HR Analytics Case Study

Submitted by:

Group Name: Team Insight

1. Kaushal Kashyap
2. Manjiri Paranjape
3. Snehil Gupta
4. Ravinder Gill

Business Understanding

- A large company named **XYZ**, employs, around 4000 employees.
- Around 15% of its employees leave the company every year and need to be replaced.
- This level attrition is rising every year.
- According to management this is not a favorable condition and impacts the business deliverables.

Problem and Goal Analysis

Problem:

The business is facing below mentioned problems due to the attrition.

1. **Delayed timelines** → Project time line gets delayed, impacts good will of the company among clients.
2. **New Recruitment** → To maintain team size company needs to recruit new talent
3. **Job Training** → Training of newly recruited employee consumes time and resources.

Goals:

1. Understand the main factors that are responsible for this attrition.
2. Improvements required at the workplace to curb the attrition.
3. Factors that can improve the employee satisfaction.
4. Determine the most important variables that need to be addressed right away.

Data Understanding

Data Source: Data is provided in ***.csv file*** format in the form of five files:

1. employee_survey_data → Data collected over employees through a survey across the company
2. manager_survey_data → Data collected over managers through survey across the company
3. in_time → Employee's reaching time to their company, source may be log registers or biometric machine or card scan sensor.
4. out_time → Employee's leaving time from the company, source may be log registers or biometric machine or card scan sensor.
5. general_data → Data containing employee's personal details like age, years of experience, education, gender, department, business travel etc.

Data Understanding:

1. The data contains the information about surveys conducted on employs, manager along with general information pertaining to the employs and office timing hours and the holidays taken by the employs over the period of past one year.
2. Data in in_time and out_time files contain the timing on weekdays only and for public holidays the value is replaced with NA.
3. Survey data file shows the rating values given by employees to the various factors like Job satisfaction, work life balance etc.

Methodology

Analysis Type	Operations Performed	Methodology/ Tools used
Data Cleaning	• Data Preparation: This includes Data Preparation, Cleansing and Formatting. Getting data ready for Logistic Regression. • Remove columns that are not required.	▪ R - studio
EDA	• Visually identify effect of various factors on the employee attrition. • Plot bar charts, histograms, box plots to visually identify any trend. • Plot correlation matrix to identify the relation ship between two variables. • Outlier treatment.	▪ Univariate analysis, ▪ Bi Variate analusis, ▪ Segmented analysis. ▪ Derived Matrices.
Modelling	• Develop a regression model for Attrition probability • Split the data in to training and test data. • Train the data with training data.	▪ Model based on logistic regression. ▪ Step AIC ▪ Variable Inflation factor
Model Evaluation	• Test the model or predict the attrition probability using the testing data • Compare the results from testing data i.e. predicted results with actual values. • Optimize the model	▪ Confusion Matrix ▪ Find the Sensitivity and Specificity ▪ Gain and Lift Charts ▪ KS Statistics

