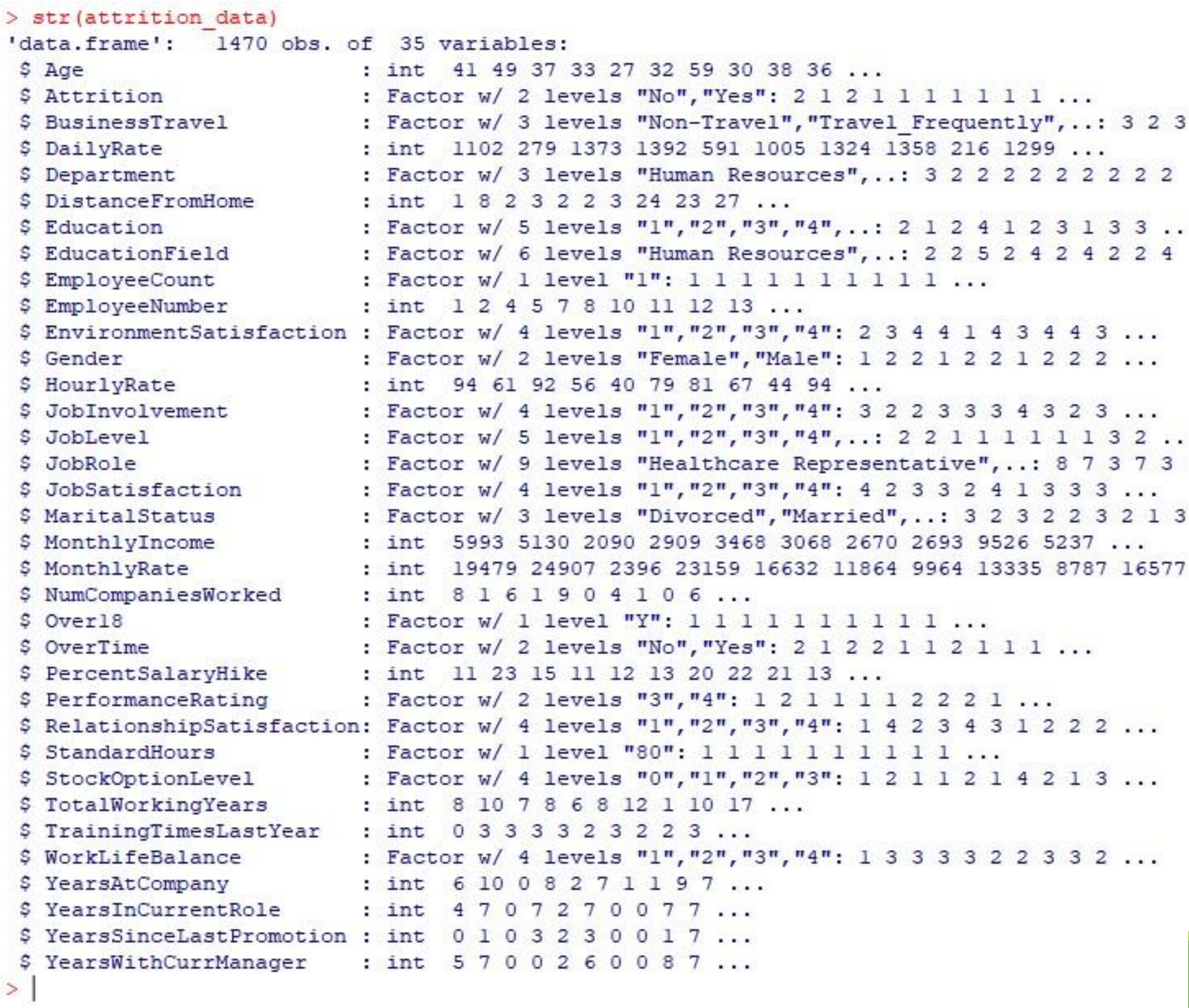
PROJECT PRESENTATION ON EMPLOYMENT ATTRITION DATA ANALYSIS

READING DATA

- 1470 Rows
- 34 Variables
- Mixed data types
- Contains categorical and numeric data types
- 'Attrition' is The Target variable
- Target variable is categorical and having two class "YES" and "NO"
- By observing data it can be conclude that the data is LABLED Data



```
> str(attrition_data)
'data.frame':	1470 obs. of  35 variables:
 $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
 $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
 $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 2 3
 $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
 $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2
 $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
 $ Education               : Factor w/ 5 levels "1","2","3","4",..: 2 1 2 4 1 2 3 1 3 3 ..
 $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4
 $ EmployeeCount           : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...
 $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
 $ EnvironmentSatisfaction : Factor w/ 4 levels "1","2","3","4": 2 3 4 4 1 4 3 4 4 3 ...
 $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
 $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
 $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 3 2 2 3 3 3 4 3 2 3 ...
 $ JobLevel                : Factor w/ 5 levels "1","2","3","4",..: 2 2 1 1 1 1 1 1 3 2 ..
 $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3
 $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4 1 3 3 3 ...
 $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3
 $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
 $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577
 $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
 $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
 $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...
 $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
 $ PerformanceRating       : Factor w/ 2 levels "3","4": 1 2 1 1 1 1 2 2 2 1 ...
 $ RelationshipSatisfaction: Factor w/ 4 levels "1","2","3","4": 1 4 2 3 4 3 1 2 2 2 ...
 $ StandardHours           : Factor w/ 1 level "80": 1 1 1 1 1 1 1 1 1 1 ...
 $ StockOptionLevel        : Factor w/ 4 levels "0","1","2","3": 1 2 1 1 2 1 4 2 1 3 ...
 $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
 $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
 $ WorkLifeBalance         : Factor w/ 4 levels "1","2","3","4": 1 3 3 3 3 2 2 3 3 2 ...
 $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
 $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
 $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
 $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
> |
```

CATEGORICAL DATA DEFINITIONS

Education -> 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'

EnvironmentSatisfaction -> 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

JobInvolvement ->1 'Low' 2 'Medium' 3 'High' 4 'Very High'

JobSatisfaction -> 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

PerformanceRating -> 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'

RelationshipSatisfaction -> 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

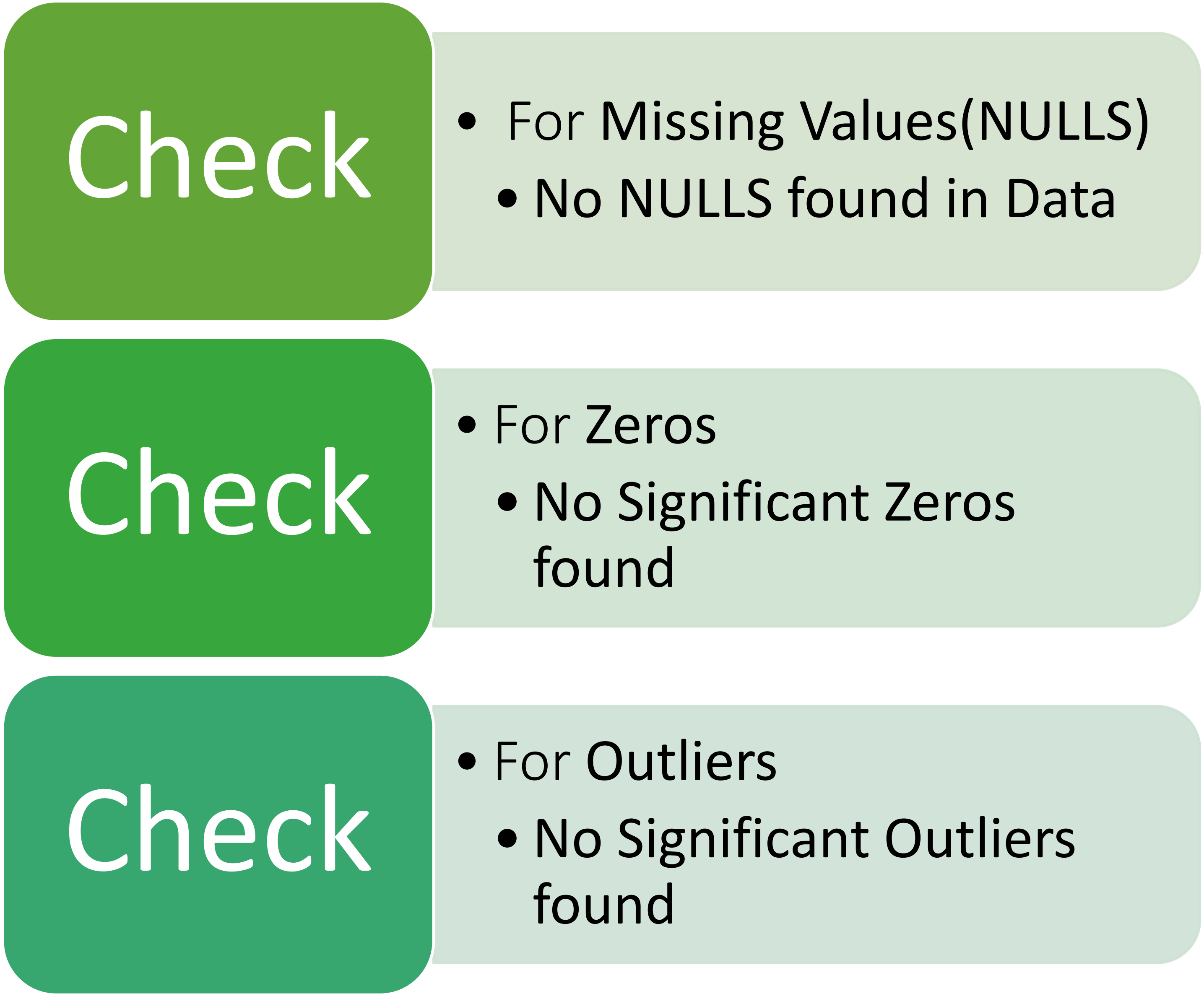
WorkLifeBalance -> 1 'Bad' 2 'Good' 3 'Better' 4 'Best'

LIBRARIES USED FOR ANALYSIS

- CARET
- CORRPLOT
- GGPLOT2
- RPART
- ROCR
- PROC
- DATA.TABLE
- Random Forest

Exploratory Data Analysis

VARIABLES CHECK



Check

- For Missing Values(NULLS)
  - No NULLS found in Data

Check

- For Zeros
  - No Significant Zeros found

Check

- For Outliers
  - No Significant Outliers found

UNIVARIATE ANALYSIS

Five Nos. Summary

Age Min. :18.00 1st Qu. :30.00 Median :36.00 Mean :36.92 3rd Qu. :43.00 Max. :60.00	DailyRate Min. : 102.0 1st Qu. : 465.0 Median : 802.0 Mean : 802.5 3rd Qu. :1157.0 Max. :1499.0	DistanceFromHome Min. : 1.000 1st Qu. : 2.000 Median : 7.000 Mean : 9.193 3rd Qu. :14.000 Max. :29.000	EmployeeNumber Min. : 1.0 1st Qu. : 491.2 Median :1020.5 Mean :1024.9 3rd Qu. :1555.8 Max. :2068.0	HourlyRate Min. : 30.00 1st Qu. : 48.00 Median : 66.00 Mean : 65.89 3rd Qu. : 83.75 Max. :100.00
MonthlyIncome Min. : 1009 1st Qu. : 2911 Median : 4919 Mean : 6503 3rd Qu. : 8379 Max. :19999	MonthlyRate Min. : 2094 1st Qu. : 8047 Median :14236 Mean :14313 3rd Qu. :20462 Max. :26999	NumCompanyWork Min. :0.000 1st Qu. :1.000 Median :2.000 Mean :2.693 3rd Qu. :4.000 Max. :9.000	PercentSalaryHike Min. :11.00 1st Qu. :12.00 Median :14.00 Mean :15.21 3rd Qu. :18.00 Max. :25.00	TotalWorkingYears Min. : 0.00 1st Qu. : 6.00 Median :10.00 Mean :11.28 3rd Qu. :15.00 Max. :40.00
TrainingTimesLastYr. Min. :0.000 1st Qu. :2.000 Median :3.000 Mean :2.799 3rd Qu. :3.000 Max. :6.000	YearsAtCompany Min. : 0.000 1st Qu. : 3.000 Median : 5.000 Mean : 7.008 3rd Qu. : 9.000 Max. :40.000	YearsInCurrentRole Min. : 0.000 1st Qu. : 2.000 Median : 3.000 Mean : 4.229 3rd Qu. : 7.000 Max. :18.000	YrSinceLastPromtion Min. : 0.000 1st Qu. : 0.000 Median : 1.000 Mean : 2.188 3rd Qu. : 3.000 Max. :15.000	YeWithCurManager Min. : 0.000 1st Qu. : 2.000 Median : 3.000 Mean : 4.123 3rd Qu. : 7.000 Max. :17.000

BIVARIATE ANALYSIS

Chi-sq. Test

A hypothesis test designed to test for a statistically significant relationship between categorical variables organized in a **bivariate** table

Used to determine the association between two variable

NULL HYPOTHESIS (H_0)

Two Categorical variable are independent/No relation exist

ALTERNATE HYPOTHESIS (H_1)

Two Categorical variable are not independent/relation exist

	Features	P-value
1.	BusinessTravel	5.609e-06
2.	MaritalStatus	9.455511e-11
3.	Gender	0.2906
4.	EnvironmentSatisfaction	0.0005563
5.	StockOptionLevel	4.379390e-13
6.	RelationshipSatisfaction	0.155
7.	PerformanceRating	0.9901
8.	JobLevel	6.634685e-15
9.	Department	0.004526
10.	WorkLifeBalance	0.0009726
11.	JobInvolvement	2.863181e-06
12.	JobRole	2.752e-15
13.	JobSatisfaction	0.0005563
14.	OverTime	2.2e-16
15.	EducationField	0.007496
16.	Education	0.5455

INDEPENDENT/NO RELATION EXIST BETWEEN-

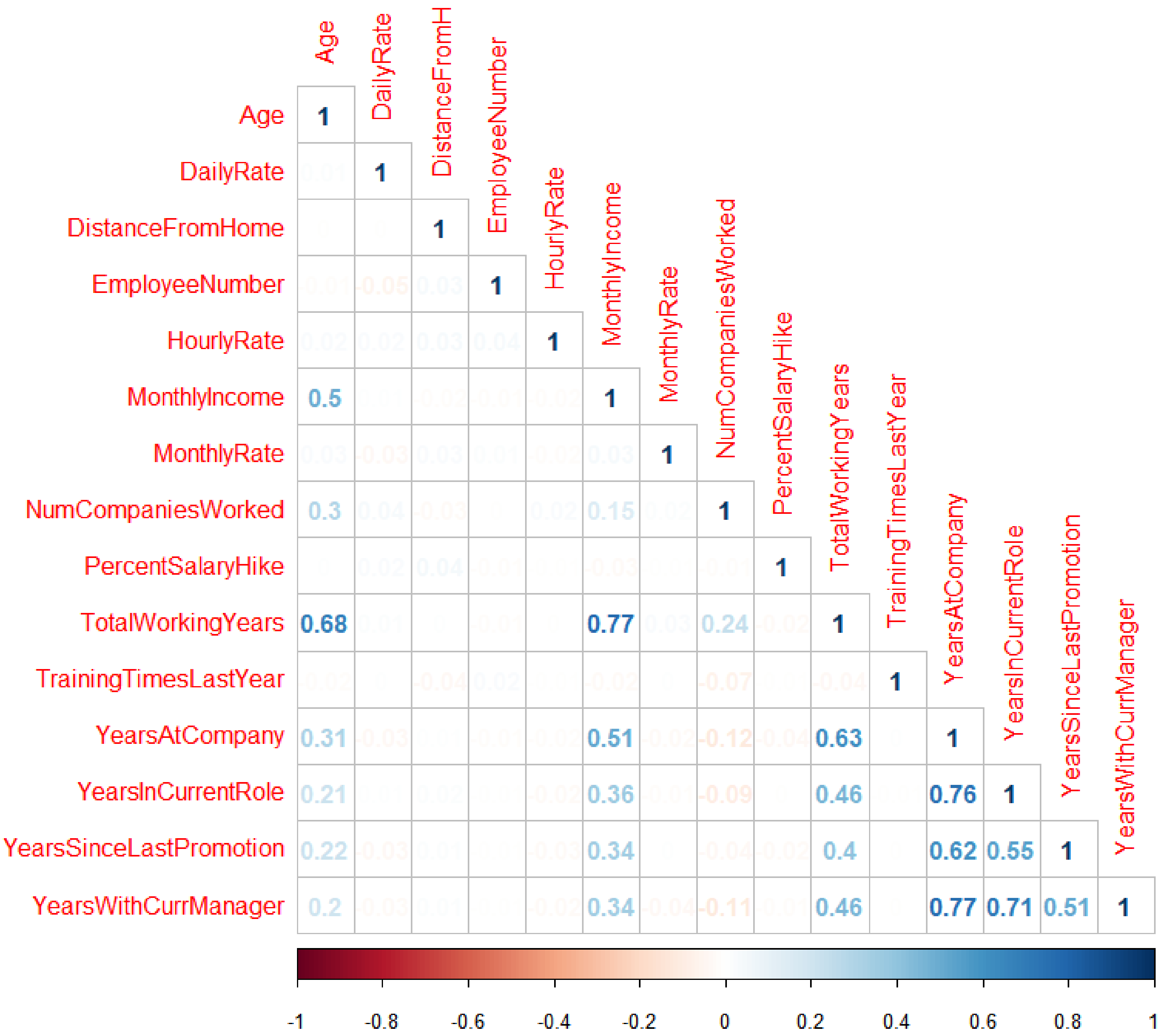
- EnvironmentSatisfaction and Attrition
- JobSatisfaction and Attrition
- OverTime and Attrition
- EducationField and Attrition
- Department and Attrition
- WorkLifeBalance and Attrition
- JobRole and Attrition are

