HR Analytics Case Study

Group Name:Upgrad Pune- Bhubaneswar Cohorts

1. Rishabh Shrivastava
2. Vinod Jha
3. Saurav Kumar
4. Amitabha Banerjee

Business Understanding & objectives from the Data analysis

Business Underdstanding -

❑ This is a case of large Company XYZ who employs around 4000 people wants to carry out this case study to find out the reasons for attrition in their company. Every year around 15% of it's employees leaves the organization which impacts project deliverables and timelines thus resulting in reputation loss among customers and partners

❑ Attrition is mainly because of two reasons –

- Employees leaving on their own voluntarily because of good opportunities outside
- Employees fired from their jobs

Business Objectives -

The objective of analysis is to understand the major factors the company should focus on to curb attrition which is not a healthy sign of a good organization. Also, they want to know which of these variables is most important and needs to be addressed right away.

Business Constraints –

Only one year of data is available for analysis

Major deliverables from Data analysis-

❑ Identify the major factors influencing the attrition

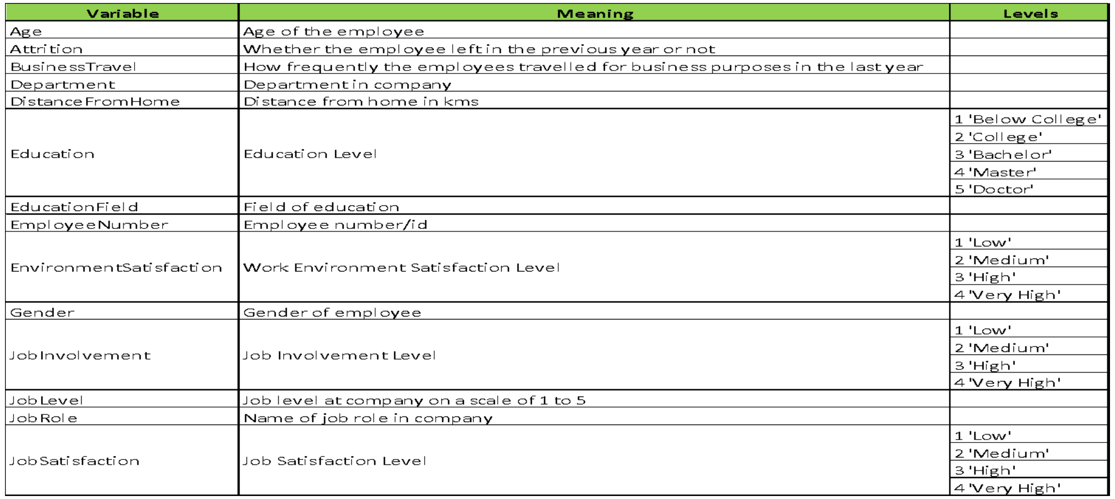
❑ Identify the correlation between various attributes with attrition and represent visually by plotting graphs (univariate & bivariate analysis)

❑ Model the probability of attrition using logistic regression and highlight the best model to management

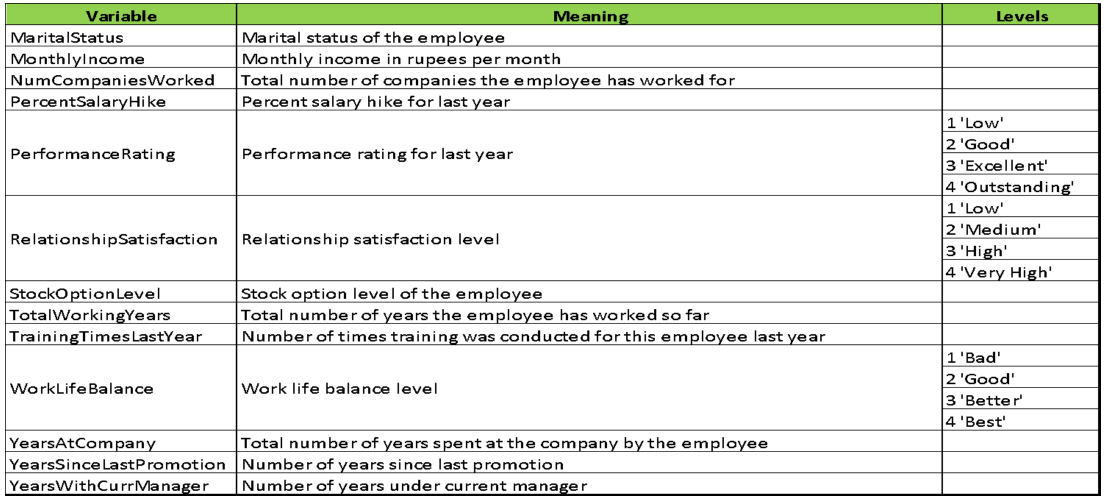
❑ Recommend and share with the management the major driving factors behind attrition cases and steps to reduce attrition . Management can then utilize this knowledge to understand what changes they should make to their workplace, in order to get most of their employees to stay.

Data Exploration & Explanation

Relevant variables for our analysis from the Data Dictionary



Variable	Meaning	Levels
Age	Age of the employee	
Attrition	Whether the employee left in the previous year or not	
BusinessTravel	How frequently the employees travelled for business purposes in the last year	
Department	Department in company	
DistanceFromHome	Distance from home in kms	
Education	Education Level	1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
EducationField	Field of education	
EmployeeNumber	Employee number/id	
EnvironmentSatisfaction	Work Environment Satisfaction Level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
Gender	Gender of employee	
JobInvolvement	Job Involvement Level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
JobLevel	Job level at company on a scale of 1 to 5	
JobRole	Name of job role in company	
JobSatisfaction	Job Satisfaction Level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'