Data Preparation and Cleaning

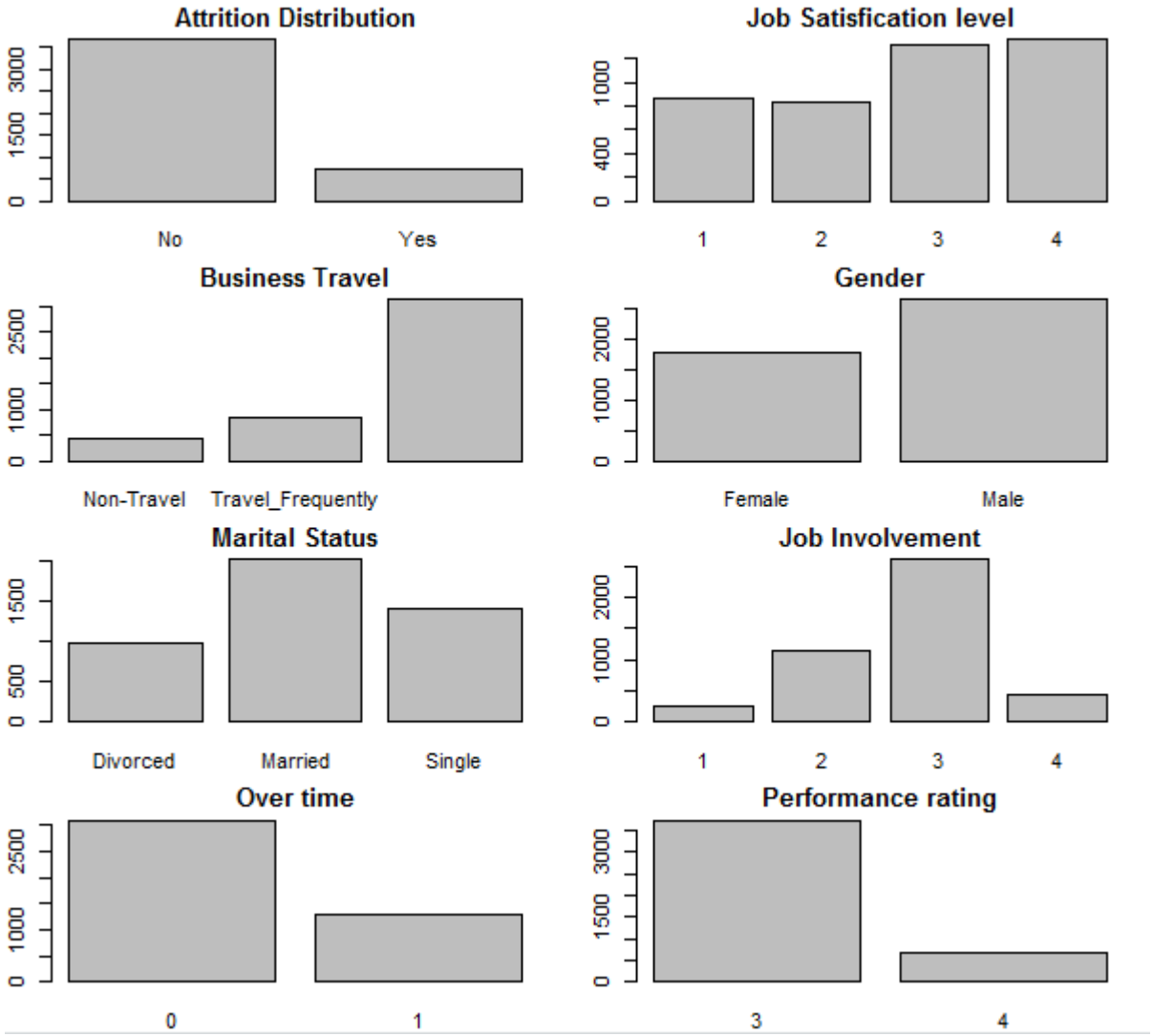
For Data cleaning and preparation following steps were performed in sequence.

1. First remove the public holiday columns i.e. remove the columns with all NA values from **in_time** and **out_time**
2. From **in_time** and **out_time** data calculated the mean working hour of each employee.
3. Check all the sheets for duplicate values and merge all the files in to one consolidated data sheet
4. Derive new column metrics like Over time, Inadequate time, No of leaves taken by each employee.
5. Treating the NA values replace them with mean or median values.
6. This data is ready for Exploratory Data Analysis.

Tools Used:

- R - language

EDA – Data Distribution

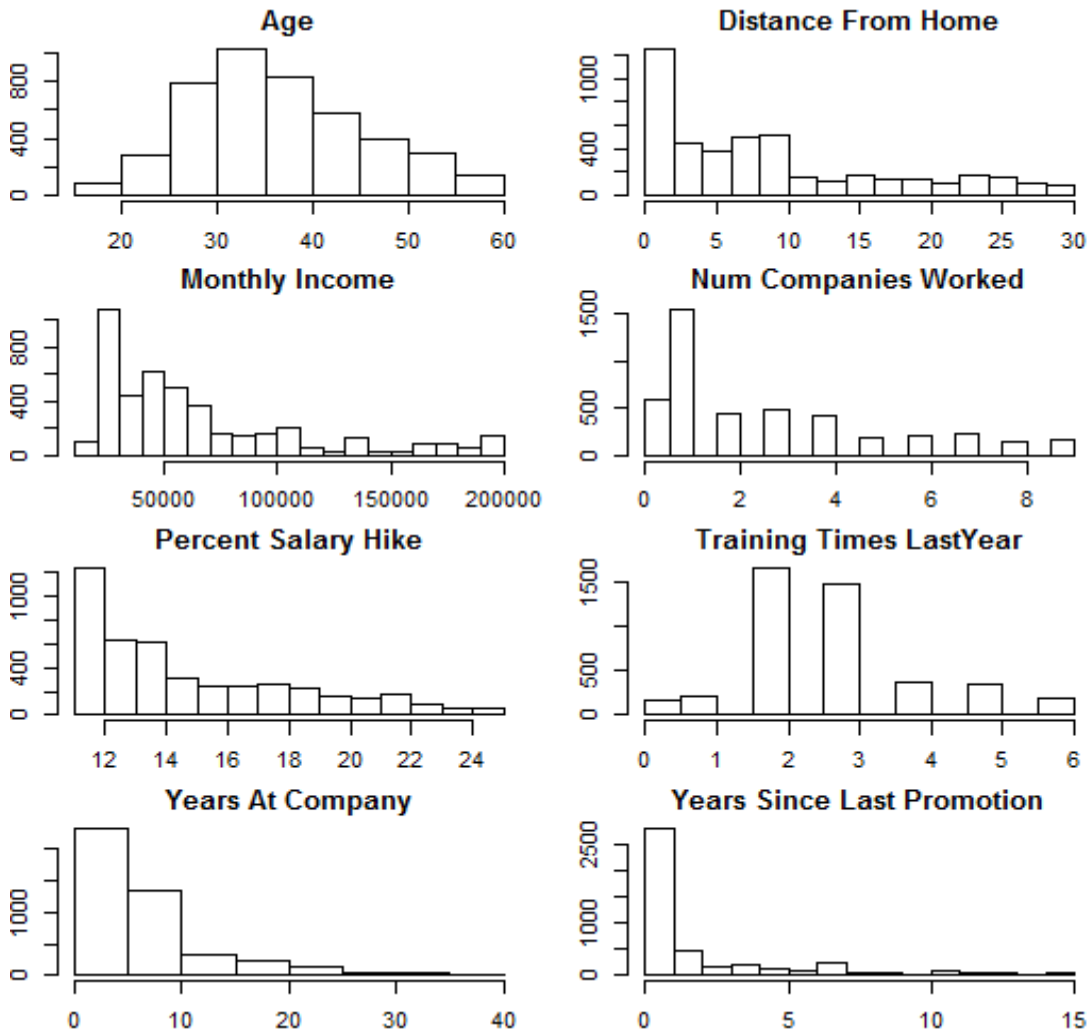


Plotting the Categorical variable:

Bar graphs describes the distribution of various categorical variables like:

- Attrition
- Job Satisfaction
- Business Travel
- Gender
- Marital Status
- Job Involvement
- Over time
- Performance Rating