MULTICOLLINEARITY

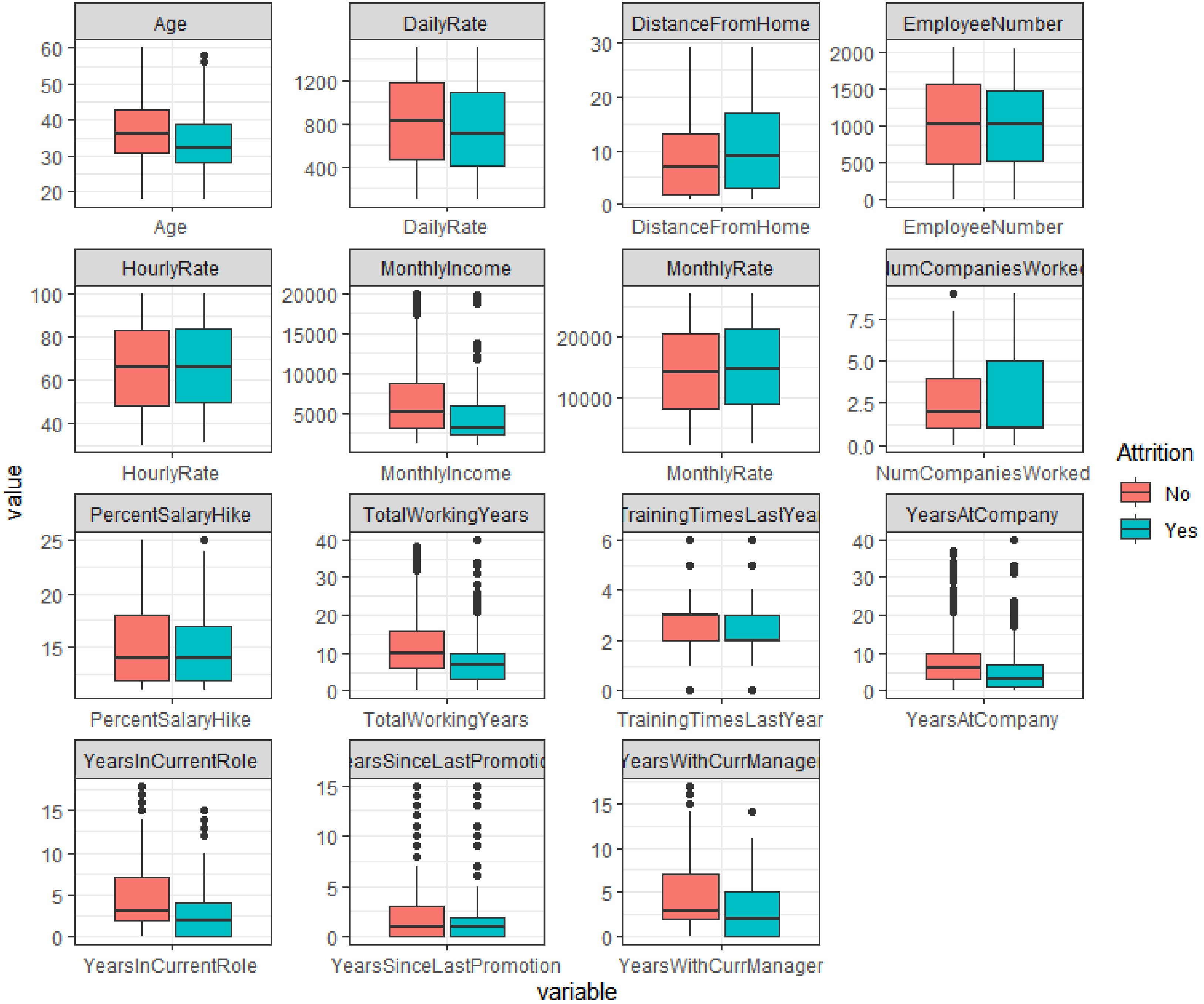
TotalWorkingYear X Monthly Income -> 0.77