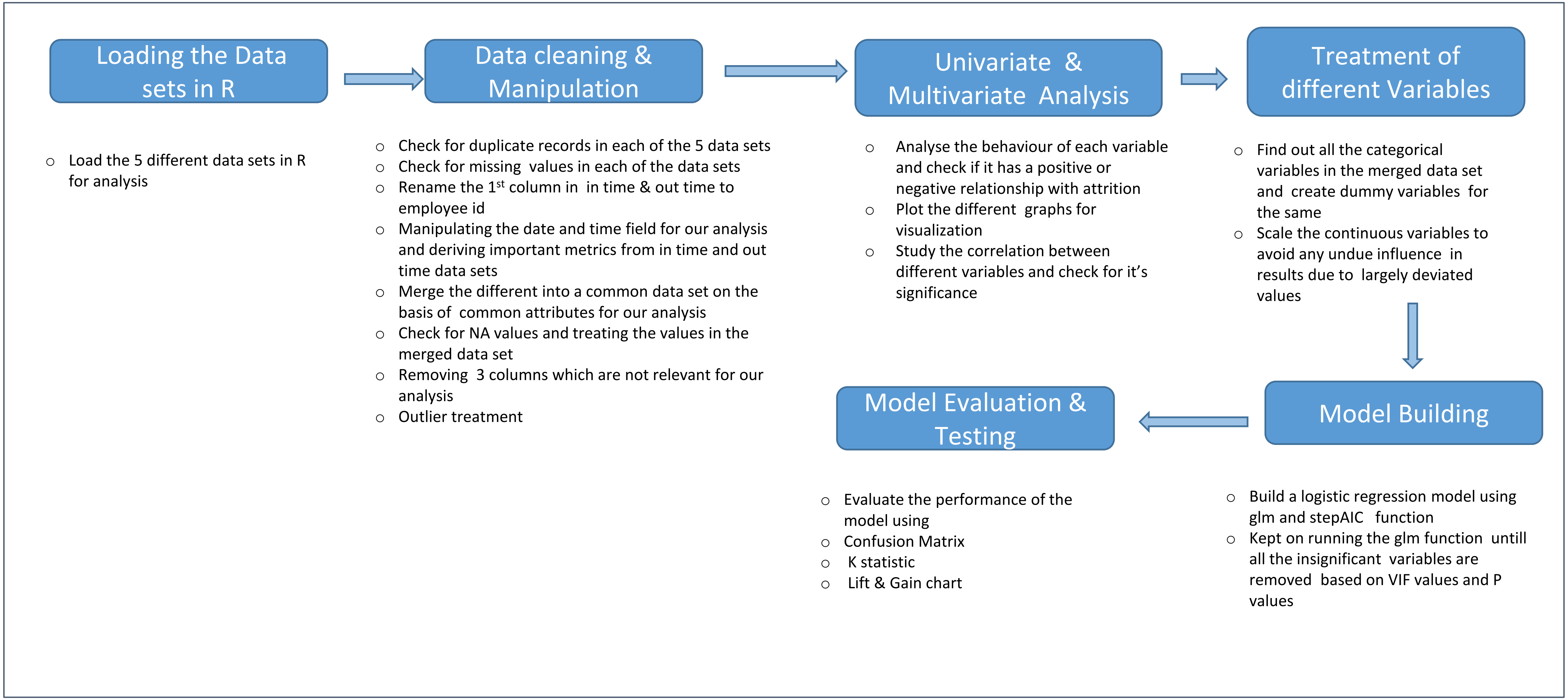


Variable	Meaning	Levels
MaritalStatus	Marital status of the employee	
MonthlyIncome	Monthly income in rupees per month	
NumCompaniesWorked	Total number of companies the employee has worked for	
PercentSalaryHike	Percent salary hike for last year	
PerformanceRating	Performance rating for last year	1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'
RelationshipSatisfaction	Relationship satisfaction level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
StockOptionLevel	Stock option level of the employee	
TotalWorkingYears	Total number of years the employee has worked so far	
TrainingTimesLastYear	Number of times training was conducted for this employee last year	
WorkLifeBalance	Work life balance level	1 'Bad' 2 'Good' 3 'Better' 4 'Best'
YearsAtCompany	Total number of years spent at the company by the employee	
YearsSinceLastPromotion	Number of years since last promotion	
YearsWithCurrManager	Number of years under current manager	

Observations -

- ❑ The company has provided us with 5 datasets for analysis –
- ❑ General Data – This contains the details of each employee. We have multiple attributes related to each employee
- ❑ Employee survey Data – It has information on employee satisfaction levels
- ❑ Manager Survey data- Provides information about each employee performance appraisal by their respective managers
- ❑ In time & Out time – Provides information of punch in and punch out time for each employee. Basically how much time the employee is spending at office.
- ❑ EmployeeCount, StandardHours & Over18 in the general survey data set are not relevant for our analysis as they are static in nature

Problem solving methodology- Process Flow



Loading the Data sets in R

- Load the 5 different data sets in R for analysis

Data cleaning & Manipulation

- Check for duplicate records in each of the 5 data sets
- Check for missing values in each of the data sets
- Rename the 1st column in in time & out time to employee id
- Manipulating the date and time field for our analysis and deriving important metrics from in time and out time data sets
- Merge the different into a common data set on the basis of common attributes for our analysis
- Check for NA values and treating the values in the merged data set
- Removing 3 columns which are not relevant for our analysis
- Outlier treatment

Univariate & Multivariate Analysis

- Analyse the behaviour of each variable and check if it has a positive or negative relationship with attrition
- Plot the different graphs for visualization
- Study the correlation between different variables and check for it's significance

Treatment of different Variables

- Find out all the categorical variables in the merged data set and create dummy variables for the same
- Scale the continuous variables to avoid any undue influence in results due to largely deviated values

Model Building

- Build a logistic regression model using glm and stepAIC   function
- Kept on running the glm function untill all the insignificant variables are removed based on VIF values and P values

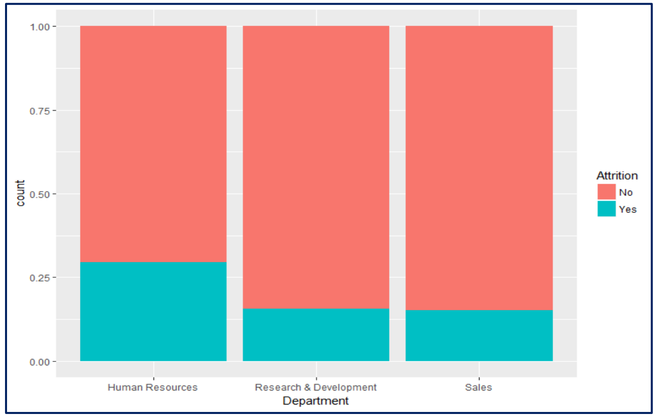
Model Evaluation & Testing

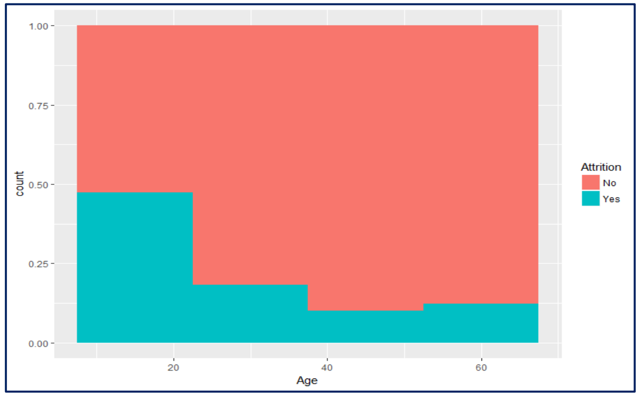
- Evaluate the performance of the model using
- Confusion Matrix
- K statistic
- Lift & Gain chart