EDA – Data Distribution

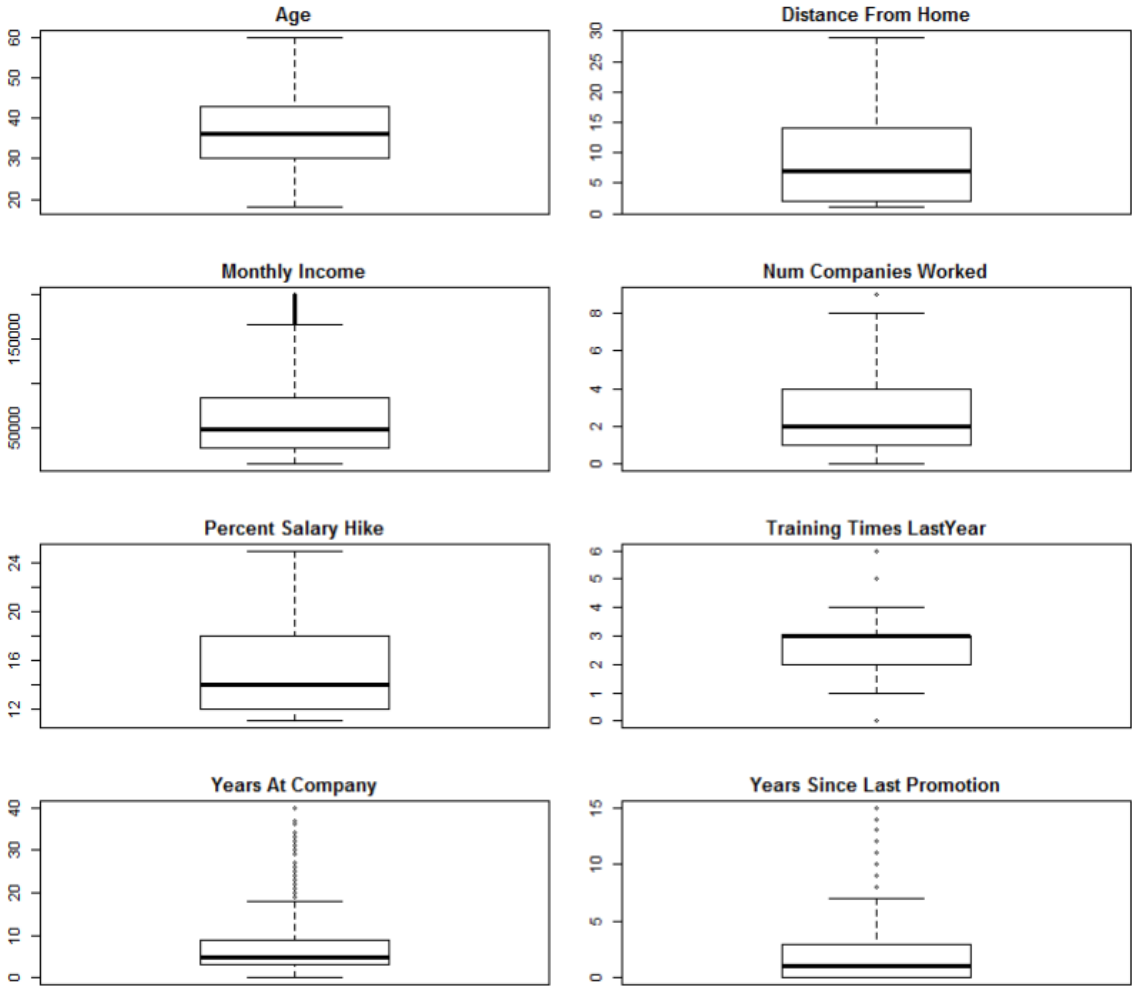


- Plotting the continuous variable

Histograms describe the distribution of numerical data like:

- Age
- Monthly Income
- Distance From Home
- Percentage Salary Hike
- Training Time Last Year
- Years at company
- Years since last promotion
- Number of companies worked

EDA – Data Distribution



- Box plots of various parameters like
  1. monthly income.
  2. Years at company.
  3. Training time last Year.
  4. Years since last promotion.

shows how data is distributed about the mean along with quartiles and presence of outliers.

- But it is not good to remove outliers as they represent the company population where such data is bound to exist

EDA – Correlation Matrix

Correlation Matrix between numerical variable shows the relationship between various numerical variable.

- Years at Company shows strong correlation with years with current manager.
- Years at Company also shows good correlation with years since last promotion.
- Similarly age shows string correlation with total working years.
- Rest correlation are not that significant.

Modelling

1. Continuous data is standardized using scale method to remove the influence of variable with high values.
2. For categorical variable dummy variables are created as per requirement.
3. Data is split into two categories training data and test data to train and test the model.
4. Attrition Probability Model is developed based on logistic regression using the training data.
5. Logistic regression model was built in R using the function glm().
6. Model is optimized using these two algorithms:
    a) Stepwise variable selection based on AIC [using stepAIC()]
    b) Backward variable selection based on VIF and p-value.
7. Iteratively eliminate the variables based up on level of significance.
8. After the whole process we get the final Logistic Regression Model ready for evaluation.