YearsWithCurrentManager X YearsAtCompany ->0.77

YearsInCurrentRole X YearsAtCompany -> 0.76

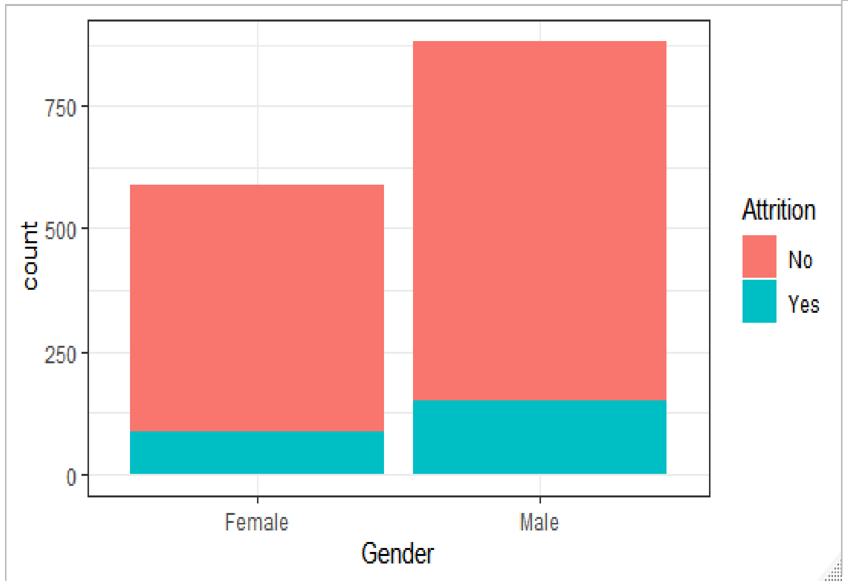


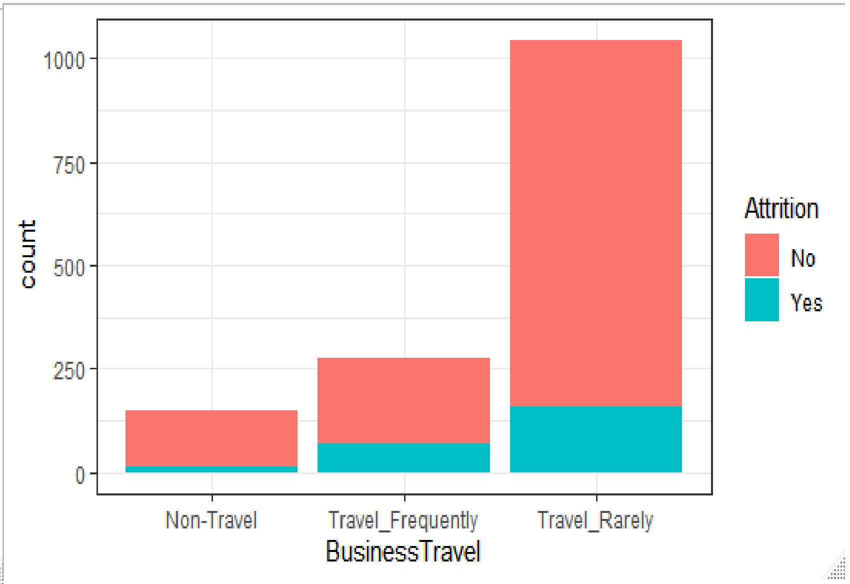
PLOTING DATA/ VISULISATION

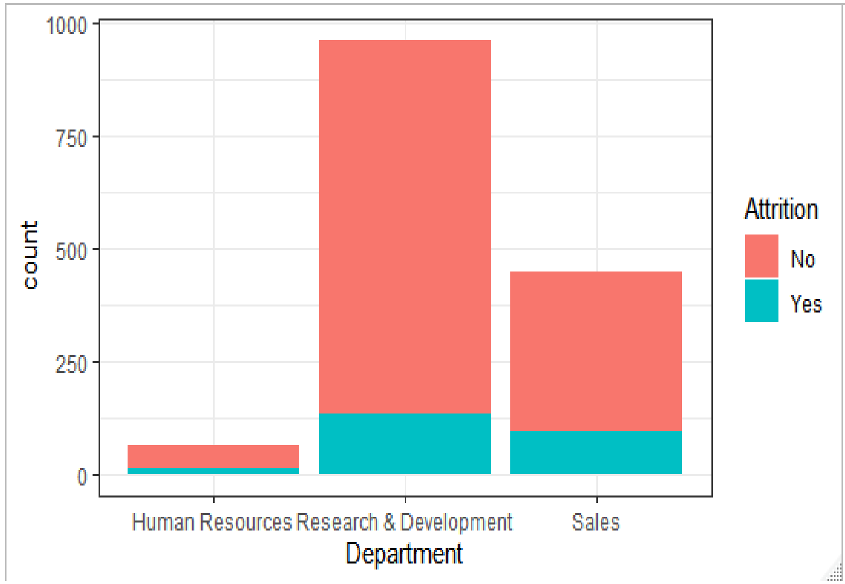


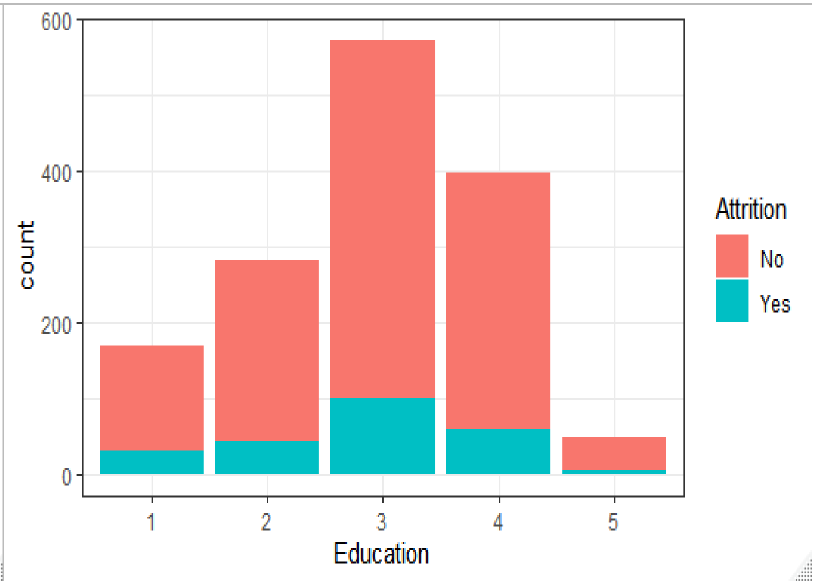
PLOTING DATA/ VISULISATION

Categorical Data Vs. Target Data

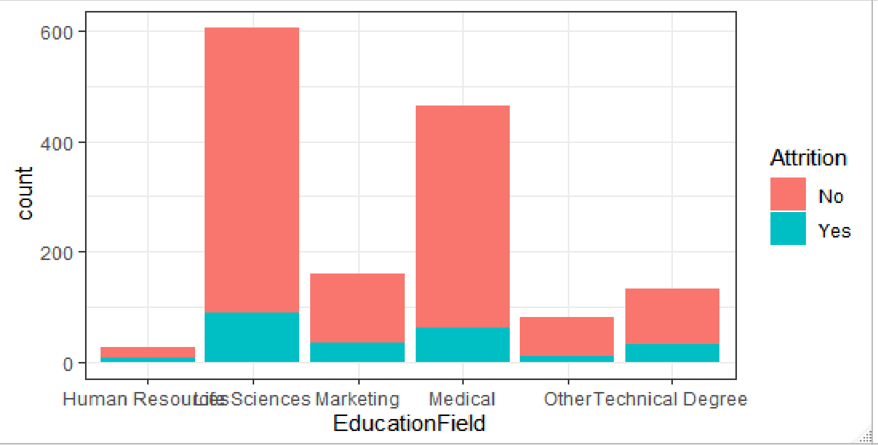


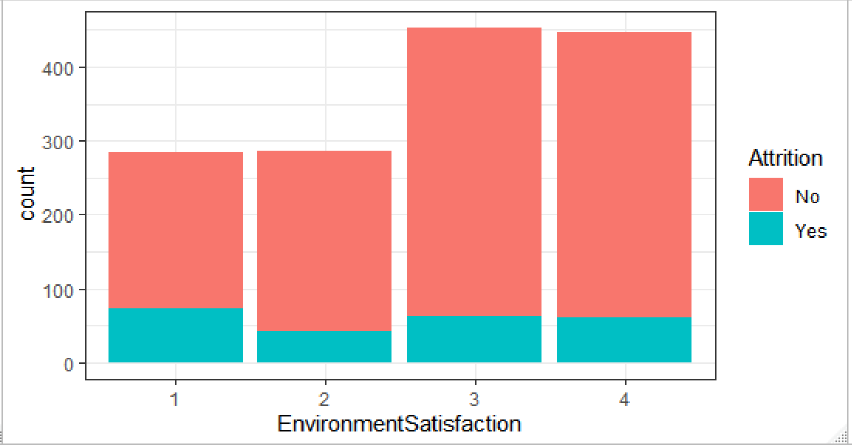


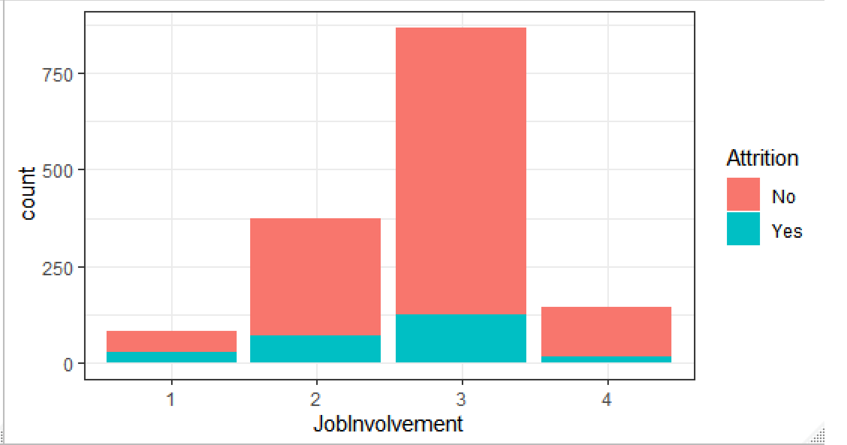


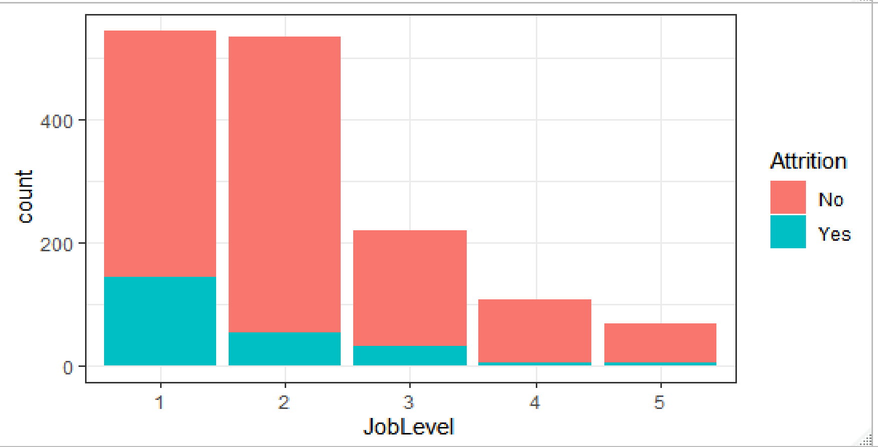


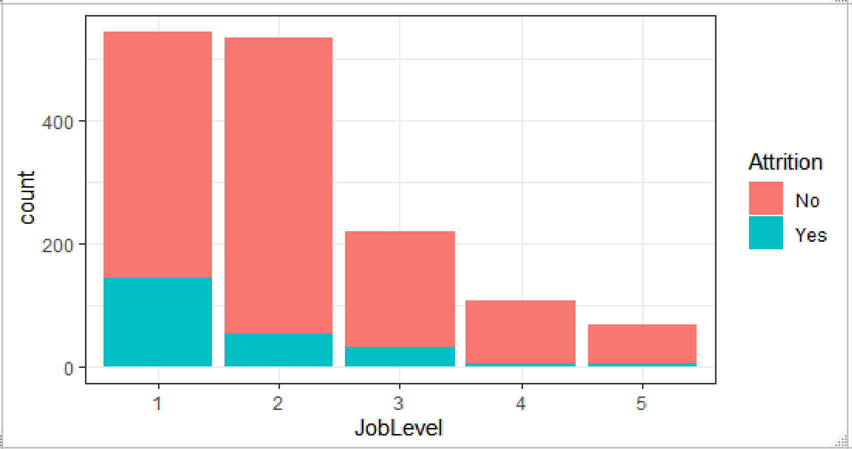
Data Visualisation Continues…

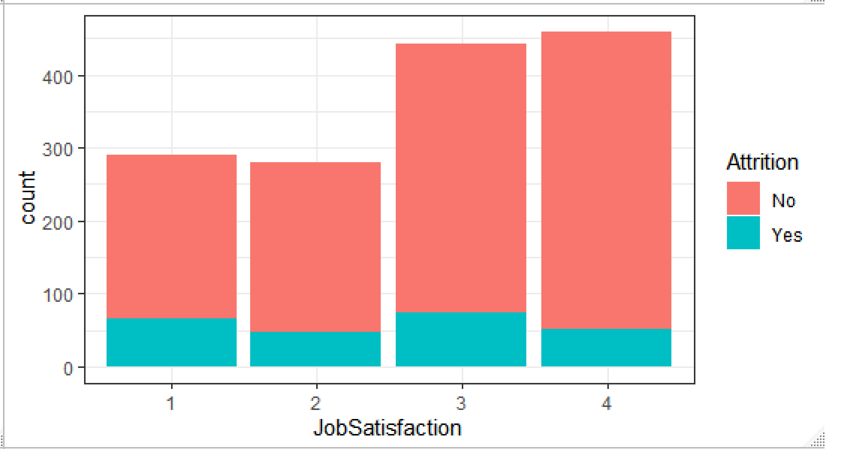


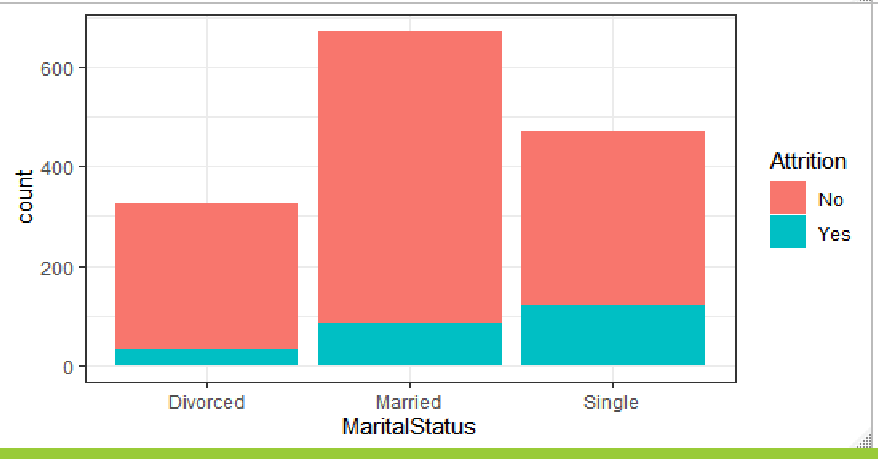


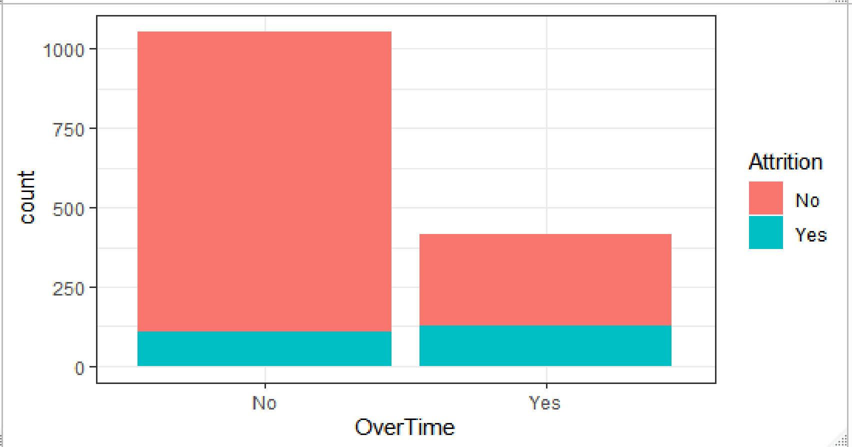


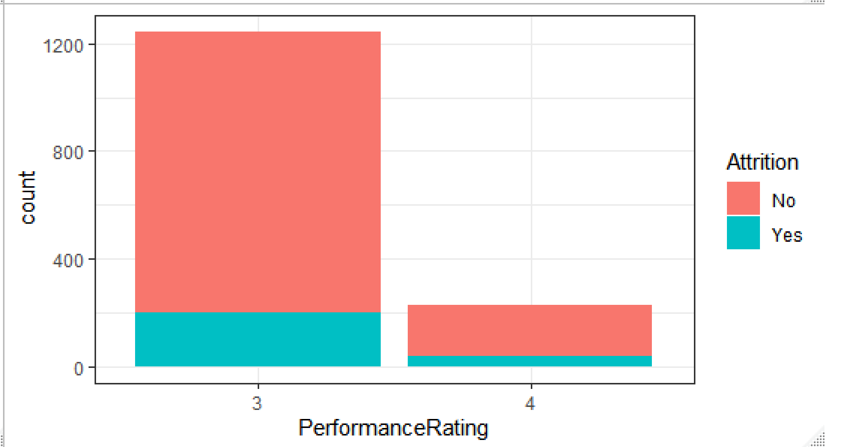




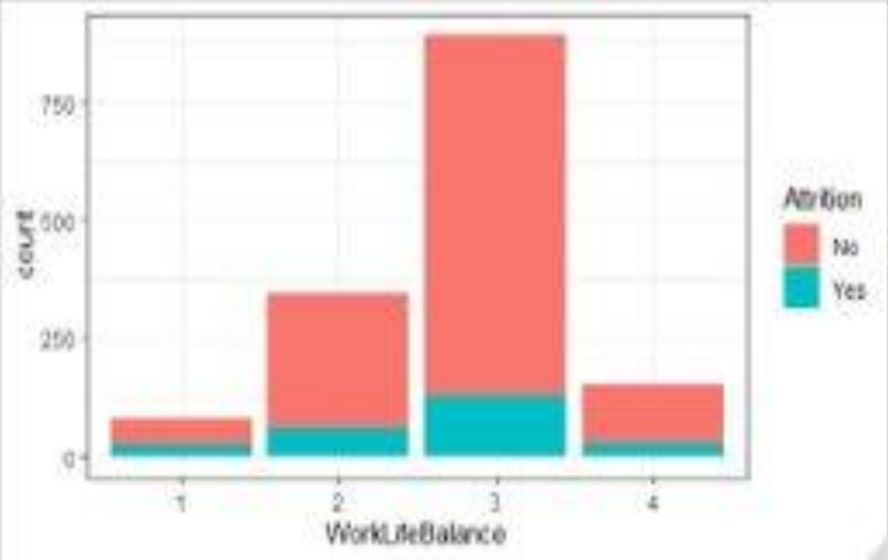


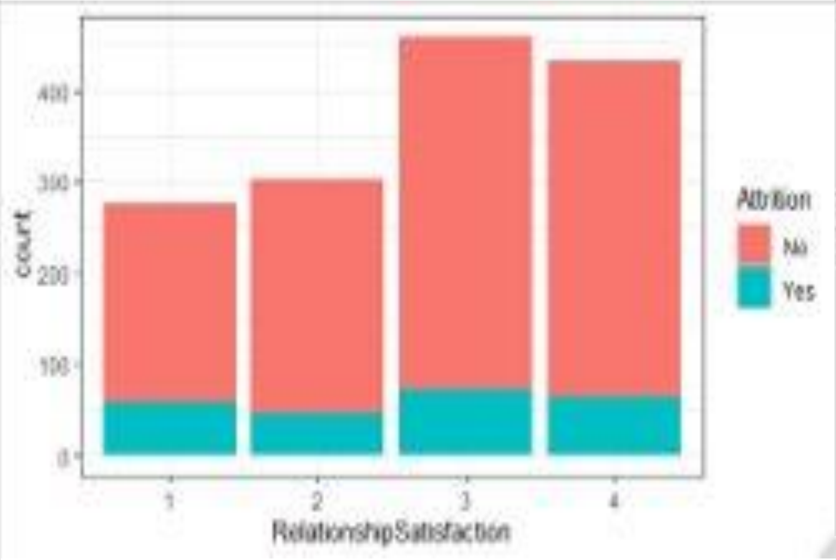


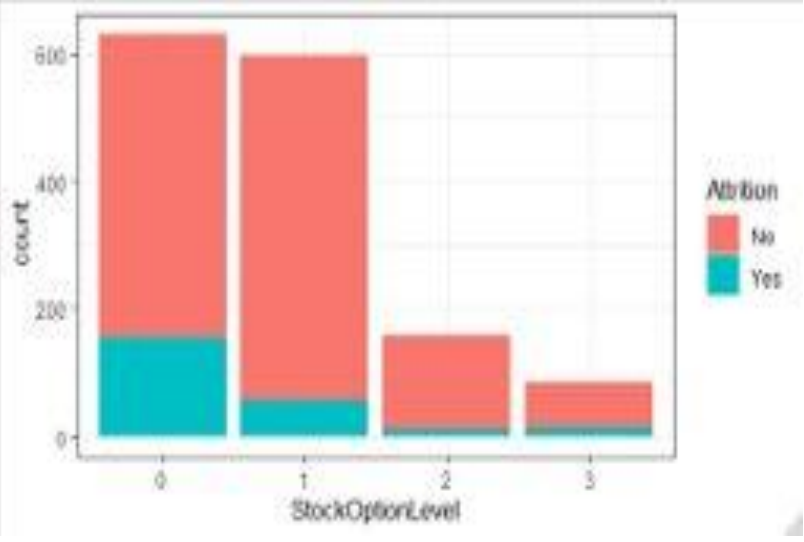




Data Visualization Continues...







EDA CONCLUSION

- The employees who are least educated are more likely(18%) to leave the company and the employees who are highly qualified are less likely (10%) to leave the company

- Lower the "Environment Satisfaction" higher the attrition rate(25%)

- Department : The worker in Research & Development are more likely to stay then the workers on other department.

- Education Field : The workers with Human Resources and Technical Degree are more likely to quit then employees from other fields of educations.

- Gender : The Male are more likely to quit.

- Job Role : The workers in Laboratory Technician, Sales Representative, and Human Resources are more likely to quit the workers in other positions.

- Marital Status : The workers who have Single marital status are more likely to quit the Married, and Divorced.

- Over Time : The workers who work more hours are likely to quit than others.

Model Building

- BASELINE OF MODEL
- SPLITING DATA IN TO TRAIN TEST DATASET
- SELECTION OF ALGORITHMS
- MODEL BUILDING
- PREDICTION
- CROSS-VALIDATION
- OPTIMISATION OF OUTPUT

MODEL BASELINE

BASELINE

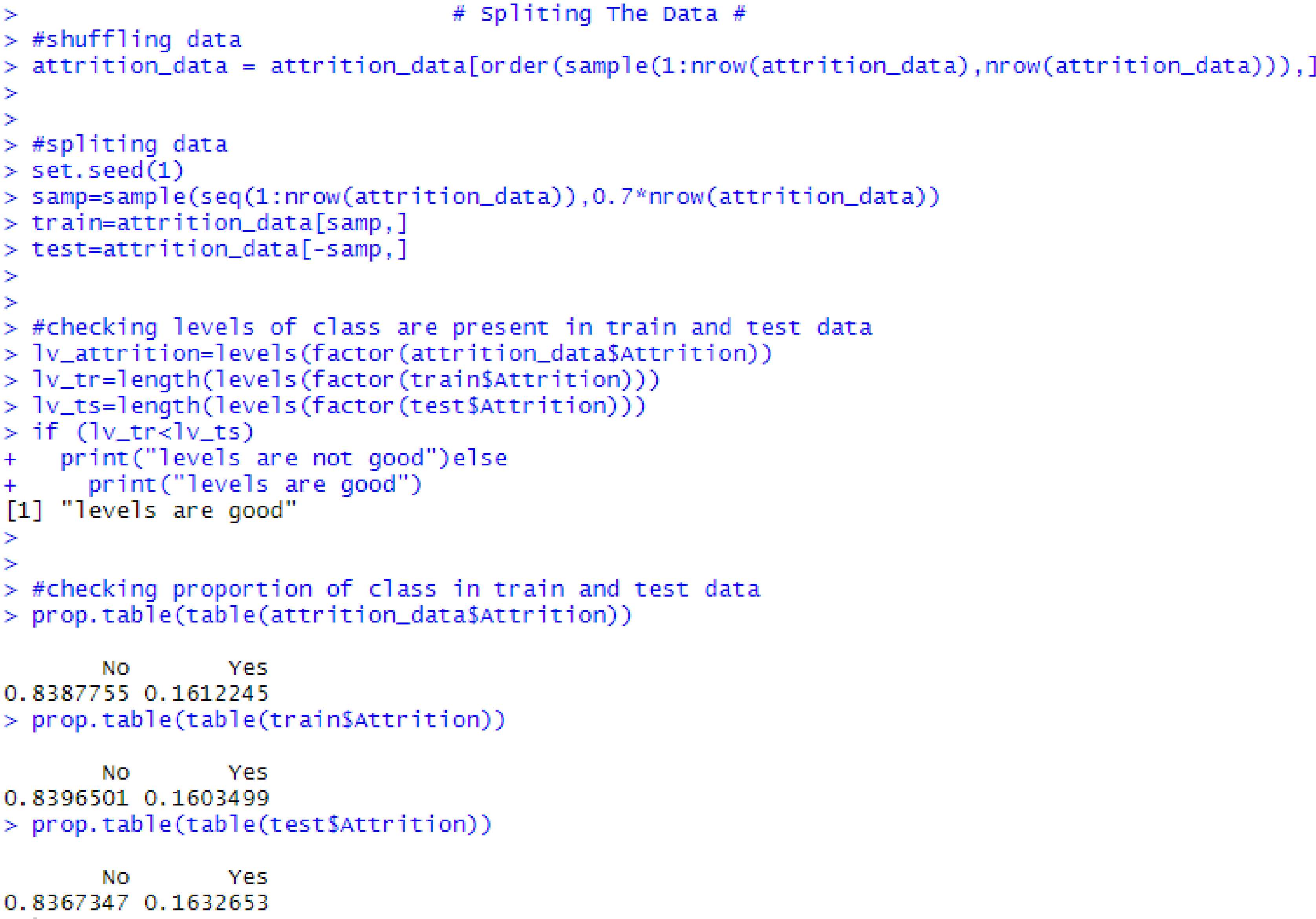
AND

PROBLEM STATEMENT

- As the Data is the Labeled data with categorical target variable supervised classification will be done to do prediction
- From data it can be seen that out of 1470 observation 1233 labeled with 'No' and 237 have churned labeled with 'Yes'
- 'No' is the most frequent outcome for all observations
- If we consider to predict 'No' as a Standard baseline of the model 1233 out of 1470 outcome would be correct with accuracy of 83.87%
- While model building 83.87% of accuracy will be taken as Baseline of model and try achieve accuracy more than 83.87%
- 'No' will be taken as positive class while building model and predicting

SPLITTING DATA TO TRAIN TEST DATASET

- Shuffling data
- Splitting the Data in the ratio of 70:30
- Ensuring levels of class are present in train and test data
- Ensuring that the target variable is well balanced in train test data set



```
>                                        # Spliting The Data #
> #shuffling data
> attrition_data = attrition_data[order(sample(1:nrow(attrition_data),nrow(attrition_data))),]
>
>
> #spliting data
> set.seed(1)
> samp=sample(seq(1:nrow(attrition_data)),0.7*nrow(attrition_data))
> train=attrition_data[samp,]
> test=attrition_data[-samp,]
>
>
> #checking levels of class are present in train and test data
> lv_attrition=levels(factor(attrition_data$Attrition))
> lv_tr=length(levels(factor(train$Attrition)))
> lv_ts=length(levels(factor(test$Attrition)))
> if (lv_tr<lv_ts)
+   print("levels are not good")else
+     print("levels are good")
[1] "levels are good"
>
>
> #checking proportion of class in train and test data
> prop.table(table(attrition_data$Attrition))

       No       Yes 
0.8387755 0.1612245 
> prop.table(table(train$Attrition))

       No       Yes 
0.8396501 0.1603499 
> prop.table(table(test$Attrition))

       No       Yes 
0.8367347 0.1632653 
```

SELECTION OF ALGORITHMS

Logistic Regression

Decision Tree

Random forest

Logistic Regression

- Easier **to implement**
- Very good **Discrimination Tool**
- Performs **well** with the dataset is **linearly separable**
- Gives a **measure of how relevant a predictor** is
- Also gives **direction of association** (positive or negative)
- Can derive **confidence level** about its prediction

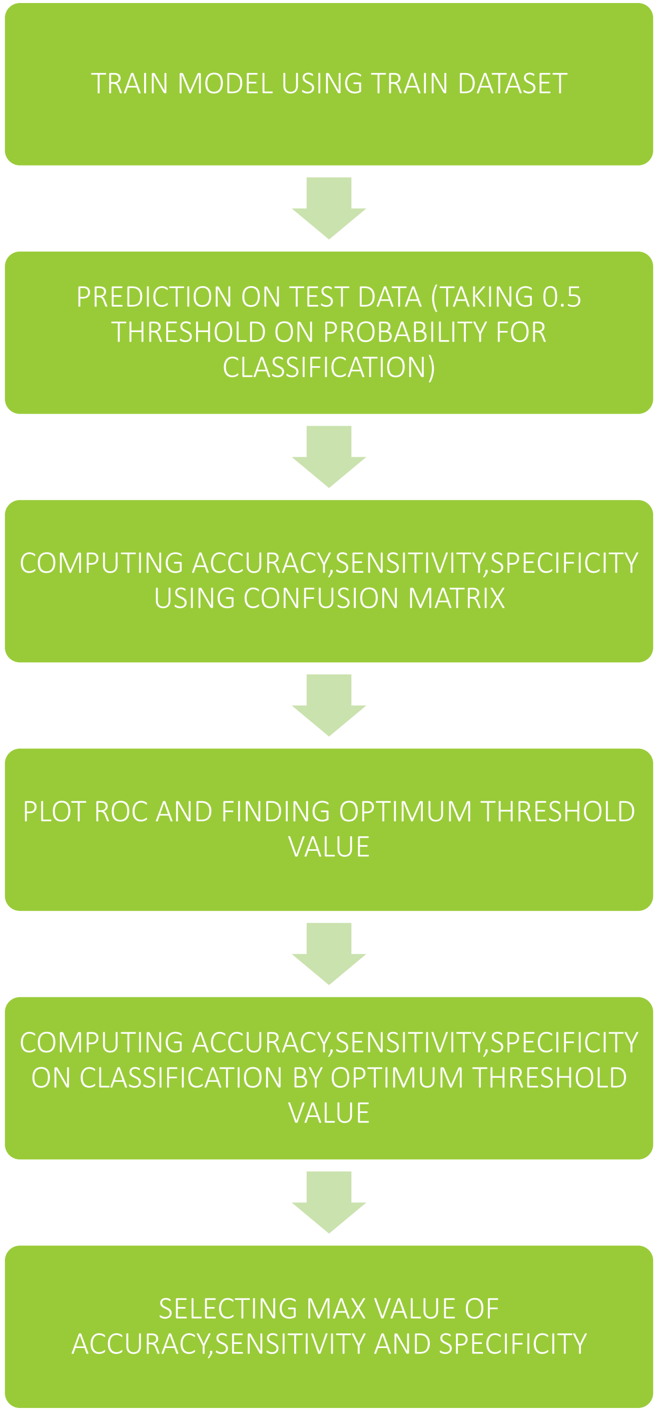
Decision Tree

- Works good **when** there is large set of categorical values in Data
- **No preprocessing** needed
- No assumptions on distribution of data
- Supports **automatic feature interaction** whereas KNN cant
- it deals collinearity better than SVM
- No need of Scaling the data

Random Forest

- Very **stable**
- Uses **Ensemble Learning** technique
- Is more robust and accurate
- Works **well** with both categorical and continuous variables
- No feature scaling required
- It **reduces overfitting** problem and **the variance**

MODEL BUILDING



TRAIN MODEL USING TRAIN DATASET

PREDICTION ON TEST DATA (TAKING 0.5 THRESHOLD ON PROBABILITY FOR CLASSIFICATION)

COMPUTING ACCURACY,SENSITIVITY,SPECIFICITY USING CONFUSION MATRIX

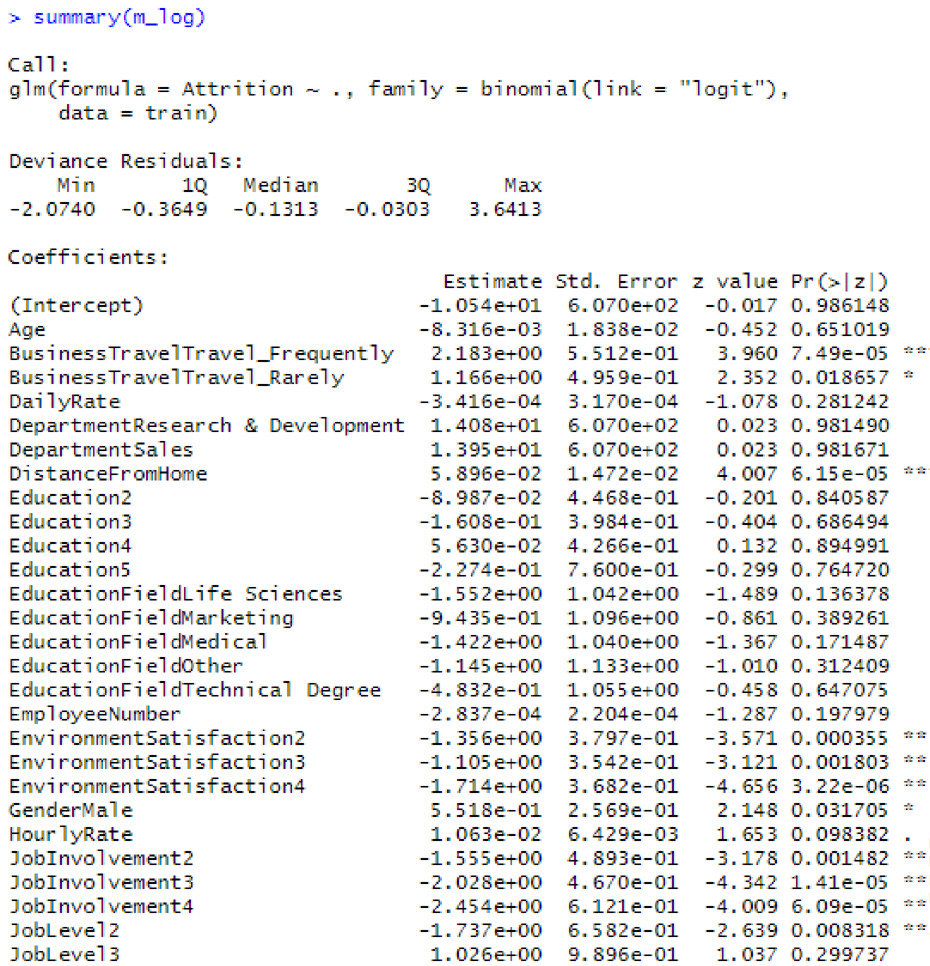
PLOT ROC AND FINDING OPTIMUM THRESHOLD VALUE