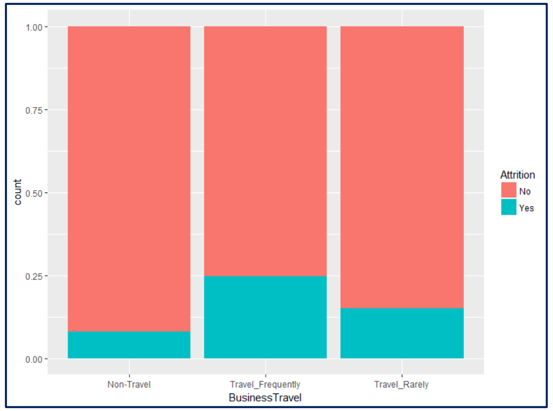
Data Cleaning & Manipulation

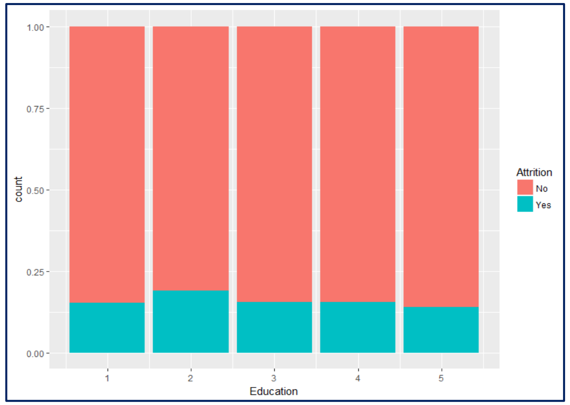
- Checked for duplicate records and missing values, found there are none
- Checked for columns which are categorical variables in nature
- Employee survey data had 83 NA’s, replaced them with the mode value
- General_data had 28 NA’s in total (19-Numcompaniesworked & 9-TotalWorking Years). Since it’s fairly small as compared to 4410 hence removed the NA’s post merging the data sets
- Analyzed the in time and out time data sets, there are large number of NA’s. Same number of NA values in both data sets, which might be there because of the absence of person from office on particular dates or because of leaves
- As per our analysis the company has given 12 leaves for the entire year and rest of the NA’s can be attributed to employee personal leaves
- Merged both in_time and out_time and derive few new metrics for finding out the average time spent by employee in office each working day and if any employee is doing overtime(working for >8 hours). Office_stay data is merged with the master data set
- Merged all the data sets to merged_employee_data on the basis of employee id for our analysis
- Removed all the unnecessary columns which are not relevant for our analysis like EmployeeCount, Over18 & Standard hours which have static value and the remaining NA’s in the merged final data set.
- Converted eligible categorical variable to factors for creating dummy variables as a prerequisite for modelling

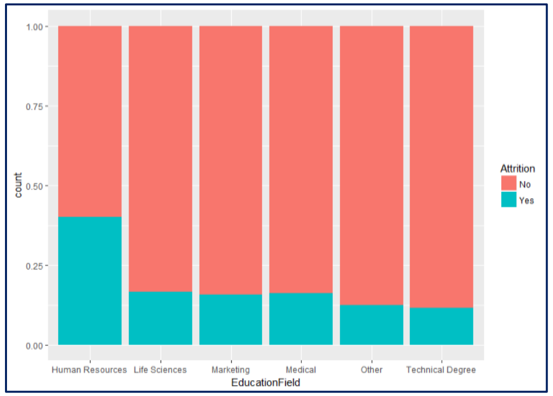
Univariate/Bivariate Analysis of Data







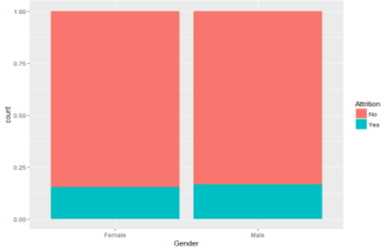


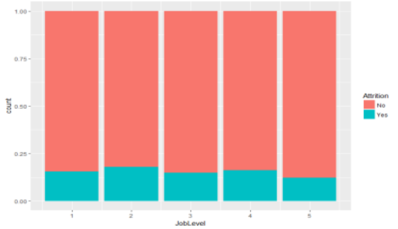


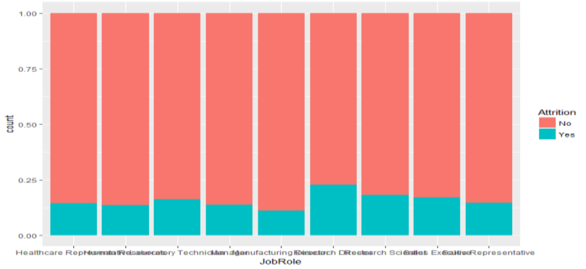
Observations –

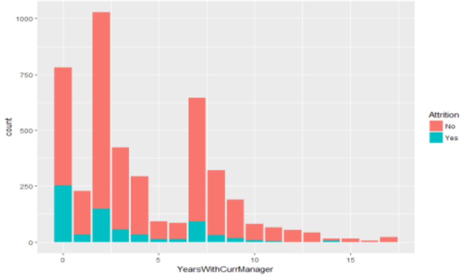
- ✓ HR department has high attrition rate but not significant enough to conclude anything
- ✓ Lower age groups are showing high attrition rates
- ✓ Those who travel frequently have higher attrition rates
- ✓ Education is not a factor for attrition