Modelling

1. After 18 iteration, the final Logistic Regression model is obtained with 17 variables
2. There variables can be considered to have influential effect on employee attrition

```
model_18<-glm(formula = Attrition ~ NumCompaniesWorked + TotalWorkingYears +
                YearsSinceLastPromotion + YearsWithCurrManager + Over_time +
                EnvironmentSatisfaction.x2 + EnvironmentSatisfaction.x3 +
                EnvironmentSatisfaction.x4 + JobSatisfaction.x2 + JobSatisfaction.x3 +
                JobSatisfaction.x4 + WorkLifeBalance.x2 + WorkLifeBalance.x3 +
                WorkLifeBalance.x4 + BusinessTravel.xTravel_Frequently +
                JobRole.xManufacturing.Director + MaritalStatus.xSingle,
              family = "binomial", data = train)
```

Model Evaluation

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7927  -0.5381  -0.3317  -0.1573   4.0037

Coefficients:
                                Estimate Std. Error z value             Pr(>|z|)
(Intercept)                     -0.67678    0.24354  -2.779             0.005453 **
NumCompaniesWorked               0.25504    0.05833   4.373   0.0000122735444965 ***
TotalWorkingYears               -0.90646    0.09655  -9.388 < 0.0000000000000002 ***
YearsSinceLastPromotion          0.59687    0.08022   7.440   0.0000000000001004 ***
YearsWithCurrManager            -0.53262    0.09026  -5.901   0.0000000036179247 ***
Over_time                        1.59987    0.11904  13.440 < 0.0000000000000002 ***
EnvironmentSatisfaction.x2      -0.88407    0.17545  -5.039   0.0000004684694500 ***
EnvironmentSatisfaction.x3      -0.91197    0.15613  -5.841   0.0000000051821291 ***
EnvironmentSatisfaction.x4      -1.12074    0.15853  -7.070   0.0000000000015540 ***
JobSatisfaction.x2              -0.66964    0.17341  -3.862             0.000113 ***
JobSatisfaction.x3              -0.58138    0.15285  -3.804             0.000143 ***
JobSatisfaction.x4              -1.29875    0.16804  -7.729   0.0000000000000109 ***
WorkLifeBalance.x2              -0.95343    0.22611  -4.217   0.0000247923443573 ***
WorkLifeBalance.x3              -1.25399    0.20944  -5.987   0.0000000021324226 ***
WorkLifeBalance.x4              -1.07565    0.26450  -4.067   0.0000476646677604 ***
BusinessTravel.xTravel_Frequently  0.83692  0.13179   6.350   0.0000000002148683 ***
JobRole.xManufacturing.Director -0.71138    0.21597  -3.294             0.000988 ***
MaritalStatus.xSingle            1.09150    0.11644   9.374 < 0.0000000000000002 ***
---
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: 2027.4  on 3069  degrees of freedom
AIC: 2063.4