COMPUTING ACCURACY,SENSITIVITY,SPECIFICITY ON CLASSIFICATION BY OPTIMUM THRESHOLD VALUE

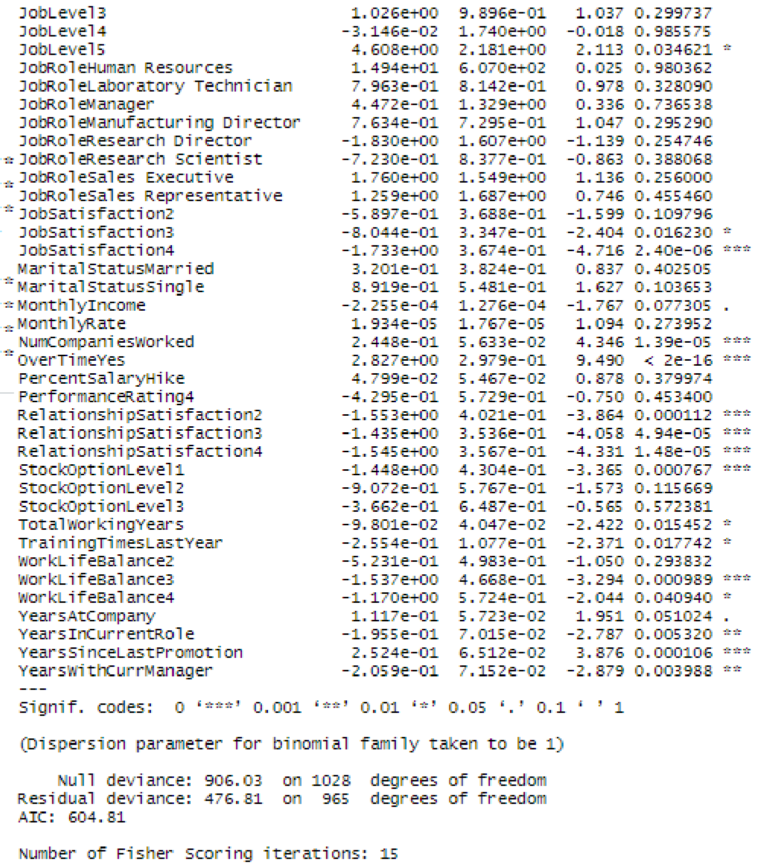
SELECTING MAX VALUE OF ACCURACY,SENSITIVITY AND SPECIFICITY

MODEL 1 Logistic Regression

- Name -> m_log
- Null deviance: 906.03 on 1028 DoF
- Residual deviance: 476.81 on 965 DoF
- AIC: 604.81



```
> summary(m_log)

Call:
glm(formula = Attrition ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-2.0740  -0.3649  -0.1313  -0.0303   3.6413

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.054e+01  6.070e+02  -0.017 0.986148
Age                             -8.316e-03  1.838e-02  -0.452 0.651019
BusinessTravelTravel_Frequently  2.183e+00  5.512e-01   3.960 7.49e-05 ***
BusinessTravelTravel_Rarely      1.166e+00  4.959e-01   2.352 0.018657 *
DailyRate                       -3.416e-04  3.170e-04  -1.078 0.281242
DepartmentResearch & Development 1.408e+01  6.070e+02   0.023 0.981490
DepartmentSales                  1.395e+01  6.070e+02   0.023 0.981671
DistanceFromHome                 5.896e-02  1.472e-02   4.007 6.15e-05 ***
Education2                      -8.987e-02  4.468e-01  -0.201 0.840587
Education3                      -1.608e-01  3.984e-01  -0.404 0.686494
Education4                       5.630e-02  4.266e-01   0.132 0.894991
Education5                      -2.274e-01  7.600e-01  -0.299 0.764720
EducationFieldLife Sciences     -1.552e+00  1.042e+00  -1.489 0.136378
EducationFieldMarketing         -9.435e-01  1.096e+00  -0.861 0.389261
EducationFieldMedical           -1.422e+00  1.040e+00  -1.367 0.171487
EducationFieldOther             -1.145e+00  1.133e+00  -1.010 0.312409
EducationFieldTechnical Degree  -4.832e-01  1.055e+00  -0.458 0.647075
EmployeeNumber                  -2.837e-04  2.204e-04  -1.287 0.197979
EnvironmentSatisfaction2        -1.356e+00  3.797e-01  -3.571 0.000355 **
EnvironmentSatisfaction3        -1.105e+00  3.542e-01  -3.121 0.001803 **
EnvironmentSatisfaction4        -1.714e+00  3.682e-01  -4.656 3.22e-06 **
GenderMale                       5.518e-01  2.569e-01   2.148 0.031705 *
HourlyRate                       1.063e-02  6.429e-03   1.653 0.098382 .
JobInvolvement2                 -1.555e+00  4.893e-01  -3.178 0.001482 **
JobInvolvement3                 -2.028e+00  4.670e-01  -4.342 1.41e-05 **
JobInvolvement4                 -2.454e+00  6.121e-01  -4.009 6.09e-05 **
JobLevel2                       -1.737e+00  6.582e-01  -2.639 0.008318 **
JobLevel3                        1.026e+00  9.896e-01   1.037 0.299737
```

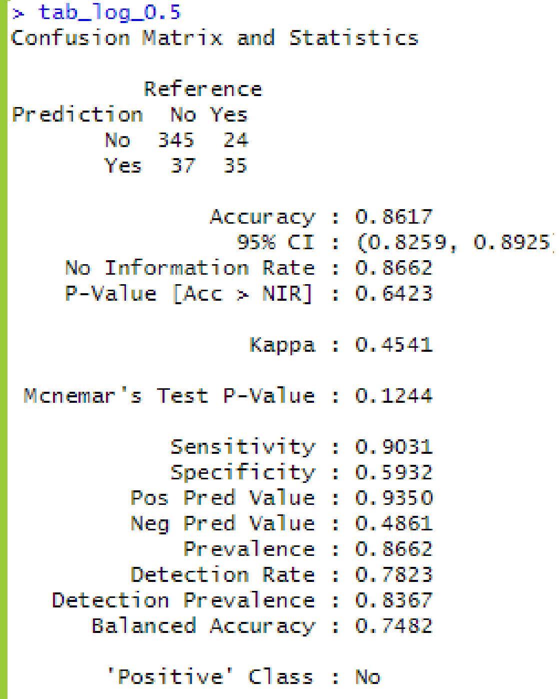


```
JobLevel3                        1.026e+00  9.896e-01   1.037 0.299737
JobLevel4                       -3.146e-02  1.740e+00  -0.018 0.985575
JobLevel5                        4.608e+00  2.181e+00   2.113 0.034621 *
JobRoleHuman Resources           1.494e+01  6.070e+02   0.025 0.980362
JobRoleLaboratory Technician     7.963e-01  8.142e-01   0.978 0.328090
JobRoleManager                   4.472e-01  1.329e+00   0.336 0.736538
JobRoleManufacturing Director    7.634e-01  7.295e-01   1.047 0.295290
JobRoleResearch Director        -1.830e+00  1.607e+00  -1.139 0.254746
JobRoleResearch Scientist       -7.230e-01  8.377e-01  -0.863 0.388068
JobRoleSales Executive           1.760e+00  1.549e+00   1.136 0.256000
JobRoleSales Representative      1.259e+00  1.687e+00   0.746 0.455460
JobSatisfaction2                -5.897e-01  3.688e-01  -1.599 0.109796
JobSatisfaction3                -8.044e-01  3.347e-01  -2.404 0.016230 *
JobSatisfaction4                -1.733e+00  3.674e-01  -4.716 2.40e-06 ***
MaritalStatusMarried             3.201e-01  3.824e-01   0.837 0.402505
MaritalStatusSingle              8.919e-01  5.481e-01   1.627 0.103653
MonthlyIncome                   -2.255e-04  1.276e-04  -1.767 0.077305 .
MonthlyRate                      1.934e-05  1.767e-05   1.094 0.273952
NumCompaniesWorked               2.448e-01  5.633e-02   4.346 1.39e-05 ***
OverTimeYes                      2.827e+00  2.979e-01   9.490  < 2e-16 ***
PercentSalaryHike                4.799e-02  5.467e-02   0.878 0.379974
PerformanceRating4              -4.295e-01  5.729e-01  -0.750 0.453400
RelationshipSatisfaction2       -1.553e+00  4.021e-01  -3.864 0.000112 ***
RelationshipSatisfaction3       -1.435e+00  3.536e-01  -4.058 4.94e-05 ***
RelationshipSatisfaction4       -1.545e+00  3.567e-01  -4.331 1.48e-05 ***
StockOptionLevel1               -1.448e+00  4.304e-01  -3.365 0.000767 ***
StockOptionLevel2               -9.072e-01  5.767e-01  -1.573 0.115669
StockOptionLevel3               -3.662e-01  6.487e-01  -0.565 0.572381
TotalWorkingYears               -9.801e-02  4.047e-02  -2.422 0.015452 *
TrainingTimesLastYear           -2.554e-01  1.077e-01  -2.371 0.017742 *
WorkLifeBalance2                -5.231e-01  4.983e-01  -1.050 0.293832
WorkLifeBalance3                -1.537e+00  4.668e-01  -3.294 0.000989 ***
WorkLifeBalance4                -1.170e+00  5.724e-01  -2.044 0.040940 *
YearsAtCompany                   1.117e-01  5.723e-02   1.951 0.051024 .
YearsInCurrentRole              -1.955e-01  7.015e-02  -2.787 0.005320 **
YearsSinceLastPromotion          2.524e-01  6.512e-02   3.876 0.000106 ***
YearsWithCurrManager            -2.059e-01  7.152e-02  -2.879 0.003988 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 906.03  on 1028  degrees of freedom
Residual deviance: 476.81  on  965  degrees of freedom
AIC: 604.81

Number of Fisher Scoring iterations: 15
```

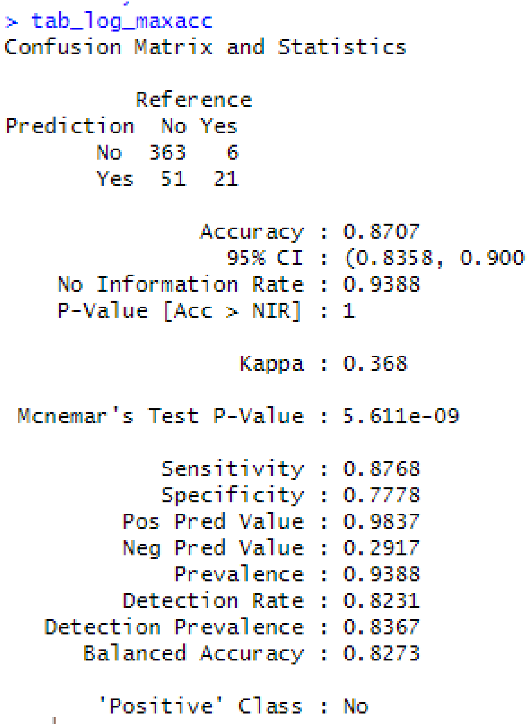
Prediction on Test data and Computing Accuracy,Sensitivity and Specificity
(Confusion Matrix)

FOR LOGISTIC REGRESSION MODEL WITH ALL VERIABLE

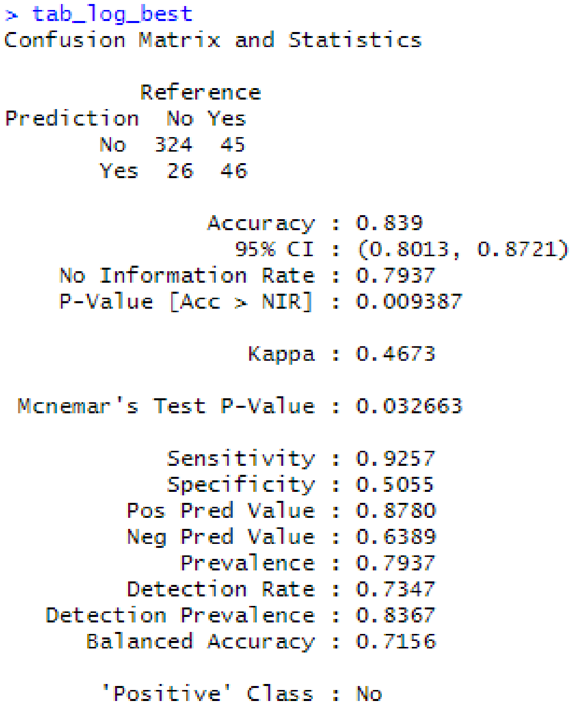
- If cutoff =0.50 Accuracy : 0.8617 Sensitivity : 0.9031 Specificity : 0.5932
- If when cutoff =0.75 Accuracy : 0.8707 Sensitivity : 0.8768 Specificity : 0.7778 #for best Accuracy***
- If when cutoff =0.30 Accuracy : 0.839 Sensitivity : 0.9275 Specificity : 0.5055 # from ROCR
- AUC = 0.8157



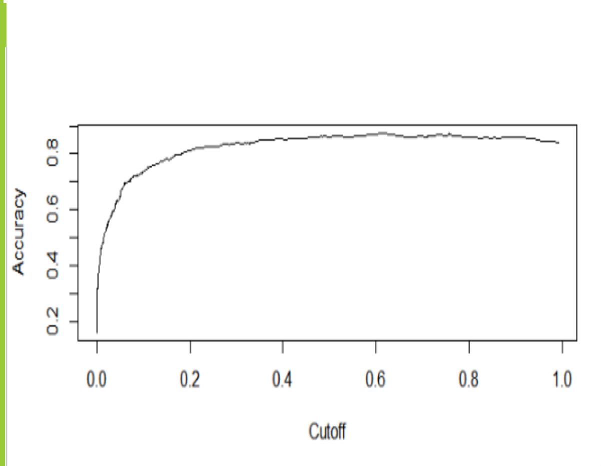
```
> tab_log_0.5
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  345  24
       Yes  37  35

               Accuracy : 0.8617
                 95% CI : (0.8259, 0.8925
    No Information Rate : 0.8662
    P-Value [Acc > NIR] : 0.6423

                  Kappa : 0.4541

 Mcnemar's Test P-Value : 0.1244

            Sensitivity : 0.9031
            Specificity : 0.5932
         Pos Pred Value : 0.9350
         Neg Pred Value : 0.4861
             Prevalence : 0.8662
         Detection Rate : 0.7823
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.7482

       'Positive' Class : No
```

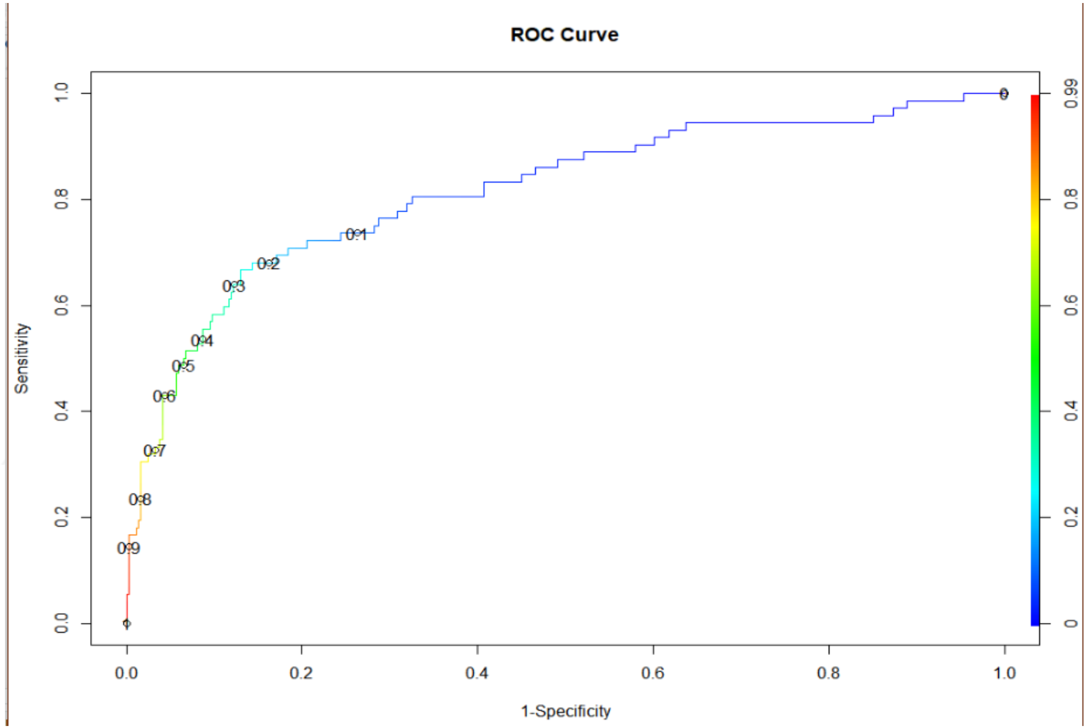


```
> tab_log_maxacc
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  363   6
       Yes  51  21

               Accuracy : 0.8707
                 95% CI : (0.8358, 0.900
    No Information Rate : 0.9388
    P-Value [Acc > NIR] : 1

                  Kappa : 0.368

 Mcnemar's Test P-Value : 5.611e-09

            Sensitivity : 0.8768
            Specificity : 0.7778
         Pos Pred Value : 0.9837
         Neg Pred Value : 0.2917
             Prevalence : 0.9388
         Detection Rate : 0.8231
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.8273

       'Positive' Class : No
```



```
> tab_log_best
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  324  45
       Yes  26  46

               Accuracy : 0.839
                 95% CI : (0.8013, 0.8721)
    No Information Rate : 0.7937
    P-Value [Acc > NIR] : 0.009387

                  Kappa : 0.4673

 Mcnemar's Test P-Value : 0.032663

            Sensitivity : 0.9257
            Specificity : 0.5055
         Pos Pred Value : 0.8780
         Neg Pred Value : 0.6389
             Prevalence : 0.7937
         Detection Rate : 0.7347
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.7156

       'Positive' Class : No
```