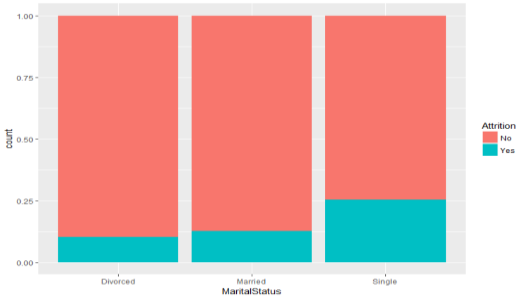
Univariate/Bivariate Analysis- Contd…

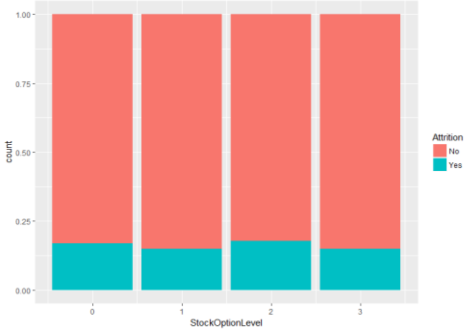


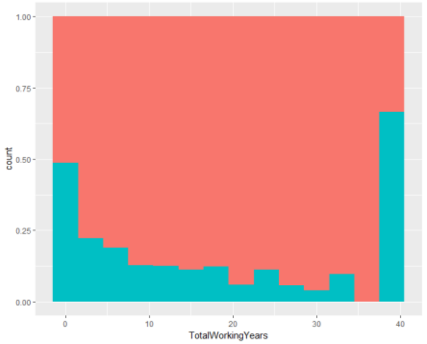


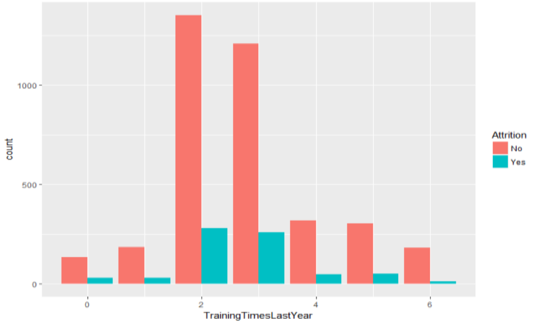


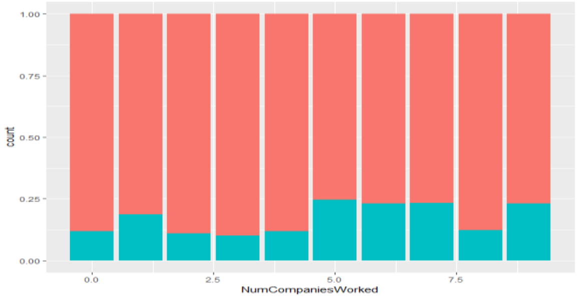


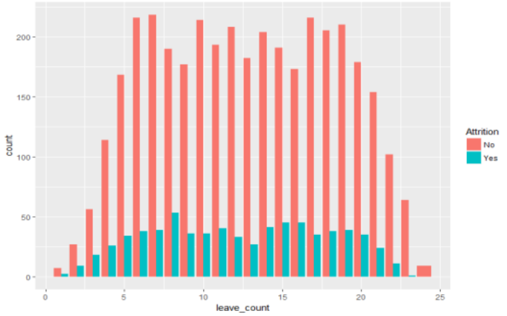








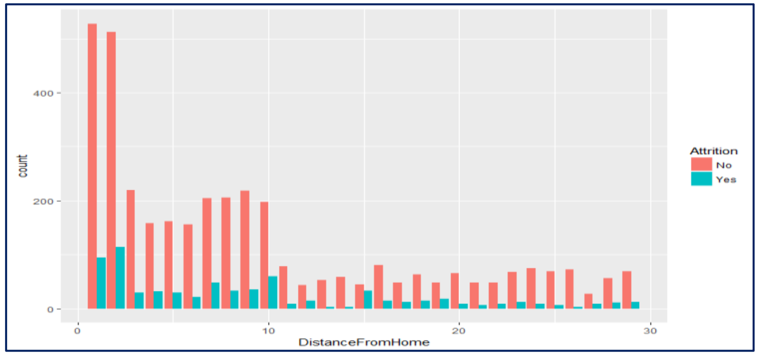


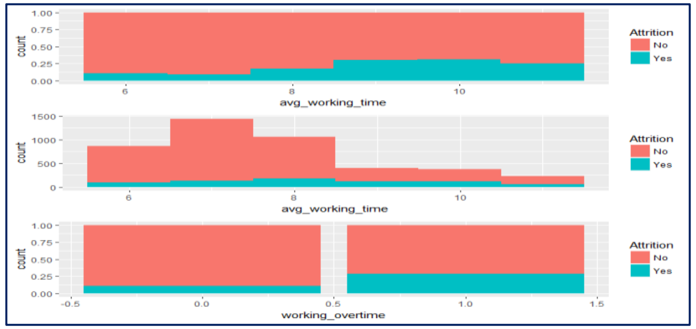


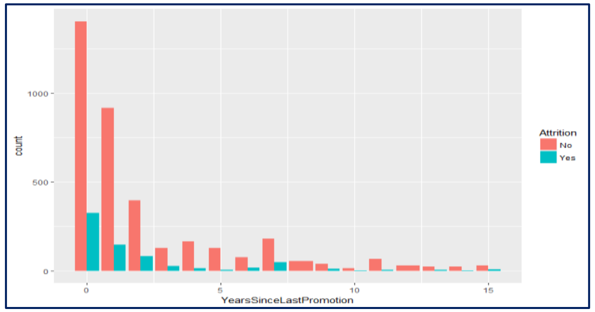
Observations –

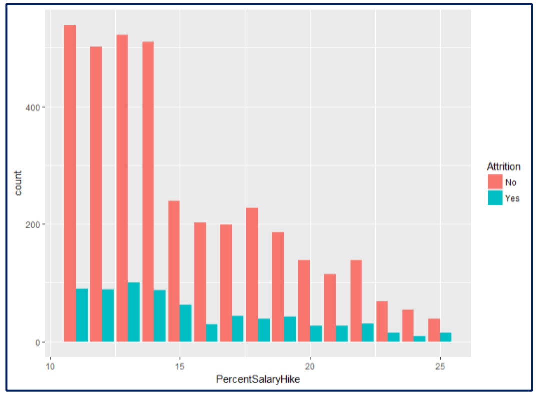
- ✓ No apparent effect of gender & Job level on attrition
- ✓ Employees who are single have high attrition rate but it is correlated with lower age
- ✓ Number of companies switched is not affecting attrition
- ✓ No apparent effect of stock option on attrition
- ✓ Employees with less years of experience and over 40 have considerable high attrition rate
- ✓ Training & Leave count also has no positive impact

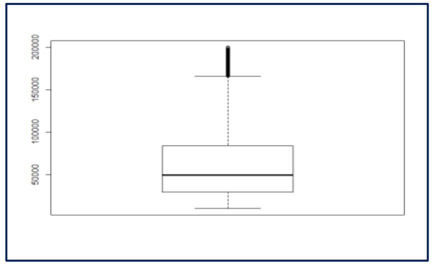
Continuous Variable Analysis-

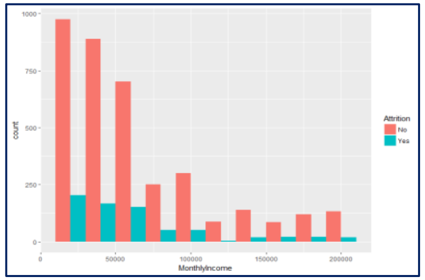


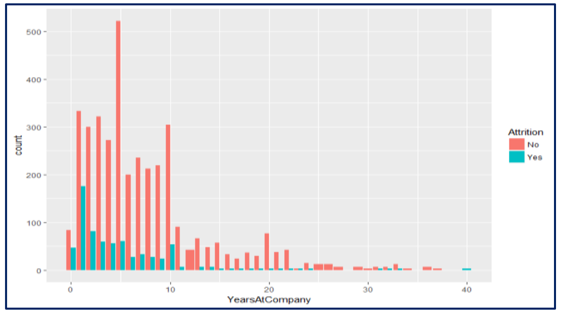












Observations –

- ✓ Distance from home is not a factor affecting attrition
- ✓ Monthly income also not a major influencing factor for attrition
- ✓ Promotion is also not impacting attrition
- ✓ Those who are doing overtime and working for more than 8 hours have more chances to leave the organization

