23	153	66.8%
4	132	31	184	80.3%
5	132	8	192	83.8%
6	132	7	199	86.9%
7	132	6	205	89.5%
8	132	8	213	93.0%
9	132	7	220	96.1%
10	132	9	229	100.0%
Total	1321	229		



Model Assessment-Lift Chart

Lift Chart

Decile	Observations	Attrition	Cum-Attrition	Gain(%Cum-Attrition)	Gain (Random Model)	Lift
1	133	77	77	33.6%	10%	3.36
2	132	53	130	56.8%	20%	2.84
3	132	23	153	66.8%	30%	2.23
4	132	31	184	80.3%	40%	2.01
5	132	8	192	83.8%	50%	1.68
6	132	7	199	86.9%	60%	1.45
7	132	6	205	89.5%	70%	1.28
8	132	8	213	93.0%	80%	1.16
9	132	7	220	96.1%	90%	1.07
10	132	9	229	100.0%	100%	1.00
Total	1321	229				



Model Conclusion and Recommendations

- The model has an increasing Gain and a decreasing Lift.
- The Model predicts more than 80% of the attritions within the 4th Decile with 74% accuracy.
- Below are the listed factors which causes employee to leave the current company and causes employee attrition.
- People switching their jobs frequently. If the number of companies worked is higher the employee is likely to leave the company.
- Years since last promotion: Employees having a considerable gap in their last promotion have higher chances to leave the company.
- Overtime: Employees spending more than the usual working hours are likely to leave the company.
- Marital Status: Singles have higher attrition rate.