ROC Curve

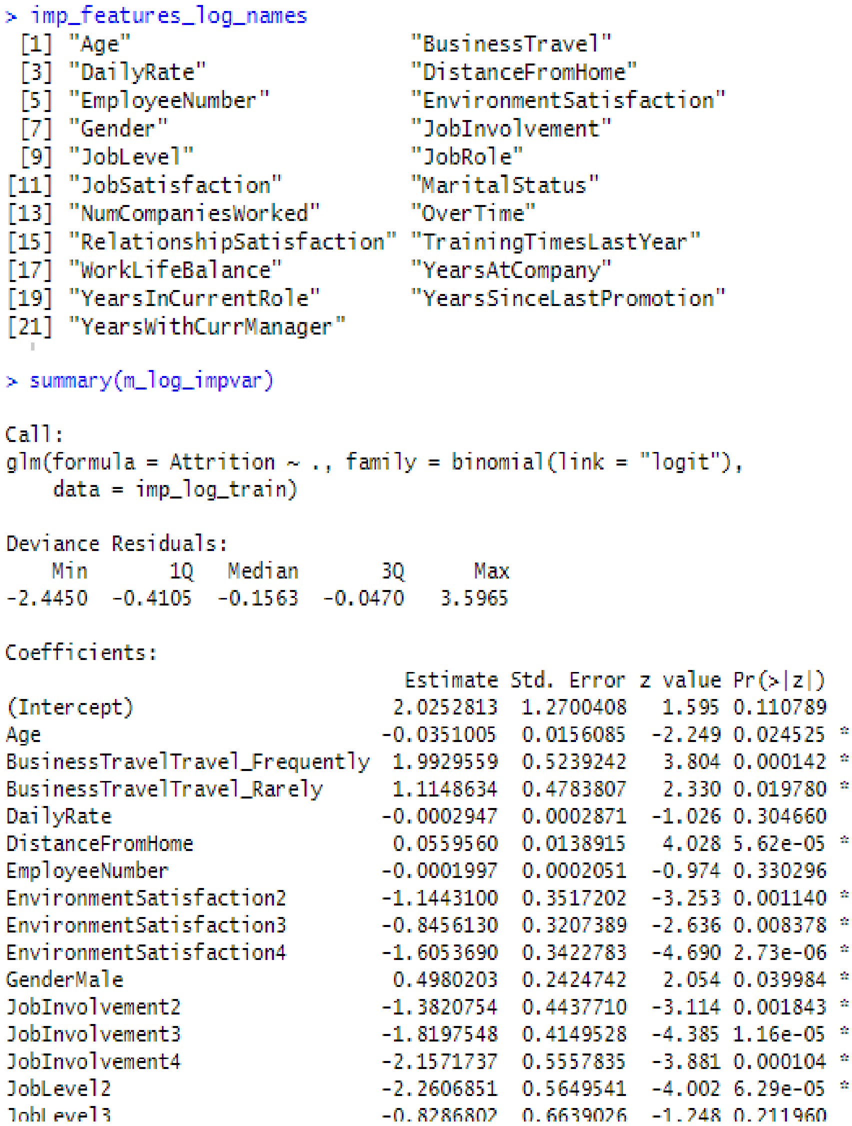
MODEL 2
Logistic Regression
(With important variable)

Selecting important Features for model based on p-value,in order to reduce the complexity of the model

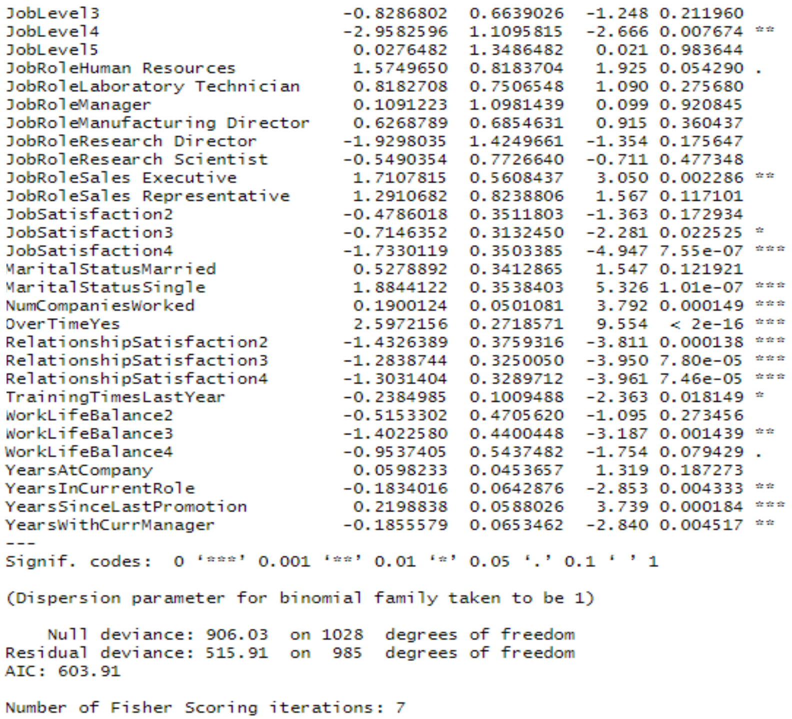
Null deviance: 906.03 on 1028 degrees of freedom

Residual deviance: 515.91 on 985 degrees of freedom

AIC: 603.91



```
> imp_features_log_names
 [1] "Age"                      "BusinessTravel"
 [3] "DailyRate"                "DistanceFromHome"
 [5] "EmployeeNumber"           "EnvironmentSatisfaction"
 [7] "Gender"                   "JobInvolvement"
 [9] "JobLevel"                 "JobRole"
[11] "JobSatisfaction"          "MaritalStatus"
[13] "NumCompaniesWorked"       "OverTime"
[15] "RelationshipSatisfaction" "TrainingTimesLastYear"
[17] "WorkLifeBalance"          "YearsAtCompany"
[19] "YearsInCurrentRole"       "YearsSinceLastPromotion"
[21] "YearsWithCurrManager"

> summary(m_log_impvar)

Call:
glm(formula = Attrition ~ ., family = binomial(link = "logit"),
    data = imp_log_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4450  -0.4105  -0.1563  -0.0470   3.5965

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.0252813  1.2700408   1.595 0.110789
Age                         -0.0351005  0.0156085  -2.249 0.024525 *
BusinessTravelTravel_Frequently 1.9929559  0.5239242   3.804 0.000142 ***
BusinessTravelTravel_Rarely  1.1148634  0.4783807   2.330 0.019780 *
DailyRate                   -0.0002947  0.0002871  -1.026 0.304660
DistanceFromHome             0.0559560  0.0138915   4.028 5.62e-05 ***
EmployeeNumber              -0.0001997  0.0002051  -0.974 0.330296
EnvironmentSatisfaction2    -1.1443100  0.3517202  -3.253 0.001140 **
EnvironmentSatisfaction3    -0.8456130  0.3207389  -2.636 0.008378 **
EnvironmentSatisfaction4    -1.6053690  0.3422783  -4.690 2.73e-06 ***
GenderMale                   0.4980203  0.2424742   2.054 0.039984 *
JobInvolvement2             -1.3820754  0.4437710  -3.114 0.001843 **
JobInvolvement3             -1.8197548  0.4149528  -4.385 1.16e-05 ***
JobInvolvement4             -2.1571737  0.5557835  -3.881 0.000104 ***
JobLevel2                   -2.2606851  0.5649541  -4.002 6.29e-05 ***
JobLevel3                   -0.8286802  0.6639026  -1.248 0.211960
```



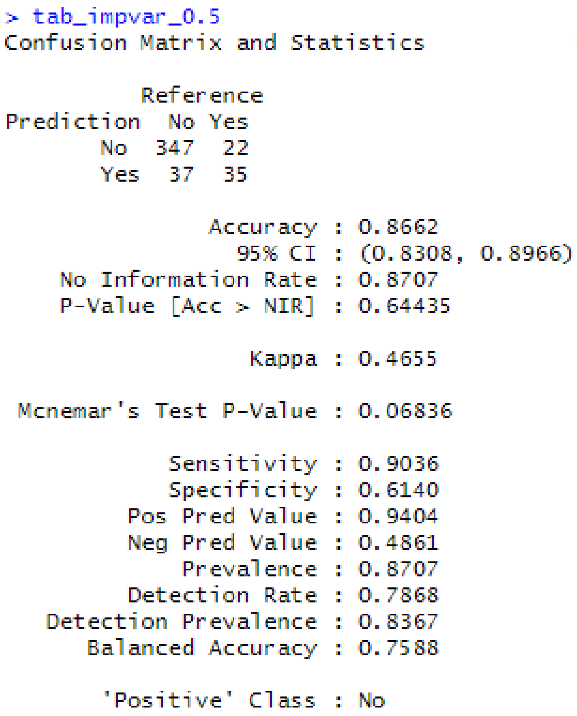
```
JobLevel3                       -0.8286802  0.6639026  -1.248 0.211960
JobLevel4                       -2.9582596  1.1095815  -2.666 0.007674 **
JobLevel5                        0.0276482  1.3486482   0.021 0.983644
JobRoleHuman Resources           1.5749650  0.8183704   1.925 0.054290 .
JobRoleLaboratory Technician     0.8182708  0.7506548   1.090 0.275680
JobRoleManager                   0.1091223  1.0981439   0.099 0.920845
JobRoleManufacturing Director    0.6268789  0.6854631   0.915 0.360437
JobRoleResearch Director        -1.9298035  1.4249661  -1.354 0.175647
JobRoleResearch Scientist       -0.5490354  0.7726640  -0.711 0.477348
JobRoleSales Executive           1.7107815  0.5608437   3.050 0.002286 **
JobRoleSales Representative      1.2910682  0.8238806   1.567 0.117101
JobSatisfaction2                -0.4786018  0.3511803  -1.363 0.172934
JobSatisfaction3                -0.7146352  0.3132450  -2.281 0.022525 *
JobSatisfaction4                -1.7330119  0.3503385  -4.947 7.55e-07 ***
MaritalStatusMarried             0.5278892  0.3412865   1.547 0.121921
MaritalStatusSingle              1.8844122  0.3538403   5.326 1.01e-07 ***
NumCompaniesWorked               0.1900124  0.0501081   3.792 0.000149 ***
OverTimeYes                      2.5972156  0.2718571   9.554  < 2e-16 ***
RelationshipSatisfaction2       -1.4326389  0.3759316  -3.811 0.000138 ***
RelationshipSatisfaction3       -1.2838744  0.3250050  -3.950 7.80e-05 ***
RelationshipSatisfaction4       -1.3031404  0.3289712  -3.961 7.46e-05 ***
TrainingTimesLastYear           -0.2384985  0.1009488  -2.363 0.018149 *
WorkLifeBalance2                -0.5153302  0.4705620  -1.095 0.273456
WorkLifeBalance3                -1.4022580  0.4400448  -3.187 0.001439 **
WorkLifeBalance4                -0.9537405  0.5437482  -1.754 0.079429 .
YearsAtCompany                   0.0598233  0.0453657   1.319 0.187273
YearsInCurrentRole              -0.1834016  0.0642876  -2.853 0.004333 **
YearsSinceLastPromotion          0.2198838  0.0588026   3.739 0.000184 ***
YearsWithCurrManager            -0.1855579  0.0653462  -2.840 0.004517 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 906.03  on 1028  degrees of freedom
Residual deviance: 515.91  on  985  degrees of freedom
AIC: 603.91

Number of Fisher Scoring iterations: 7
```

Prediction on Test data and Computing Accuracy, Sensitivity and Specificity

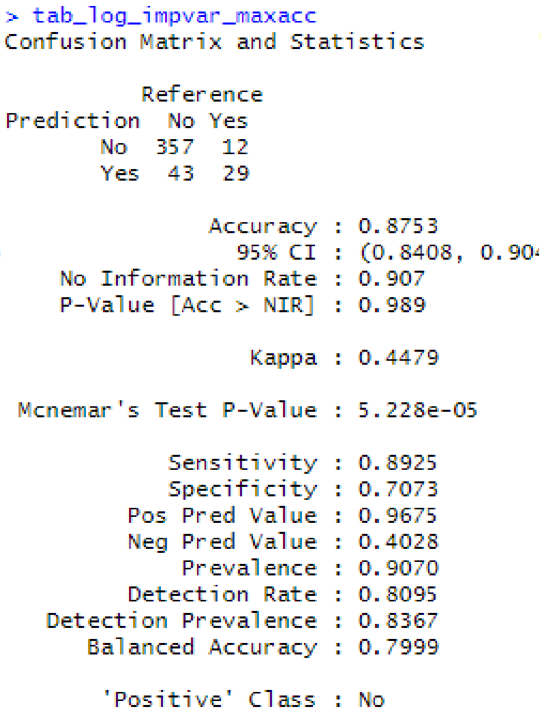
(Confusion Matrix)

FOR LOGISTIC REGESSTION WITH IMPORTANT VARIABLE

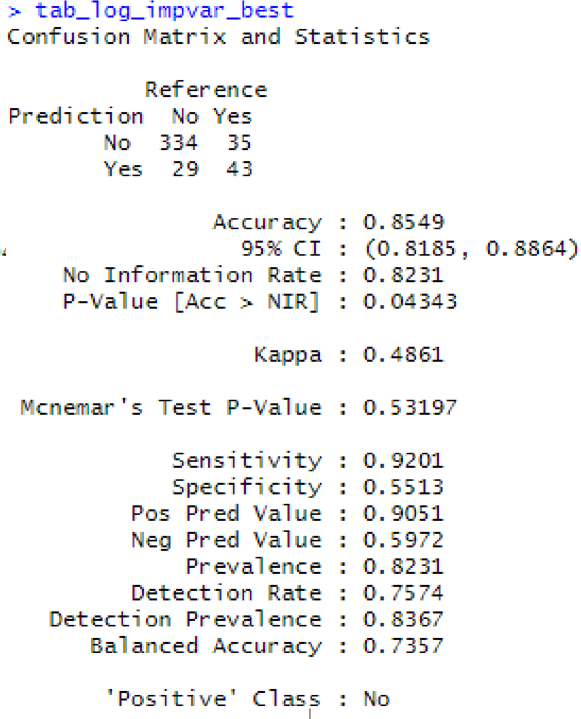
- If CUTOFF :0.5 Accuracy : 0.8662 Sensitivity : 0.9036 Specificity : 0.6140
- If CUTOFF :0.6 Accuracy : 0.8753 Sensitivity : 0.8925 Specificity : 0.7073 for maximum acc
- If CUTOFF :0.3 Accuracy : 0.8549 Sensitivity : 0.9201 Specificity : 0.5513***
- AUC = 0.8504



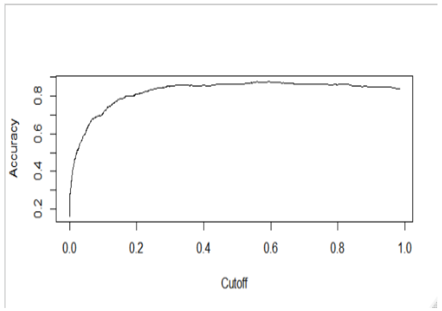
```
> tab_impvar_0.5
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  347  22
       Yes  37  35

               Accuracy : 0.8662
                 95% CI : (0.8308, 0.8966)
    No Information Rate : 0.8707
    P-Value [Acc > NIR] : 0.64435

                  Kappa : 0.4655

 Mcnemar's Test P-Value : 0.06836

            Sensitivity : 0.9036
            Specificity : 0.6140
         Pos Pred Value : 0.9404
         Neg Pred Value : 0.4861
             Prevalence : 0.8707
         Detection Rate : 0.7868
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.7588

       'Positive' Class : No
```

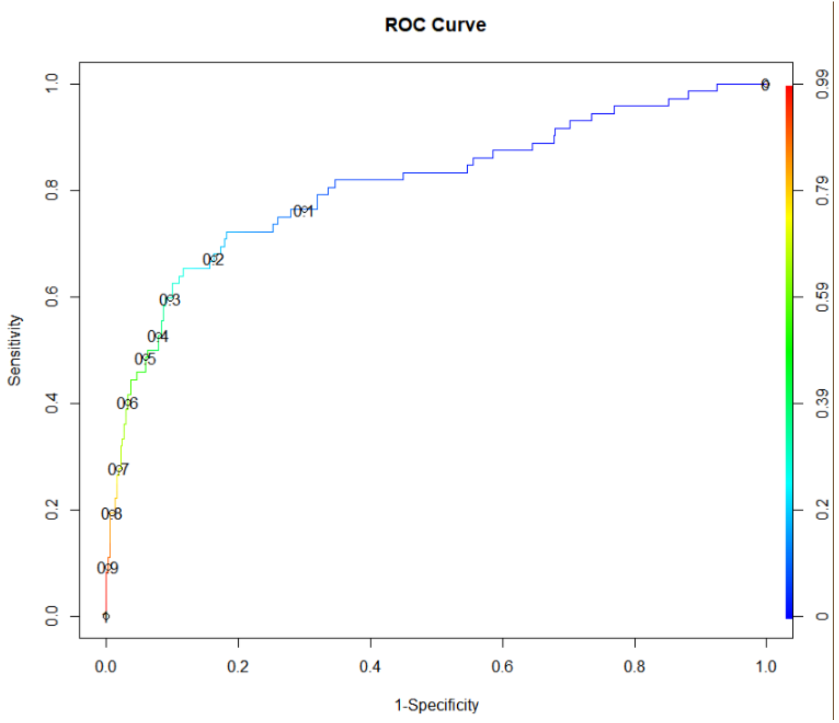


```
> tab_log_impvar_maxacc
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  357  12
       Yes  43  29

               Accuracy : 0.8753
                 95% CI : (0.8408, 0.90
    No Information Rate : 0.907
    P-Value [Acc > NIR] : 0.989

                  Kappa : 0.4479

 Mcnemar's Test P-Value : 5.228e-05

            Sensitivity : 0.8925
            Specificity : 0.7073
         Pos Pred Value : 0.9675
         Neg Pred Value : 0.4028
             Prevalence : 0.9070
         Detection Rate : 0.8095
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.7999

       'Positive' Class : No
```