Number of Fisher Scoring iterations: 6
```

1. This is the final Logistic Regression Model.
2. All the variables in the model has low VIF and low p values.
3. The final model is used to make predictions using the test data and the results are saved as **predicted values**
4. Since we already know the results of testing data which are called **actual values**.
5. Comparison between **actual values** and **predicted values** tells the accuracy of the model.
6. This comparison is depicted if the tabular form called **Confusion Matrix.**

Model Accuracy – Confusion Matrix

- Confusion Matrix at cut off >= 0.4 i.e. 40%

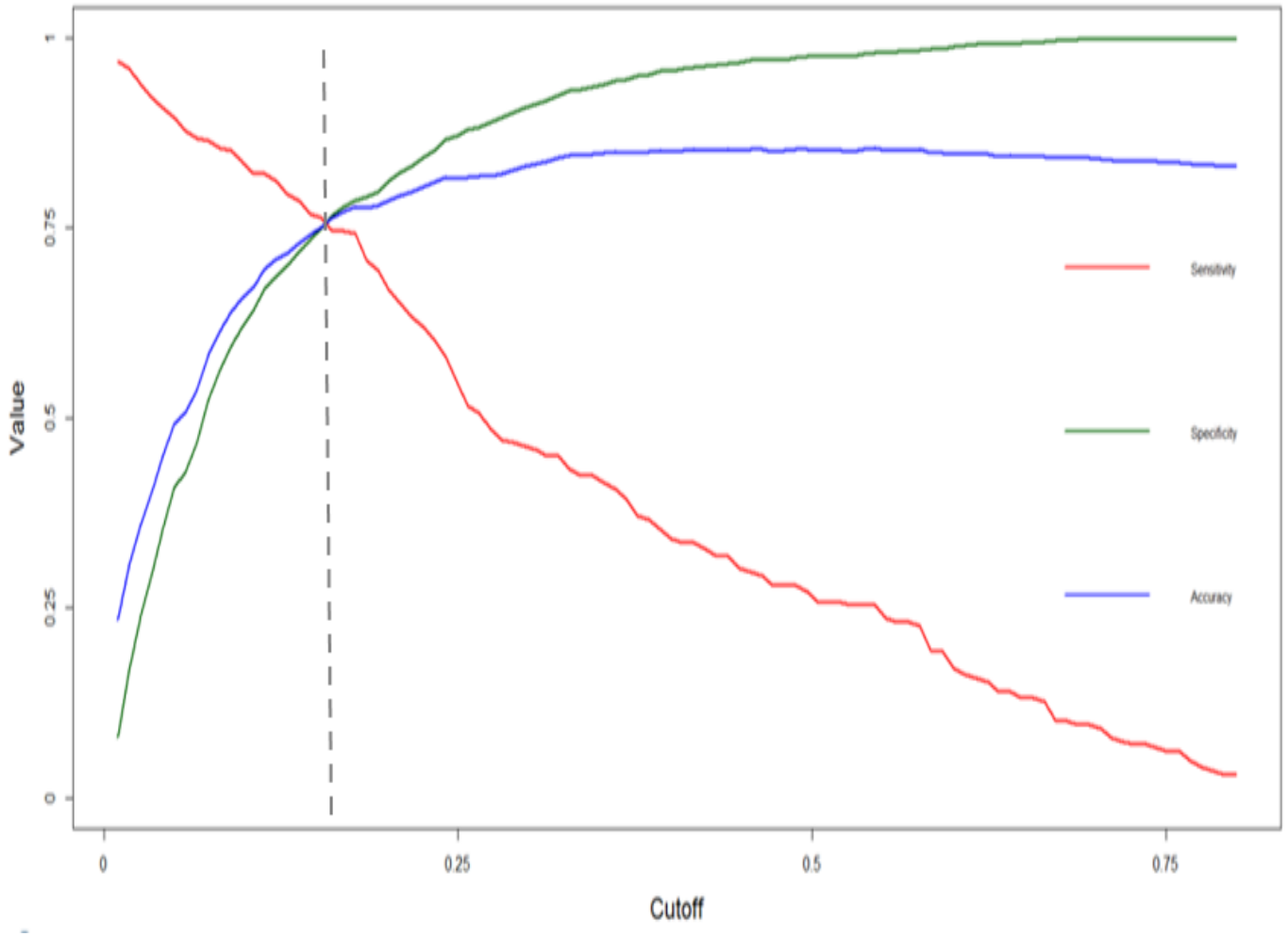
```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1047  151
       Yes   47   78

               Accuracy : 0.8503
                 95% CI : (0.83, 0.8691)
    No Information Rate : 0.8269
    P-Value [Acc > NIR] : 0.01223

                  Kappa : 0.3628
 Mcnemar's Test P-Value : 2.482e-13

            Sensitivity : 0.34061
            Specificity : 0.95704
         Pos Pred Value : 0.62400
         Neg Pred Value : 0.87396
             Prevalence : 0.17309
         Detection Rate : 0.05896
   Detection Prevalence : 0.09448
      Balanced Accuracy : 0.64882