Model Building

Pre requisites for modelling –

❑ _Created the dummy variables for each categorical variables and other input variables which were to be fed to the model

❑ Scaled the continuous variables which could have undue influence on the results

❑ Dependent variable in this case is attrition which has to be predicted using logistic regression model

❑ Data split (training: testing) was done in the ratio of 70:30

Model Building–

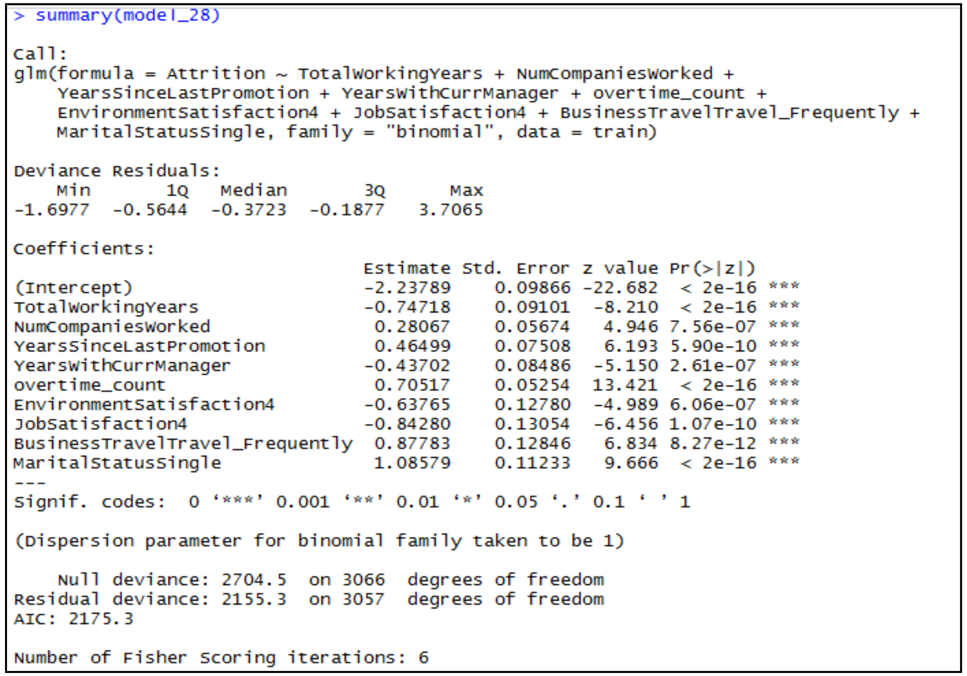
❑ Used glm function in R to build our logistic regression model for prediction

❑ Used stepAIC function as per the standard to build an optimal model

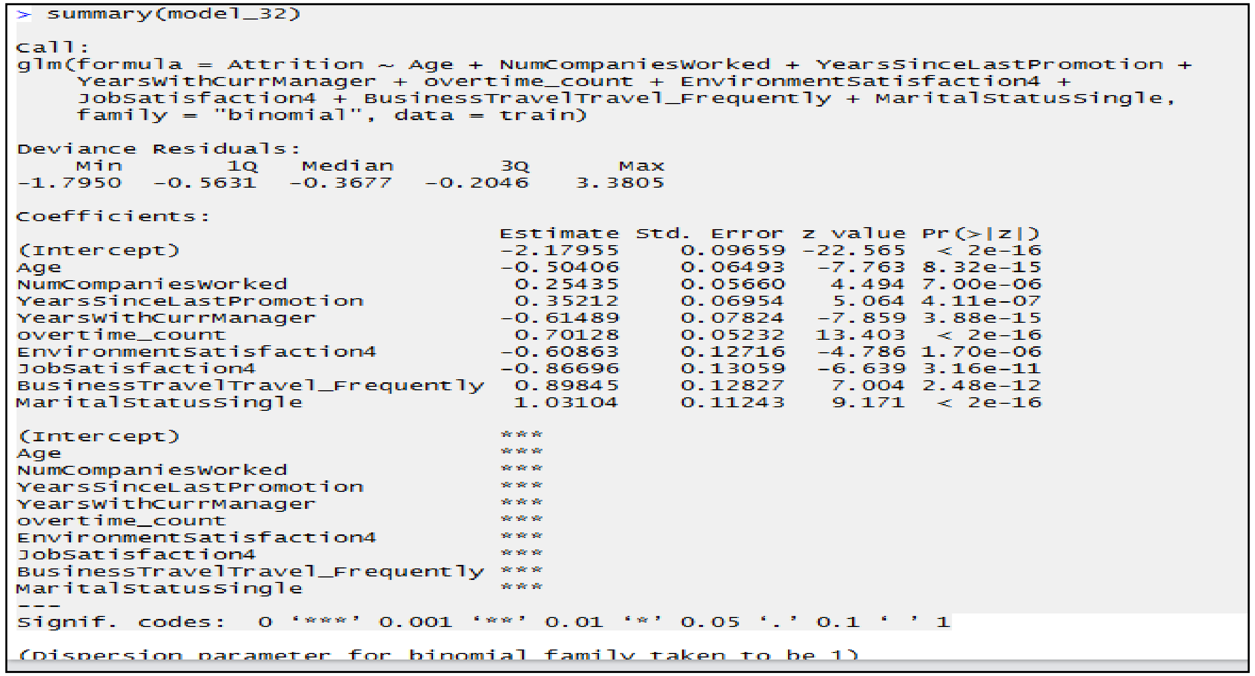
❑ Insignificant variables at each step was removed with the help of VIF (Variable inflation factor) values and P values