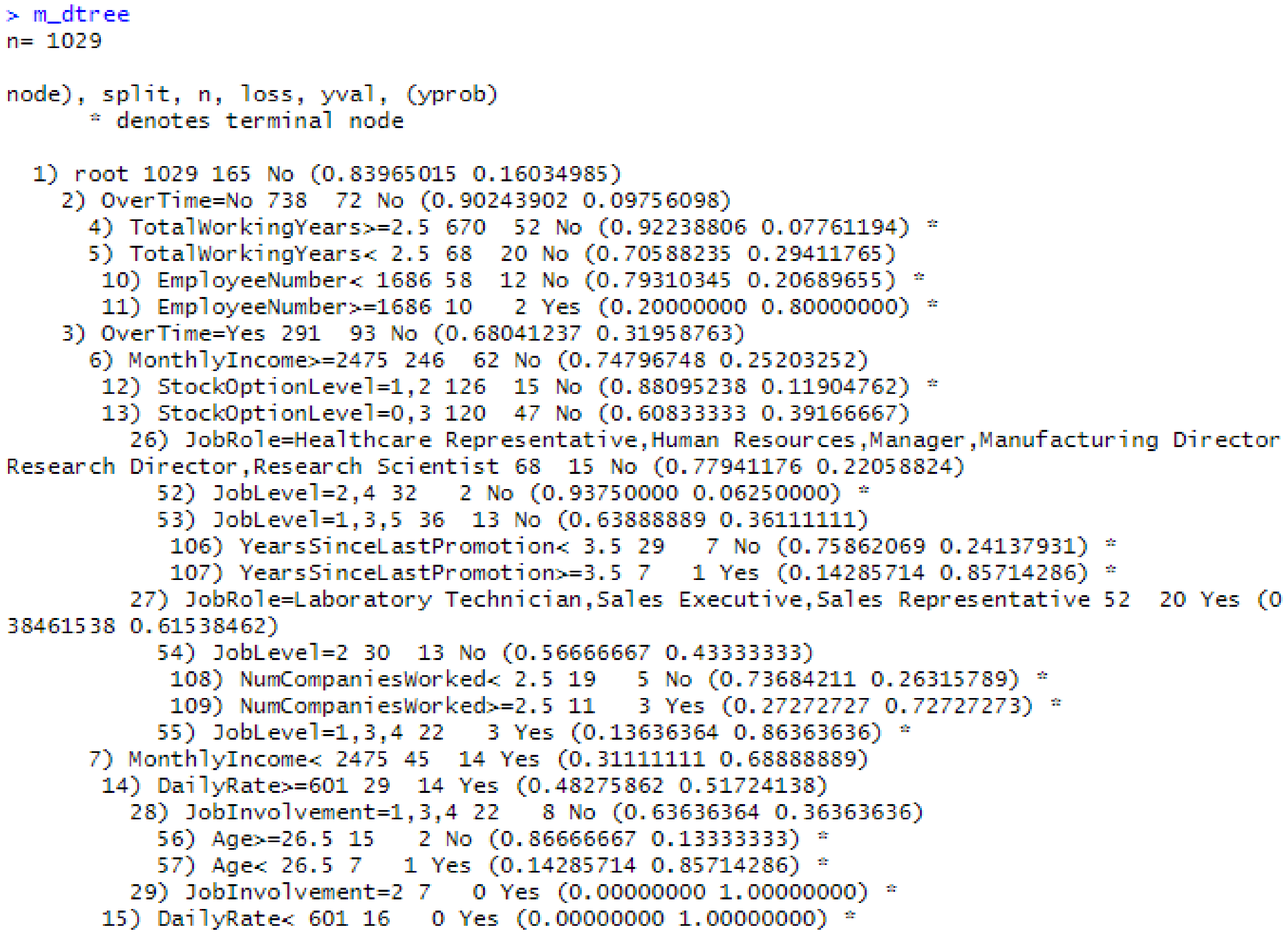
```
> tab_log_impvar_best
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  334  35
       Yes  29  43

               Accuracy : 0.8549
                 95% CI : (0.8185, 0.8864)
    No Information Rate : 0.8231
    P-Value [Acc > NIR] : 0.04343

                  Kappa : 0.4861

 Mcnemar's Test P-Value : 0.53197

            Sensitivity : 0.9201
            Specificity : 0.5513
         Pos Pred Value : 0.9051
         Neg Pred Value : 0.5972
             Prevalence : 0.8231
         Detection Rate : 0.7574
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.7357

       'Positive' Class : No
```





MODEL 3
Decision tree

- Node number 1 = 1029 obs.
- complexity param=0.05151515
- Predicted class = No
- expected loss=0.1603499
- P(node) =1
- class counts: 864   165
- probabilities: 0.840   0.160



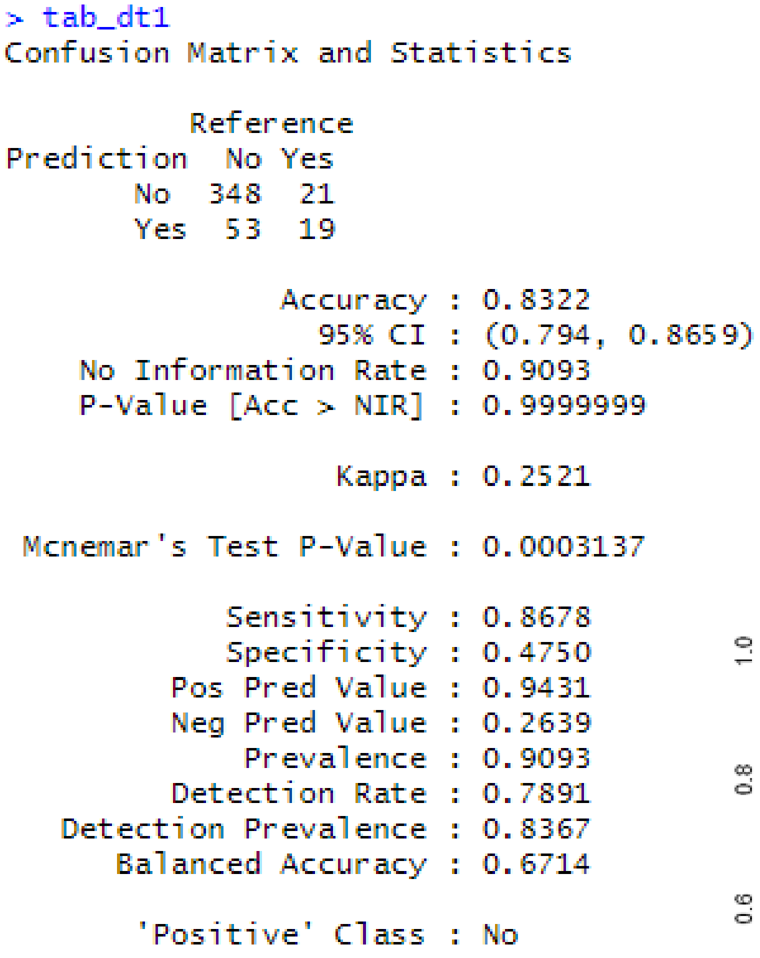
```
> m_dtree
n= 1029

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 1029 165 No (0.83965015 0.16034985)
    2) OverTime=No 738  72 No (0.90243902 0.09756098)
      4) TotalWorkingYears>=2.5 670  52 No (0.92238806 0.07761194) *
      5) TotalWorkingYears< 2.5 68  20 No (0.70588235 0.29411765)
       10) EmployeeNumber< 1686 58  12 No (0.79310345 0.20689655) *
       11) EmployeeNumber>=1686 10   2 Yes (0.20000000 0.80000000) *
    3) OverTime=Yes 291  93 No (0.68041237 0.31958763)
      6) MonthlyIncome>=2475 246  62 No (0.74796748 0.25203252)
       12) StockOptionLevel=1,2 126  15 No (0.88095238 0.11904762) *
       13) StockOptionLevel=0,3 120  47 No (0.60833333 0.39166667)
         26) JobRole=Healthcare Representative,Human Resources,Manager,Manufacturing Director
Research Director,Research Scientist 68  15 No (0.77941176 0.22058824)
           52) JobLevel=2,4 32   2 No (0.93750000 0.06250000) *
           53) JobLevel=1,3,5 36  13 No (0.63888889 0.36111111)
            106) YearsSinceLastPromotion< 3.5 29   7 No (0.75862069 0.24137931) *
            107) YearsSinceLastPromotion>=3.5 7   1 Yes (0.14285714 0.85714286) *
         27) JobRole=Laboratory Technician,Sales Executive,Sales Representative 52  20 Yes (0
38461538 0.61538462)
           54) JobLevel=2 30  13 No (0.56666667 0.43333333)
            108) NumCompaniesWorked< 2.5 19   5 No (0.73684211 0.26315789) *
            109) NumCompaniesWorked>=2.5 11   3 Yes (0.27272727 0.72727273) *
           55) JobLevel=1,3,4 22   3 Yes (0.13636364 0.86363636) *
      7) MonthlyIncome< 2475 45  14 Yes (0.31111111 0.68888889)
       14) DailyRate>=601 29  14 Yes (0.48275862 0.51724138)
         28) JobInvolvement=1,3,4 22   8 No (0.63636364 0.36363636)
           56) Age>=26.5 15   2 No (0.86666667 0.13333333) *
           57) Age< 26.5 7   1 Yes (0.14285714 0.85714286) *
         29) JobInvolvement=2 7   0 Yes (0.00000000 1.00000000) *
       15) DailyRate< 601 16   0 Yes (0.00000000 1.00000000) *
```

Prediction on Test data and Computing Accuracy,Sensitivity and Specificity

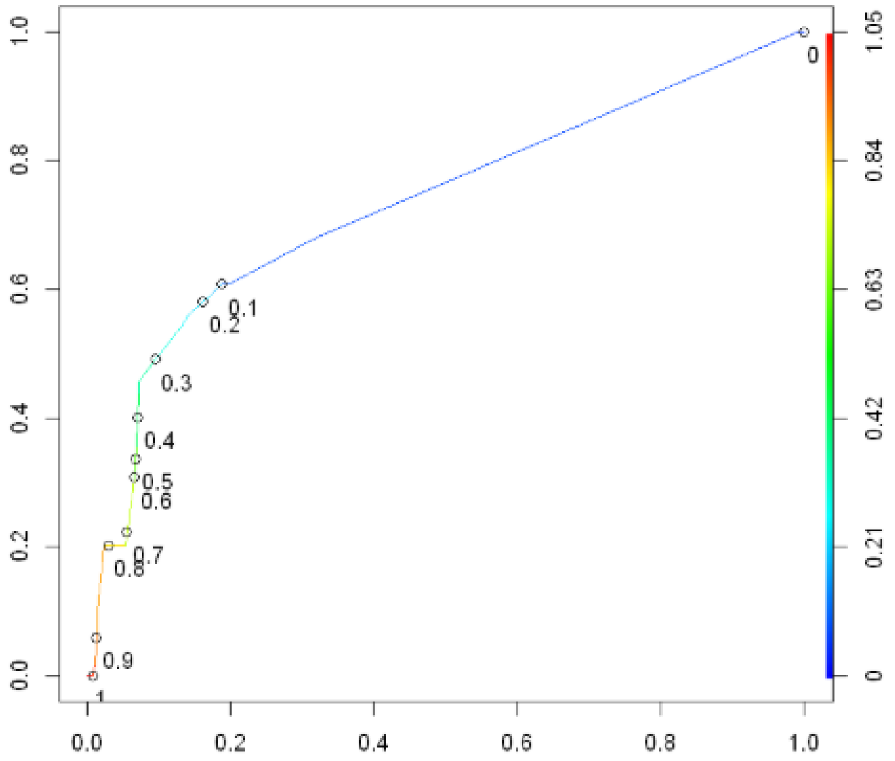
(Confusion Matrix)

FOR DECISION TREE MODELS WITH ALL VARIABLE

- Accuracy = 0.8322
- Sensitivity = 0.8678
- Specificity = 0.4750***
- AUC = 0.71

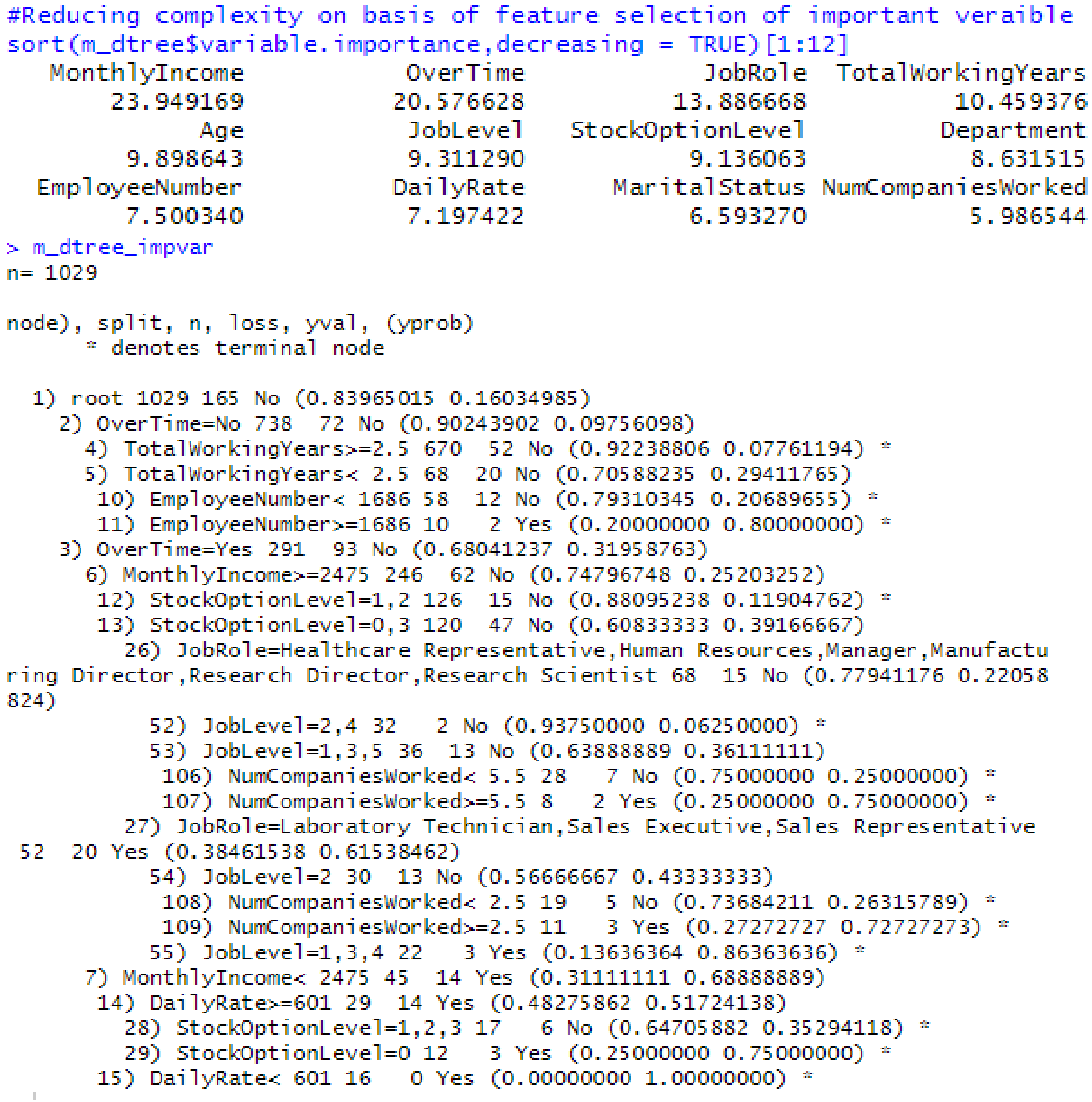


```
> tab_dt1
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  348  21
       Yes  53  19

               Accuracy : 0.8322
                 95% CI : (0.794, 0.8659)
    No Information Rate : 0.9093
    P-Value [Acc > NIR] : 0.9999999

                  Kappa : 0.2521

 Mcnemar's Test P-Value : 0.0003137

            Sensitivity : 0.8678
            Specificity : 0.4750
         Pos Pred Value : 0.9431
         Neg Pred Value : 0.2639
             Prevalence : 0.9093
         Detection Rate : 0.7891
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.6714

       'Positive' Class : No
```



MODEL 4
Decision Tree
(With important variable)

- Node number 1 = 1029 obs. complexity param=0.05151515
- predicted class=No
- expected loss=0.1603499
- P(node) =1
- class counts: 864 165
- probabilities: 0.840 0.160



```
#Reducing complexity on basis of feature selection of important veraible
sort(m_dtree$variable.importance,decreasing = TRUE)[1:12]
   MonthlyIncome           OverTime            JobRole  TotalWorkingYears
       23.949169          20.576628          13.886668          10.459376
             Age           JobLevel   StockOptionLevel         Department
        9.898643           9.311290           9.136063           8.631515
  EmployeeNumber          DailyRate      MaritalStatus NumCompaniesWorked
        7.500340           7.197422           6.593270           5.986544
> m_dtree_impvar
n= 1029

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 1029 165 No (0.83965015 0.16034985)
    2) OverTime=No 738  72 No (0.90243902 0.09756098)
      4) TotalWorkingYears>=2.5 670  52 No (0.92238806 0.07761194) *
      5) TotalWorkingYears< 2.5 68  20 No (0.70588235 0.29411765)
       10) EmployeeNumber< 1686 58  12 No (0.79310345 0.20689655) *
       11) EmployeeNumber>=1686 10   2 Yes (0.20000000 0.80000000) *
    3) OverTime=Yes 291  93 No (0.68041237 0.31958763)
      6) MonthlyIncome>=2475 246  62 No (0.74796748 0.25203252)
       12) StockOptionLevel=1,2 126  15 No (0.88095238 0.11904762) *
       13) StockOptionLevel=0,3 120  47 No (0.60833333 0.39166667)
         26) JobRole=Healthcare Representative,Human Resources,Manager,Manufactu
ring Director,Research Director,Research Scientist 68  15 No (0.77941176 0.22058
824)
           52) JobLevel=2,4 32   2 No (0.93750000 0.06250000) *
           53) JobLevel=1,3,5 36  13 No (0.63888889 0.36111111)
            106) NumCompaniesWorked< 5.5 28   7 No (0.75000000 0.25000000) *
            107) NumCompaniesWorked>=5.5 8   2 Yes (0.25000000 0.75000000) *
         27) JobRole=Laboratory Technician,Sales Executive,Sales Representative
 52  20 Yes (0.38461538 0.61538462)
           54) JobLevel=2 30  13 No (0.56666667 0.43333333)
            108) NumCompaniesWorked< 2.5 19   5 No (0.73684211 0.26315789) *
            109) NumCompaniesWorked>=2.5 11   3 Yes (0.27272727 0.72727273) *
           55) JobLevel=1,3,4 22   3 Yes (0.13636364 0.86363636) *
      7) MonthlyIncome< 2475 45  14 Yes (0.31111111 0.68888889)
       14) DailyRate>=601 29  14 Yes (0.48275862 0.51724138)
         28) StockOptionLevel=1,2,3 17   6 No (0.64705882 0.35294118) *
         29) StockOptionLevel=0 12   3 Yes (0.25000000 0.75000000) *
       15) DailyRate< 601 16   0 Yes (0.00000000 1.00000000) *
```

Prediction on Test data and Computing Accuracy, Sensitivity and Specificity (Confusion Matrix)