       'Positive' Class : Yes
```

	Predicted	
Actual	No (no attrition)	Yes (attrition)
No (no attrition)	1047	151
Yes (attrition)	47	78

Metrics	
Accuracy	85%
Sensitivity	0.34
Specificity	0.95

Model is fairly accurate but sensitivity is low at cut off >= 0.4 or 40%

Model Accuracy – Sensitivity and Specificity



Optimum Cut off:

- Sensitivity and specificity indicates a model's discriminative power and both can not be optimized simultaneously.
- It depends on the business model and condition that, which metrics is important to us i.e. Sensitivity or Specificity.
- Our problem is of attrition i.e. employees leaving the organization, thus sensitivity metrics is more important to us.
- We want to predict employees who can leave the organization more accurately then who will not.
- So we can optimize the model for high sensitivity by decreasing the cut off value.
- For present model optimum cut off threshold is around 15%.

Model Accuracy – Optimum cut off

- Confusion Matrix at cut off >= 0.15 i.e. 15%

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  819  54
       Yes 275 175

               Accuracy : 0.7513
                 95% CI : (0.7271, 0.7744)
    No Information Rate : 0.8269
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3712
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7642
            Specificity : 0.7486
         Pos Pred Value : 0.3889
         Neg Pred Value : 0.9381
             Prevalence : 0.1731
         Detection Rate : 0.1323
   Detection Prevalence : 0.3401
      Balanced Accuracy : 0.7564

       'Positive' Class : Yes
```

	Predicted	
Actual	No (no attrition)	Yes (attrition)
No (no attrition)	819	54
Yes (attrition)	275	175

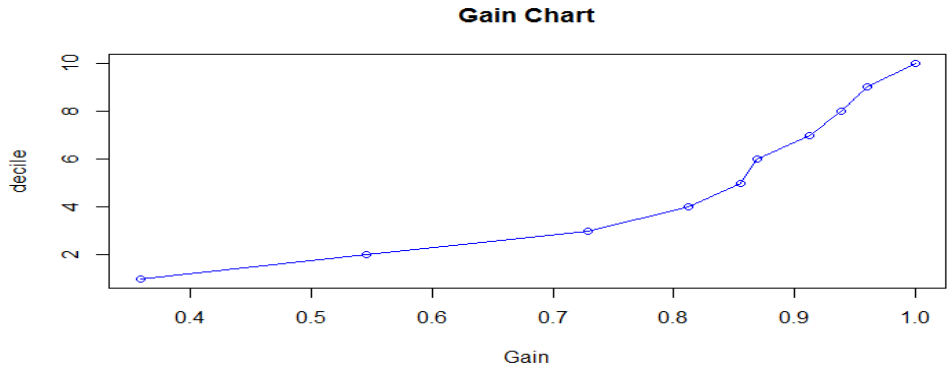
Metrics	
Accuracy	75%
Sensitivity	0.76
Specificity	0.74

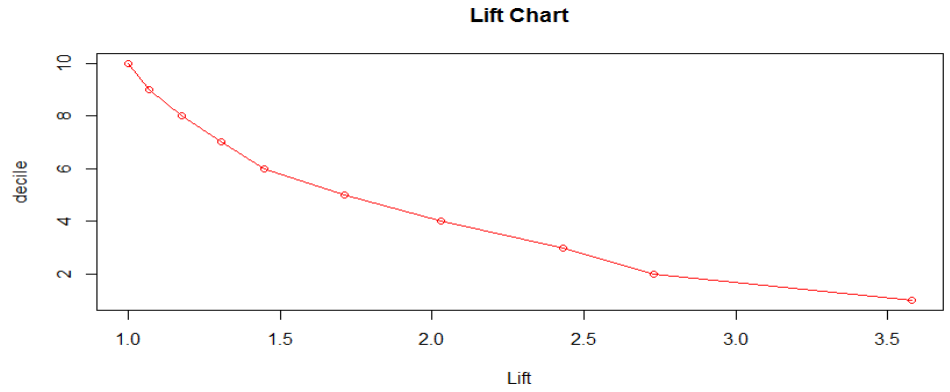
This model is fairly balanced

Model Accuracy - Gain and Lift

- Order the data based up on attrition probability in descending order.
- Prepare the gain chart which divides the test data (i.e. 1323 observations) in to 10 rows called deciles.