❑ **Model_32** is our ideal model with all the significant variables

Model Building –Contd –Final Model Selection



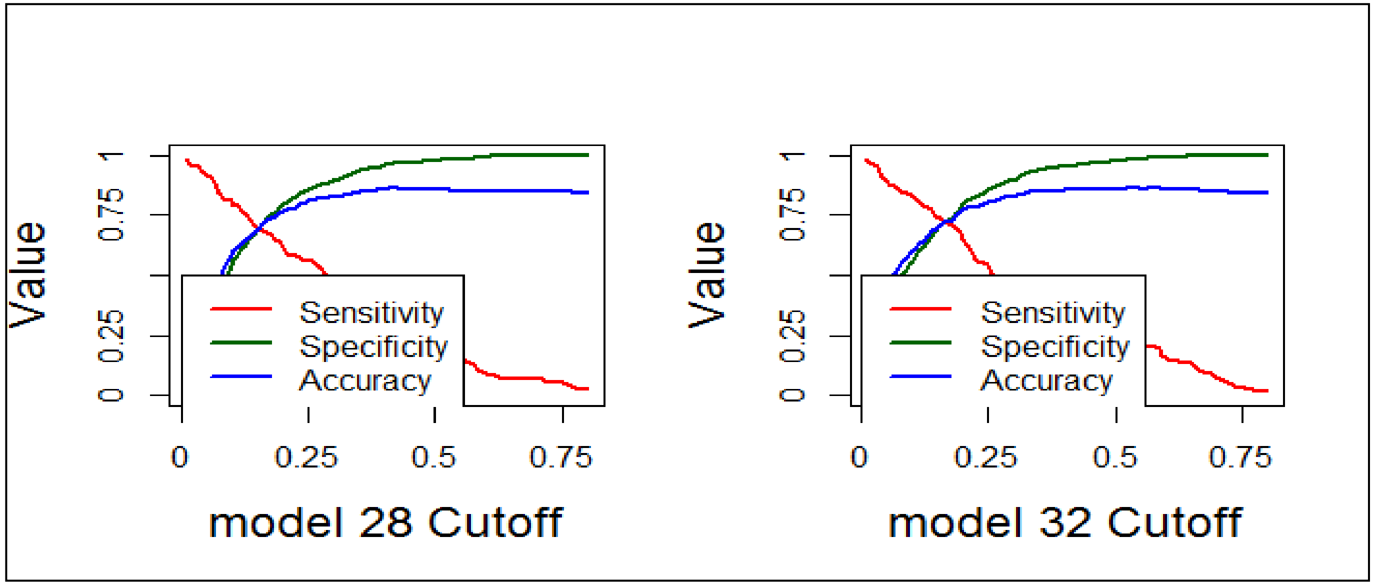
```
> summary(model_28)

Call:
glm(formula = Attrition ~ TotalWorkingYears + NumCompaniesWorked + 
    YearsSinceLastPromotion + YearsWithCurrManager + overtime_count + 
    EnvironmentSatisfaction4 + JobSatisfaction4 + BusinessTravelTravel_Frequently + 
    MaritalStatusSingle, family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6977  -0.5644  -0.3723  -0.1877   3.7065  

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     -2.23789    0.09866 -22.682  < 2e-16 ***
TotalWorkingYears               -0.74718    0.09101  -8.210  < 2e-16 ***
NumCompaniesWorked               0.28067    0.05674   4.946 7.56e-07 ***
YearsSinceLastPromotion          0.46499    0.07508   6.193 5.90e-10 ***
YearsWithCurrManager            -0.43702    0.08486  -5.150 2.61e-07 ***
overtime_count                   0.70517    0.05254  13.421  < 2e-16 ***
EnvironmentSatisfaction4        -0.63765    0.12780  -4.989 6.06e-07 ***
JobSatisfaction4                -0.84280    0.13054  -6.456 1.07e-10 ***
BusinessTravelTravel_Frequently  0.87783    0.12846   6.834 8.27e-12 ***
MaritalStatusSingle              1.08579    0.11233   9.666  < 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: 2704.5  on 3066  degrees of freedom
Residual deviance: 2155.3  on 3057  degrees of freedom
AIC: 2175.3

Number of Fisher Scoring iterations: 6
```



```
> summary(model_32)

Call:
glm(formula = Attrition ~ Age + NumCompaniesWorked + YearsSinceLastPromotion + 
    YearsWithCurrManager + overtime_count + EnvironmentSatisfaction4 + 
    JobSatisfaction4 + BusinessTravelTravel_Frequently + MaritalStatusSingle, 
    family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7950  -0.5631  -0.3677  -0.2046   3.3805  

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.17955    0.09659 -22.565  < 2e-16
Age                             -0.50406    0.06493  -7.763 8.32e-15
NumCompaniesWorked               0.25435    0.05660   4.494 7.00e-06
YearsSinceLastPromotion          0.35212    0.06954   5.064 4.11e-07
YearsWithCurrManager            -0.61489    0.07824  -7.859 3.88e-15
overtime_count                   0.70128    0.05232  13.403  < 2e-16
EnvironmentSatisfaction4        -0.60863    0.12716  -4.786 1.70e-06
JobSatisfaction4                -0.86696    0.13059  -6.639 3.16e-11
BusinessTravelTravel_Frequently  0.89845    0.12827   7.004 2.48e-12
MaritalStatusSingle              1.03104    0.11243   9.171  < 2e-16
                                   
(Intercept)                     ***
Age                             ***
NumCompaniesWorked              ***
YearsSinceLastPromotion         ***
YearsWithCurrManager            ***
overtime_count                  ***
EnvironmentSatisfaction4        ***
JobSatisfaction4                ***
BusinessTravelTravel_Frequently ***
MaritalStatusSingle             ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Observations –
Based on VIF & P values we arrived at **model_28** as the optimal model with all significant variables. Further as per business understanding we added & removed few variables and checked their effect on the model and we arrived at **model_32** as another optimal model for consideration. We evaluated both the models using K statistics and confusion matrix in the next slide for selecting the best model amongst the two.

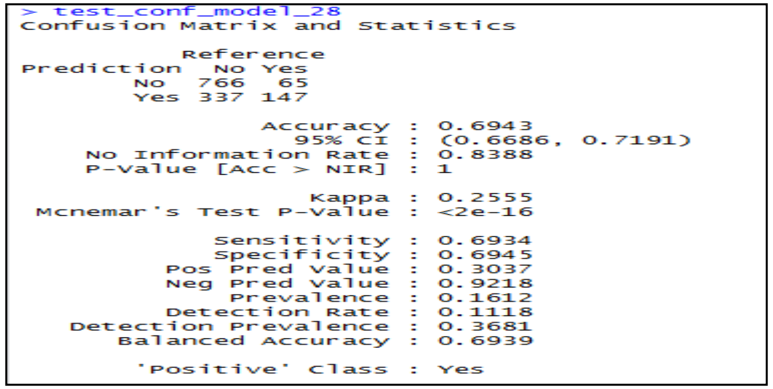
Model Evaluation-Confusion Matrix



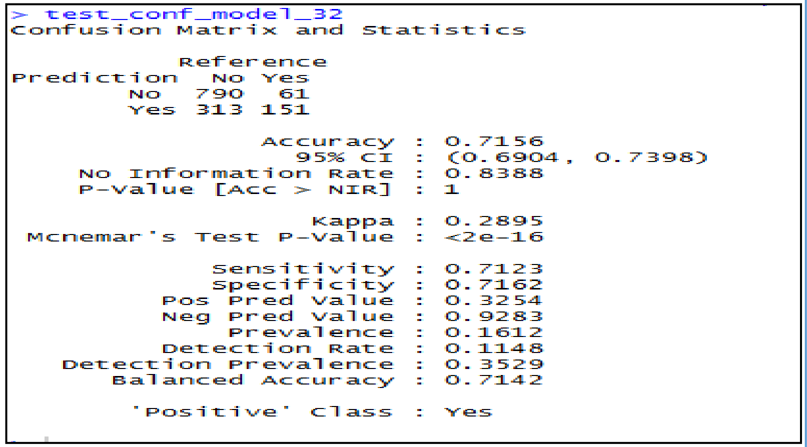
Observations –

✓ Cut off for model_28 is **- 0.1536364** & cutoff for model_32 is -**0.1616162**

✓ Sensitivity, specificity & accuracy in model_32 **(71%)** are higher as compared to model_28 **(69%)**



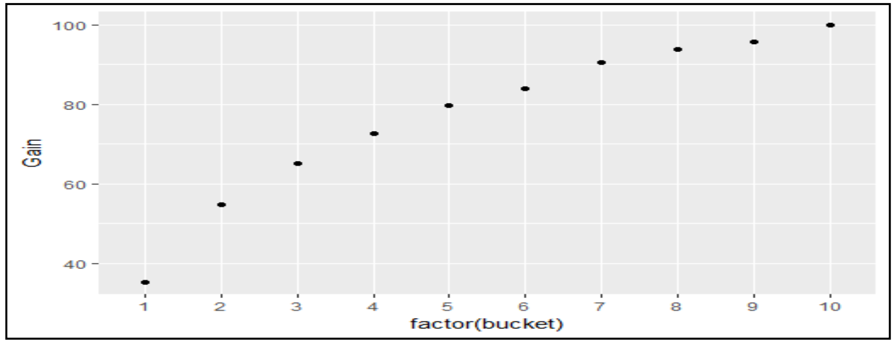
```
> test_conf_model_28
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  766  65
       Yes 337 147