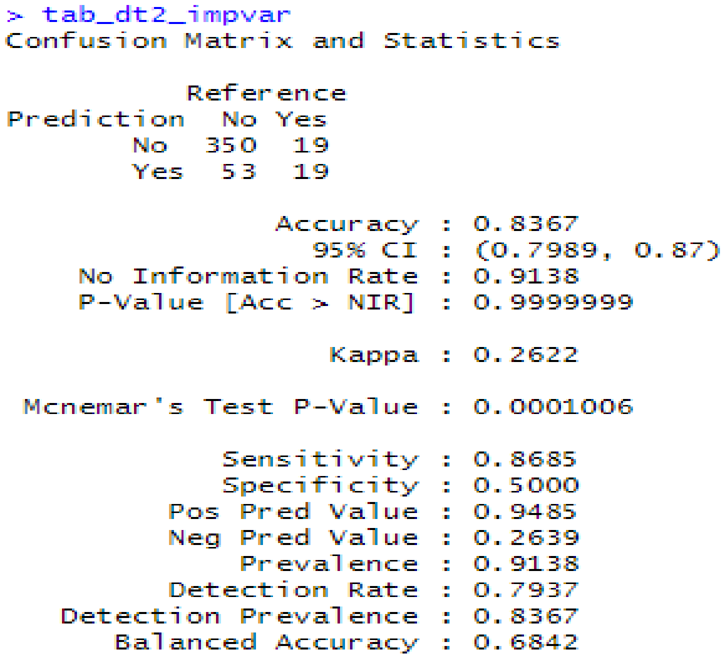
Accuracy =0.82

Sensitivity = 0.90

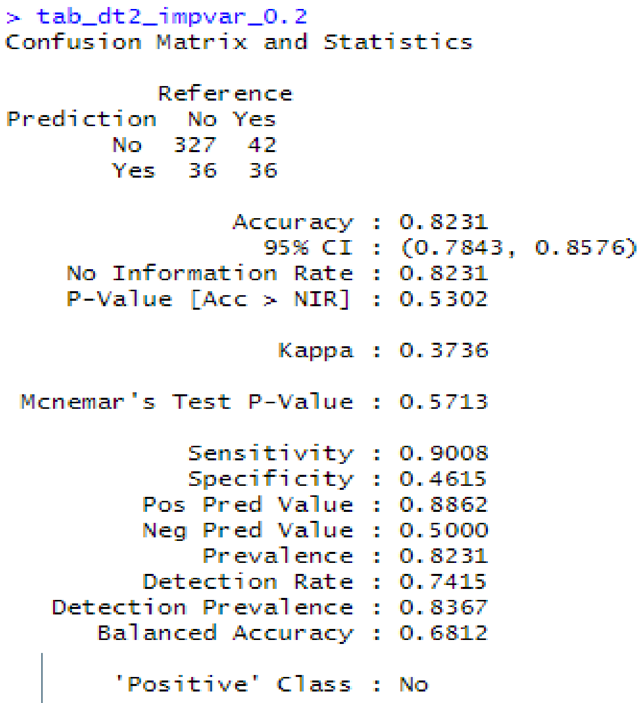
Specificity = 0.4615

*** WHEN CUTOFF TAKEN FROM ROCR

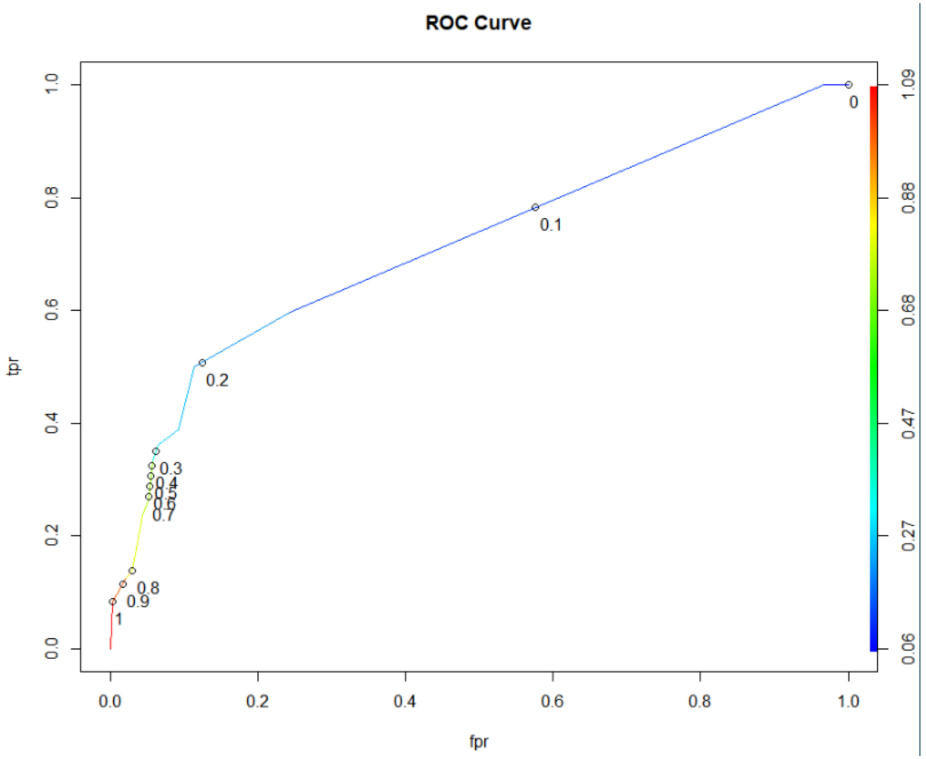
AUC = 0.7137



```
> tab_dt2_impvar
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  350  19
       Yes  53  19

               Accuracy : 0.8367
                 95% CI : (0.7989, 0.87)
    No Information Rate : 0.9138
    P-Value [Acc > NIR] : 0.9999999

                  Kappa : 0.2622

 Mcnemar's Test P-Value : 0.0001006

            Sensitivity : 0.8685
            Specificity : 0.5000
         Pos Pred Value : 0.9485
         Neg Pred Value : 0.2639
             Prevalence : 0.9138
         Detection Rate : 0.7937
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.6842
```



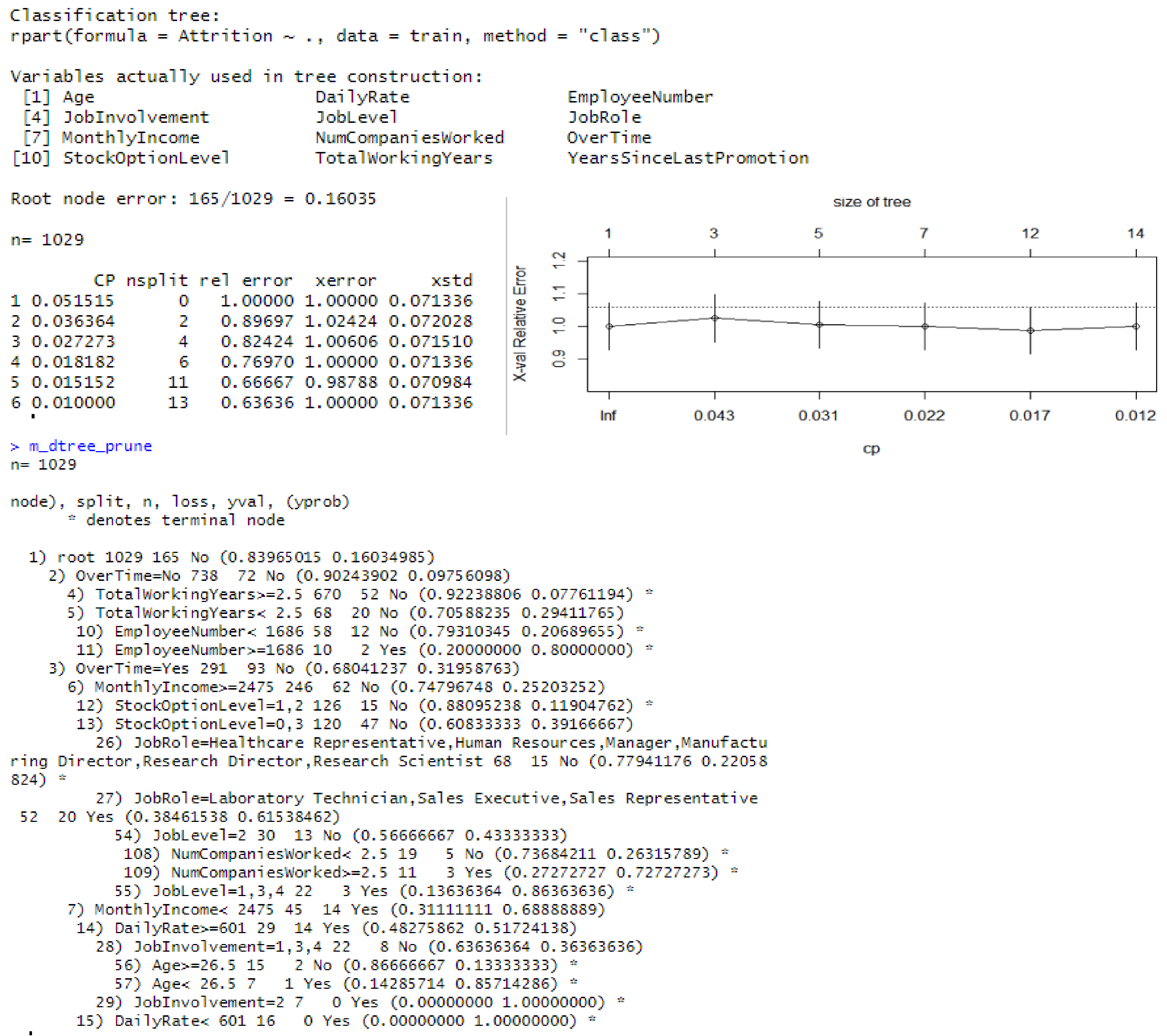
```
> tab_dt2_impvar_0.2
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  327  42
       Yes  36  36

               Accuracy : 0.8231
                 95% CI : (0.7843, 0.8576)
    No Information Rate : 0.8231
    P-Value [Acc > NIR] : 0.5302

                  Kappa : 0.3736

 Mcnemar's Test P-Value : 0.5713

            Sensitivity : 0.9008
            Specificity : 0.4615
         Pos Pred Value : 0.8862
         Neg Pred Value : 0.5000
             Prevalence : 0.8231
         Detection Rate : 0.7415
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.6812

       'Positive' Class : No
```



MODEL 5
Decision Tree

(Pruned decision tree with all variable)

Min Cp. = 0.01515



```
Classification tree:
rpart(formula = Attrition ~ ., data = train, method = "class")

Variables actually used in tree construction:
 [1] Age                   DailyRate              EmployeeNumber
 [4] JobInvolvement        JobLevel               JobRole
 [7] MonthlyIncome         NumCompaniesWorked     OverTime
[10] StockOptionLevel      TotalWorkingYears      YearsSinceLastPromotion

Root node error: 165/1029 = 0.16035

n= 1029

        CP nsplit rel error  xerror     xstd
1 0.051515      0   1.00000 1.00000 0.071336
2 0.036364      2   0.89697 1.02424 0.072028
3 0.027273      4   0.82424 1.00606 0.071510
4 0.018182      6   0.76970 1.00000 0.071336
5 0.015152     11   0.66667 0.98788 0.070984
6 0.010000     13   0.63636 1.00000 0.071336
```

```
> m_dtree_prune
n= 1029

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 1029 165 No (0.83965015 0.16034985)
    2) OverTime=No 738  72 No (0.90243902 0.09756098)
      4) TotalWorkingYears>=2.5 670  52 No (0.92238806 0.07761194) *
      5) TotalWorkingYears< 2.5 68  20 No (0.70588235 0.29411765)
       10) EmployeeNumber< 1686 58  12 No (0.79310345 0.20689655) *
       11) EmployeeNumber>=1686 10   2 Yes (0.20000000 0.80000000) *
    3) OverTime=Yes 291  93 No (0.68041237 0.31958763)
      6) MonthlyIncome>=2475 246  62 No (0.74796748 0.25203252)
       12) StockOptionLevel=1,2 126  15 No (0.88095238 0.11904762) *
       13) StockOptionLevel=0,3 120  47 No (0.60833333 0.39166667)
         26) JobRole=Healthcare Representative,Human Resources,Manager,Manufactu
ring Director,Research Director,Research Scientist 68  15 No (0.77941176 0.22058
824) *
         27) JobRole=Laboratory Technician,Sales Executive,Sales Representative
 52  20 Yes (0.38461538 0.61538462)
           54) JobLevel=2 30  13 No (0.56666667 0.43333333)
            108) NumCompaniesWorked< 2.5 19   5 No (0.73684211 0.26315789) *
            109) NumCompaniesWorked>=2.5 11   3 Yes (0.27272727 0.72727273) *
           55) JobLevel=1,3,4 22   3 Yes (0.13636364 0.86363636) *
      7) MonthlyIncome< 2475 45  14 Yes (0.31111111 0.68888889)
       14) DailyRate>=601 29  14 Yes (0.48275862 0.51724138)
         28) JobInvolvement=1,3,4 22   8 No (0.63636364 0.36363636)
           56) Age>=26.5 15   2 No (0.86666667 0.13333333) *
           57) Age< 26.5 7   1 Yes (0.14285714 0.85714286) *
         29) JobInvolvement=2 7   0 Yes (0.00000000 1.00000000) *
       15) DailyRate< 601 16   0 Yes (0.00000000 1.00000000) *
```

Decision Tree
(Pruned) contd…

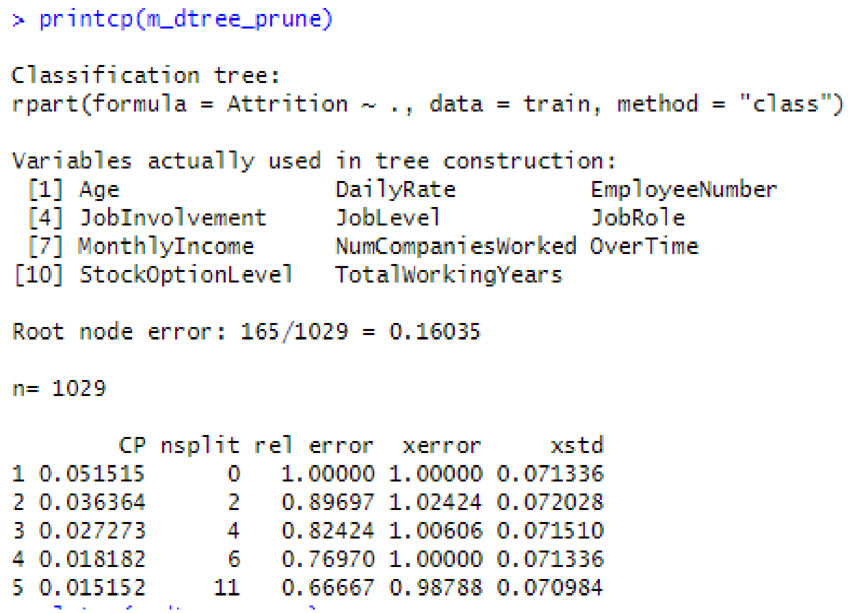
For dtree_prune

Accuracy = 0.8367

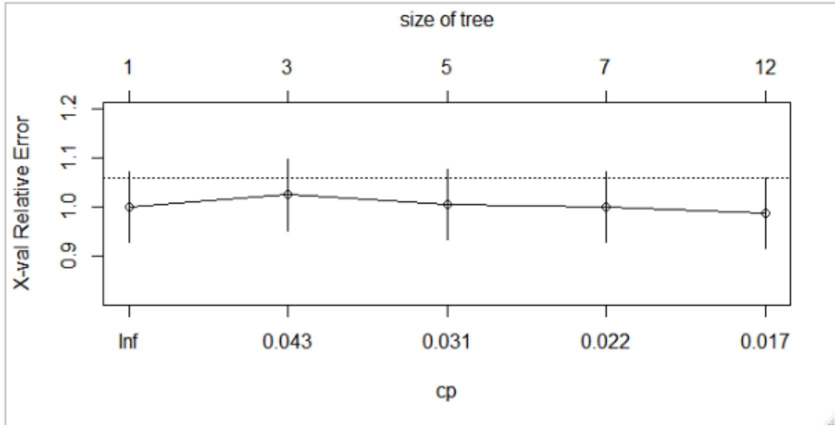
Sensitivity = 0.86

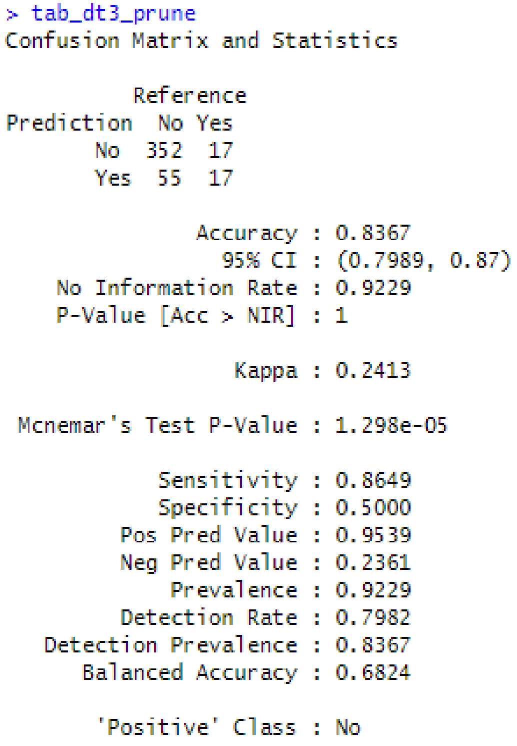
Specificity = 0.50

AUC = 0.6874

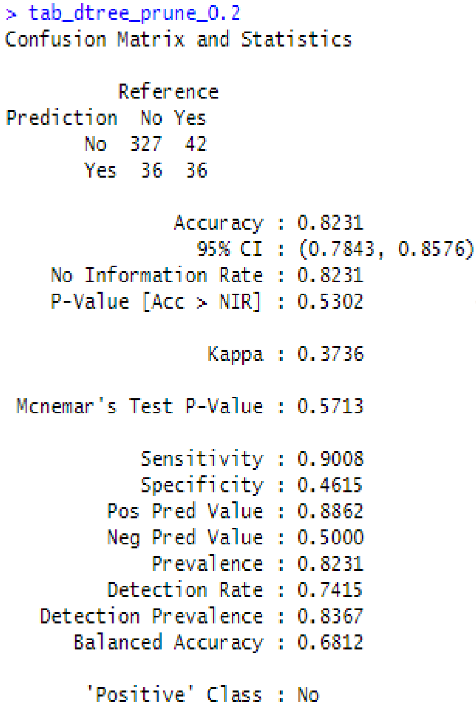


```
> printcp(m_dtree_prune)

Classification tree:
rpart(formula = Attrition ~ ., data = train, method = "class")

Variables actually used in tree construction:
 [1] Age               DailyRate         EmployeeNumber
 [4] JobInvolvement    JobLevel          JobRole
 [7] MonthlyIncome     NumCompaniesWorked OverTime
[10] StockOptionLevel  TotalWorkingYears

Root node error: 165/1029 = 0.16035

n= 1029

        CP nsplit rel error  xerror     xstd
1 0.051515      0   1.00000 1.00000 0.071336
2 0.036364      2   0.89697 1.02424 0.072028
3 0.027273      4   0.82424 1.00606 0.071510
4 0.018182      6   0.76970 1.00000 0.071336
5 0.015152     11   0.66667 0.98788 0.070984
```