Lift Chart						
Decile	Observations	Churn	Cum- Churn	Gain(%Cum-Churn)	Gain (Random Model)	Lift
1	133	82	82	35.8%	10%	3.58
2	133	43	125	54.6%	20%	2.73
3	133	42	167	72.9%	30%	2.43
4	132	19	186	81.2%	40%	2.03
5	132	10	196	85.6%	50%	1.71
6	132	3	199	86.9%	60%	1.45
7	132	10	209	91.3%	70%	1.30
8	132	6	215	93.9%	80%	1.17
9	132	5	220	96.1%	90%	1.07
10	132	9	229	100.0%	100%	1.00
Total	1323	229				





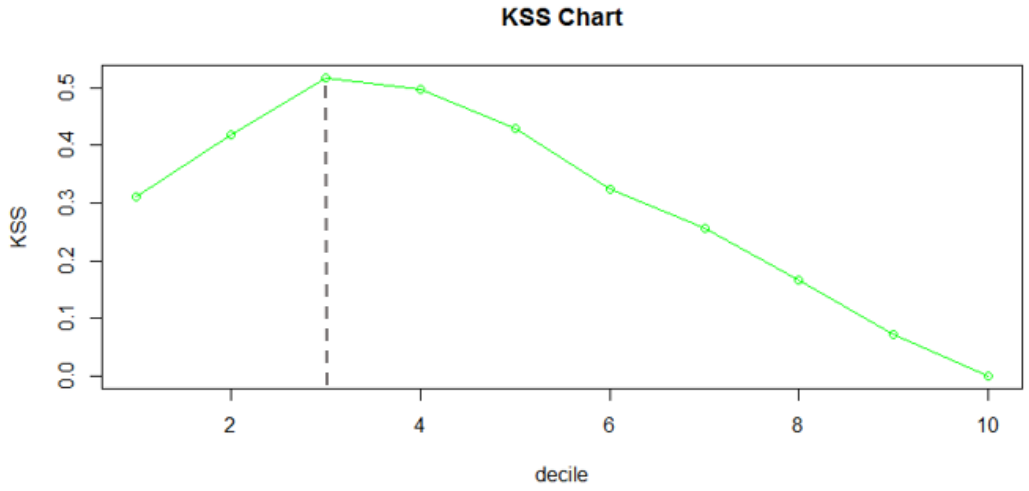
At 2nd decile we have got the maximum lift this means contact the top 20% employees in this sorted list, as that would result in contacting 54% of the employees that were likely to leave organization.

Model Accuracy - KS statistic

- A high KS statistic means that not only does the model have all attritions at the top, it has all non- attritions at the bottom.
- KS statistic is an indicator of how well your model discriminates between the two classes.
- Present model shows the KS Statistic value equals to 54% which is more that 40% and lies in 3rd decile.
- Thus model is fairly good and reliable.

KS Statistics Chart

Decile	Observations	Churn	Cum- Churn	% Cum-Churn	Non- Churn	Cum-Non-Churn	%Cum-Non-Churn	(%Cum-Churn) - (%Cum-Non-Churn)
1	133	82	82	35.8%	51	51	4.7%	31.1%
2	133	43	125	54.6%	90	141	12.9%	41.7%
3	133	42	167	72.9%	91	232	21.2%	51.7%
4	132	19	186	81.2%	113	345	31.5%	49.7%
5	132	10	196	85.6%	122	467	42.7%	42.9%
6	132	3	199	86.9%	129	596	54.5%	32.4%
7	132	10	209	91.3%	122	718	65.6%	25.6%
8	132	6	215	93.9%	126	844	77.1%	16.7%
9	132	5	220	96.1%	127	971	88.8%	7.3%
10	132	9	229	100.0%	123	1094	100.0%	0.0%
Total	1323	229			1094			



Analysis and Conclusion

1. The most influential factor is Over time i.e. the employees that work over time often are **more likely** to leave the job.
2. Followed by people with marital status as ‘Single’.
3. Third in the list are employees who tend to travel frequently.
4. Next in the list stands those kind of people who have not received their promotion since years.

1. Employees with more number of working years are **less likely** to leave the job.
2. Followed by employees who have worked for long under the current manager.
3. Factors like environment satisfaction, work life balance and job satisfaction must be addressed strongly.