               Accuracy : 0.6943
                 95% CI : (0.6686, 0.7191)
    No Information Rate : 0.8388
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2555
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6934
            Specificity : 0.6945
         Pos Pred Value : 0.3037
         Neg Pred Value : 0.9218
             Prevalence : 0.1612
         Detection Rate : 0.1118
   Detection Prevalence : 0.3681
      Balanced Accuracy : 0.6939

       'Positive' Class : Yes
```



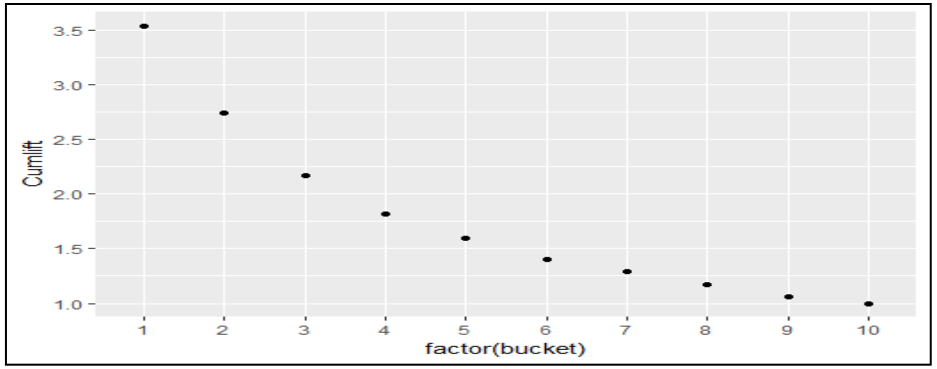
```
> test_conf_model_32
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  790  61
       Yes 313 151

               Accuracy : 0.7156
                 95% CI : (0.6904, 0.7398)
    No Information Rate : 0.8388
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2895
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7123
            Specificity : 0.7162
         Pos Pred Value : 0.3254
         Neg Pred Value : 0.9283
             Prevalence : 0.1612
         Detection Rate : 0.1148
   Detection Prevalence : 0.3529
      Balanced Accuracy : 0.7142

       'Positive' Class : Yes
```

Model Evaluation –Contd… Lift & Gain charts

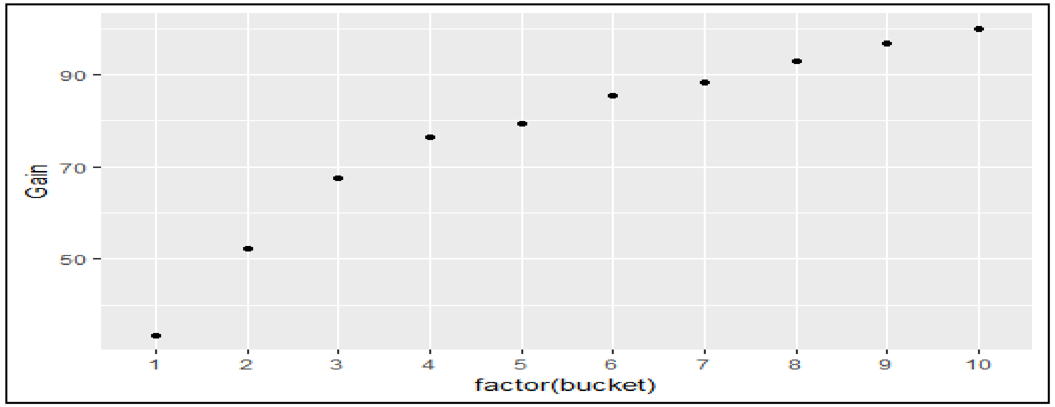
Gain Chart Model_28



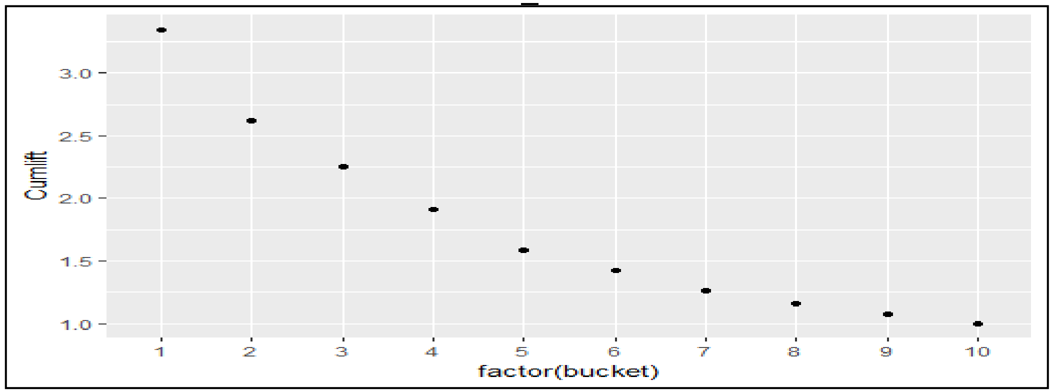
Lift Chart Model_28



Gain Chart Model_32



Lift Chart Model_32

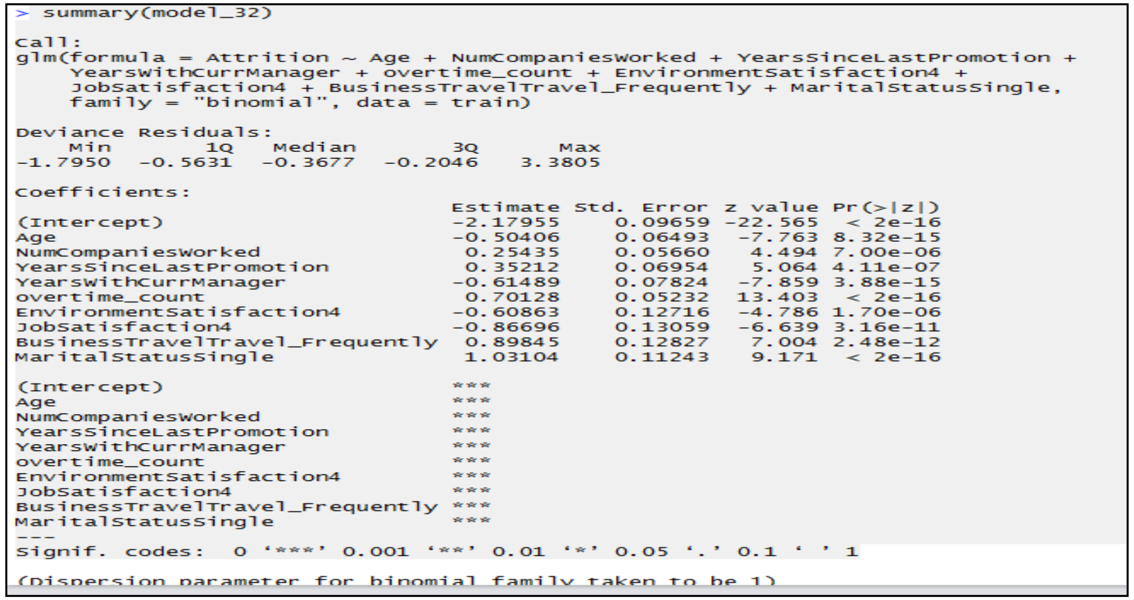


Observations –

We are able to predict **75%** of attrition in 4th decile for **model_28** whereas for **model_32** we are able to predict approximately **77%** of attrition in 4th decile. Hence, **model_32** is better in terms of lift/gain chart

Model Evaluation –Contd –Final Model Selection

Performance Measurement	Model_28	Model_32
Sensitivity	69.34%	71.2
Specificity	69.45%	71.60%
Accuracy	69.40%	71.56%
K Statistics	41.7% (3rd decile)	44.5%(3rd decile)
Lift/Gain chart	We are able to predict 75% of attrition in 4th decile	We are able to predict approx. 77% of attrition in 4th decile



```
> summary(model_32)

Call:
glm(formula = Attrition ~ Age + NumCompaniesWorked + YearsSinceLastPromotion +
    YearsWithCurrManager + overtime_count + EnvironmentSatisfaction4 +
    JobSatisfaction4 + BusinessTravelTravel_Frequently + MaritalStatusSingle,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7950  -0.5631  -0.3677  -0.2046   3.3805

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.17955    0.09659 -22.565  < 2e-16
Age                             -0.50406    0.06493  -7.763 8.32e-15
NumCompaniesWorked               0.25435    0.05660   4.494 7.00e-06
YearsSinceLastPromotion          0.35212    0.06954   5.064 4.11e-07
YearsWithCurrManager            -0.61489    0.07824  -7.859 3.88e-15
overtime_count                   0.70128    0.05232  13.403  < 2e-16
EnvironmentSatisfaction4        -0.60863    0.12716  -4.786 1.70e-06
JobSatisfaction4                -0.86696    0.13059  -6.639 3.16e-11
BusinessTravelTravel_Frequently  0.89845    0.12827   7.004 2.48e-12
MaritalStatusSingle              1.03104    0.11243   9.171  < 2e-16

(Intercept)                     ***
Age                             ***
NumCompaniesWorked              ***
YearsSinceLastPromotion         ***
YearsWithCurrManager            ***
overtime_count                  ***
EnvironmentSatisfaction4        ***
JobSatisfaction4                ***
BusinessTravelTravel_Frequently ***
MaritalStatusSingle             ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Observations –
Based on Confusion matrix, K statistics measure and lift/gain chart we are concluding that **model_32** is the best model for our prediction in this case as it will correctly and more accurately predict each case of employee attrition

Conclusion-Major Influencing factors for attrition & recommendation

Major Factors influencing attrition	Observations & Recommendations
Age	Employees who are <30 are more likely to switch jobs as per the trend . Company should focus on such employees and likely to concentrate on the more skilled ones and should provide them rewards and incentives to retain them.
Number of Companies Worked(NumCompaniesWorked)	If the employee has switched employment multiple times(>8) within his total career span then the company should ask for specific reasons for preventing attrition. Company should try to understand in skip levels the aspirations of such employees if they are satisfied or not and what is their future expectation for better employee connect
Last Promotion (YearsSinceLastPromotion)	Company should focus on those employees who were not promoted for a long period of time. Company should look into their past performance appraisals and if they are good then they should have a discussion with their respective supervisors to understand the specific reasons of delay in promotion and if possible to promote them in the upcoming cycle
Duration with Current Manager(YearsWithCurrManager)	The more the duration of the employee's tenure with the current manager the less likely he/she will leave the job. If the manager is easily approachable and interacts with the employees more often then it will create a healthy environment and will retain employees
Overtime (Overtime_Count)	Those employees who are doing overtime and are spending more than 8 hours in office more often they are more likely to quit. It can be due to work pressure and improper work life balance. Management should focus on them and should ensure that the skill matrix and talent pool is competent so that the work load is balanced. For additional work there should be attractive incentive proposals.
Environment Satisfaction(EnvironmentSatisfaction4)	Those with high satisfaction are likely to continue in their job, company should ensure that work environment is healthy and good so that employees will get motivated to work. Gifts ,bonus /allowances should be tagged to employee’s performance
Job Satisfaction(JobsSatisfaction4)	The employees with higher job satisfaction will likely to stay in the job and continue. Company should have a job rotation policy after every 2 - 2.5 year to avoid monotony in the job. Time to Time skip levels with the employees also helps
Business Travel (BusinessTravel_Travel_Frequently)	The trend shows that those who are travelling more frequently are likely to quit more as they may not be happy travelling often and are compelled to travel because of certain business reasons. Company should focus on such cases and try to have a open discussion with them to check if they are happy with their roles or they are looking for a role change.
Marital Status (MaritalStatusSingle)	The trend shows that those who are single are more likely to leave as they can easily relocate to other places if they have better offers in hand as they don't have to think about their family and children's school and other stuff