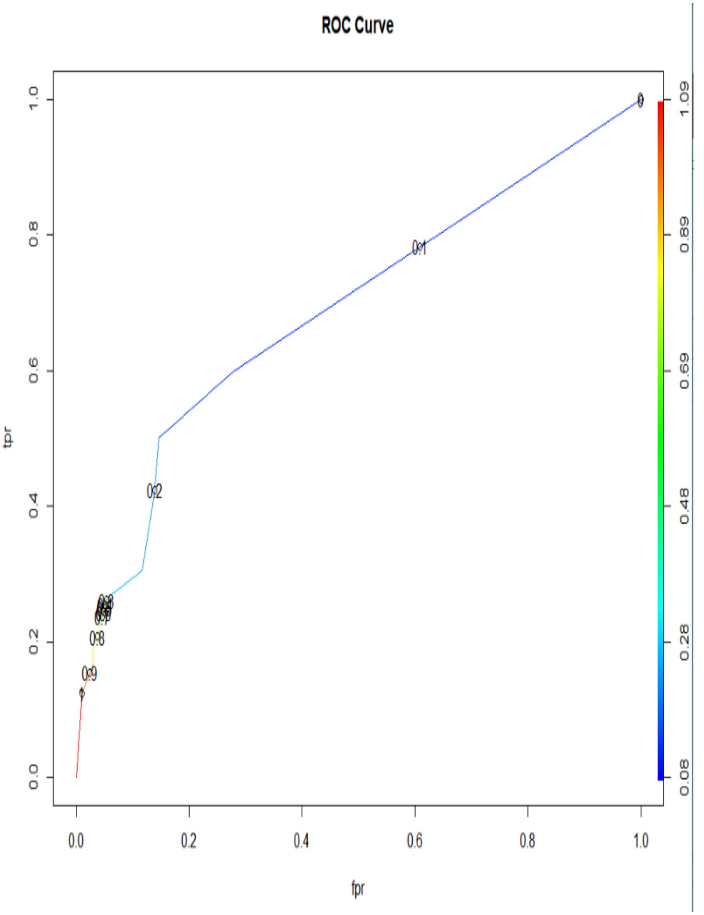




```
> tab_dt3_prune
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  352  17
       Yes  55  17

               Accuracy : 0.8367
                 95% CI : (0.7989, 0.87)
    No Information Rate : 0.9229
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2413

 Mcnemar's Test P-Value : 1.298e-05

            Sensitivity : 0.8649
            Specificity : 0.5000
         Pos Pred Value : 0.9539
         Neg Pred Value : 0.2361
             Prevalence : 0.9229
         Detection Rate : 0.7982
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.6824

       'Positive' Class : No
```

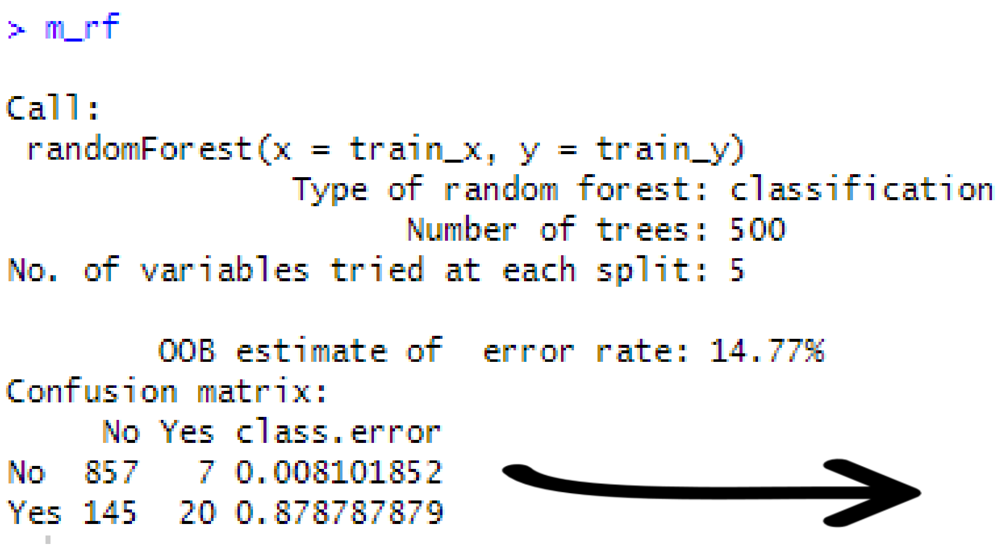


```
> tab_dtree_prune_0.2
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  327  42
       Yes  36  36

               Accuracy : 0.8231
                 95% CI : (0.7843, 0.8576)
    No Information Rate : 0.8231
    P-Value [Acc > NIR] : 0.5302

                  Kappa : 0.3736

 Mcnemar's Test P-Value : 0.5713

            Sensitivity : 0.9008
            Specificity : 0.4615
         Pos Pred Value : 0.8862
         Neg Pred Value : 0.5000
             Prevalence : 0.8231
         Detection Rate : 0.7415
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.6812

       'Positive' Class : No
```

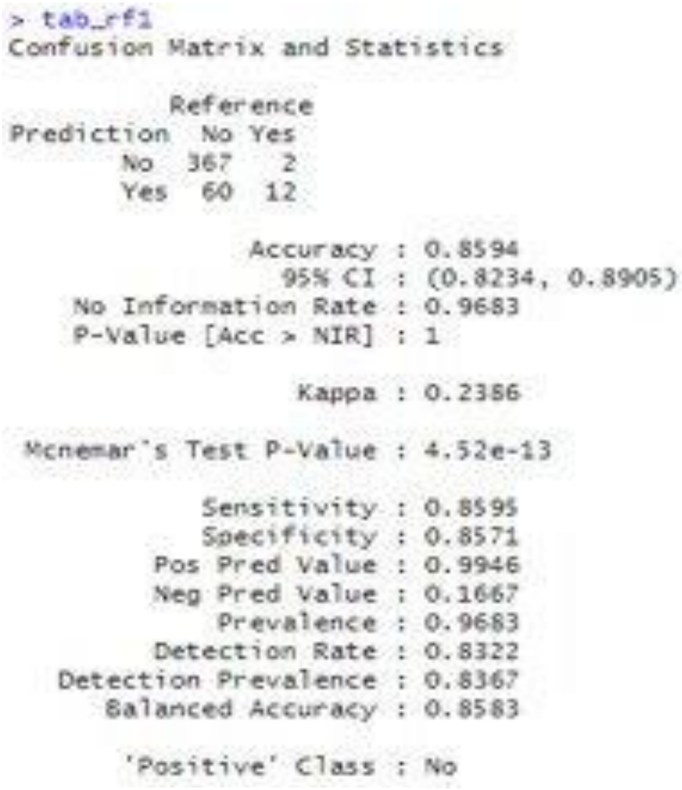


MODEL 6 RandomForest

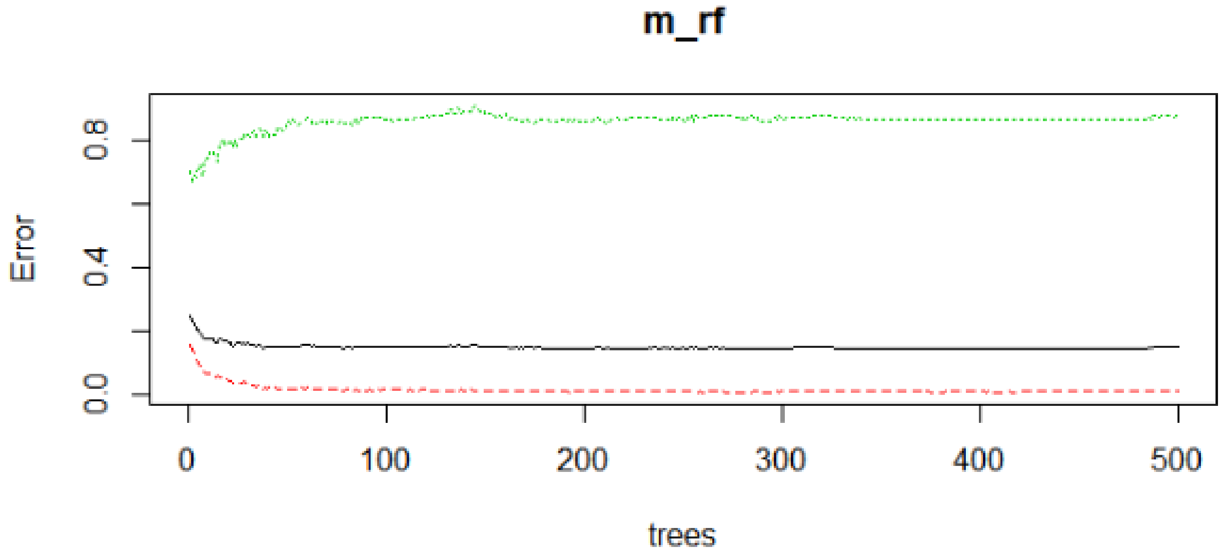
- OOB estimate of error rate: 14.77%
- No. of variables tried at each split: 5
- Number of trees: 500



```
> m_rf

Call:
 randomForest(x = train_x, y = train_y) 
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 14.77%
Confusion matrix:
     No Yes class.error
No  857   7 0.008101852
Yes 145  20 0.878787879
```



```
> tab_rf1
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  367   2
       Yes  60  12

               Accuracy : 0.8594
                 95% CI : (0.8234, 0.8905)
    No Information Rate : 0.9683
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2386

 Mcnemar's Test P-Value : 4.52e-13

            Sensitivity : 0.8595
            Specificity : 0.8571
         Pos Pred Value : 0.9946
         Neg Pred Value : 0.1667
             Prevalence : 0.9683
         Detection Rate : 0.8322
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.8583

       'Positive' Class : No
```



RandomForest

1. Finding best value of Mtry = 7

NTreeTry = 600

StepFactor = 1.2

Improve = 0.01

OOB error = 14.19%

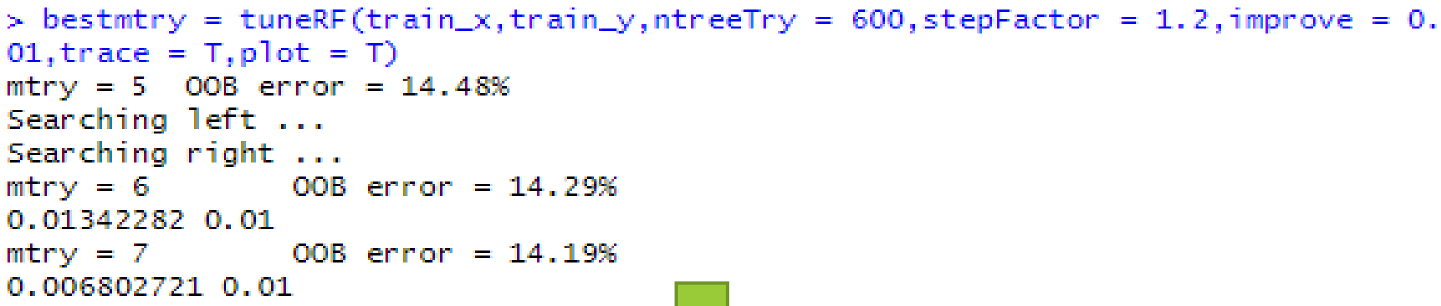
2. further finding best value of mtry =5

NtreeTry = 600

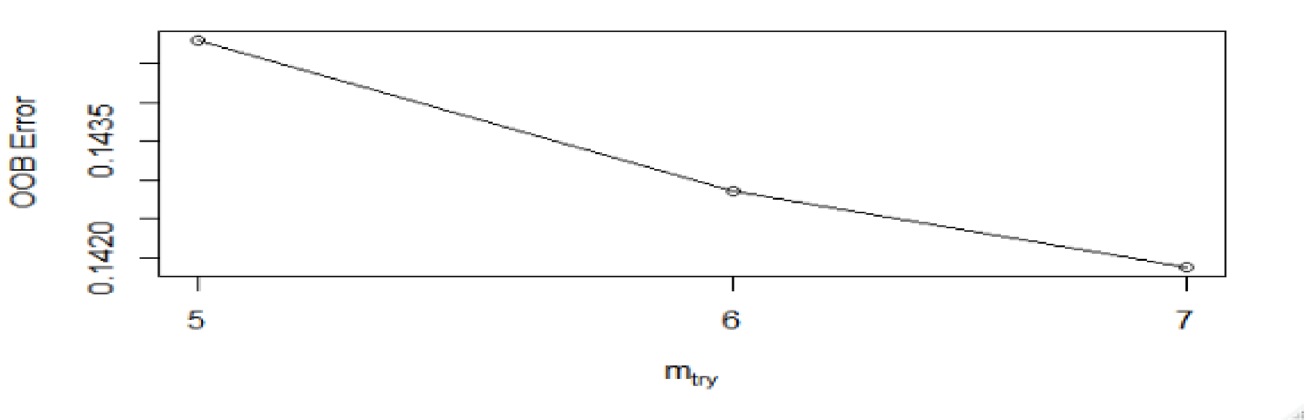
StepFactor = 1

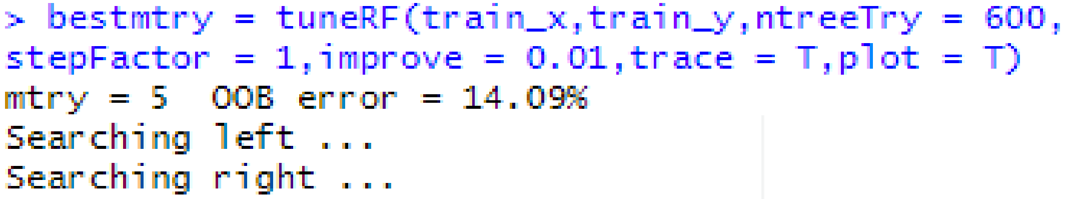
Improve = 0.01

OOB error = 14.09%

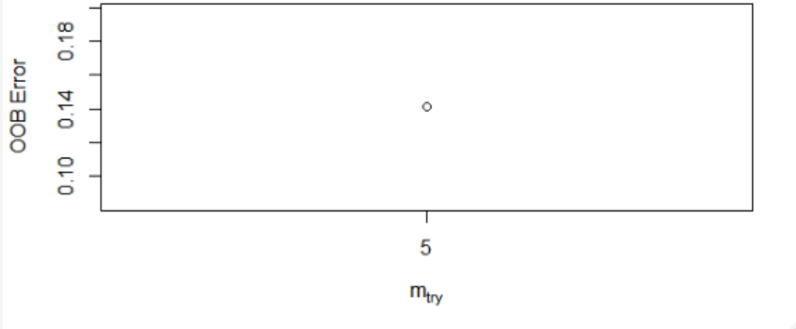


```
> bestmtry = tuneRF(train_x,train_y,ntreeTry = 600,stepFactor = 1.2,improve = 0.
01,trace = T,plot = T)
mtry = 5  OOB error = 14.48%
Searching left ...
Searching right ...
mtry = 6        OOB error = 14.29%
0.01342282 0.01
mtry = 7        OOB error = 14.19%
0.006802721 0.01
```





```
> bestmtry = tuneRF(train_x,train_y,ntreeTry = 600,
stepFactor = 1,improve = 0.01,trace = T,plot = T)
mtry = 5  OOB error = 14.09%
Searching left ...
Searching right ...
```



RF model with Best Mtry=5

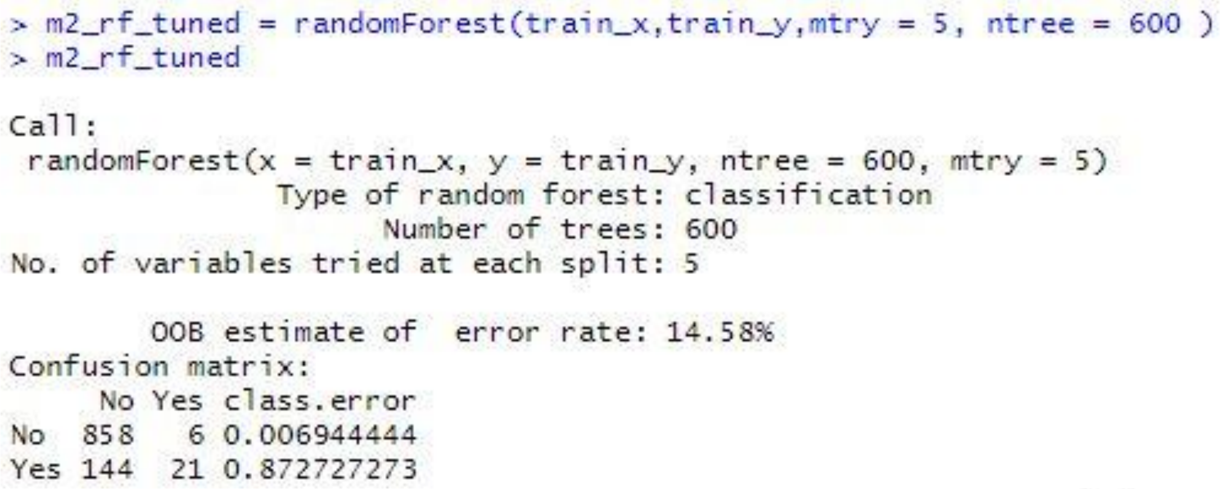
OOB error = 14.09

Accuracy = 0.8639

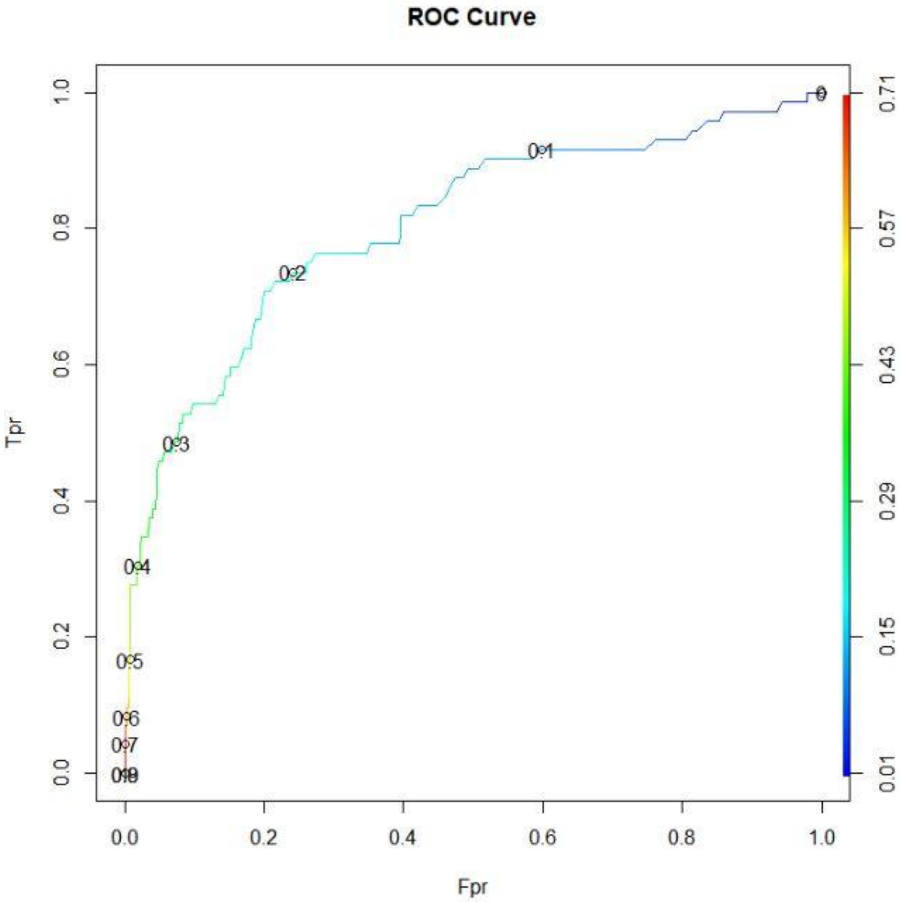
Sensitivity =0.8652

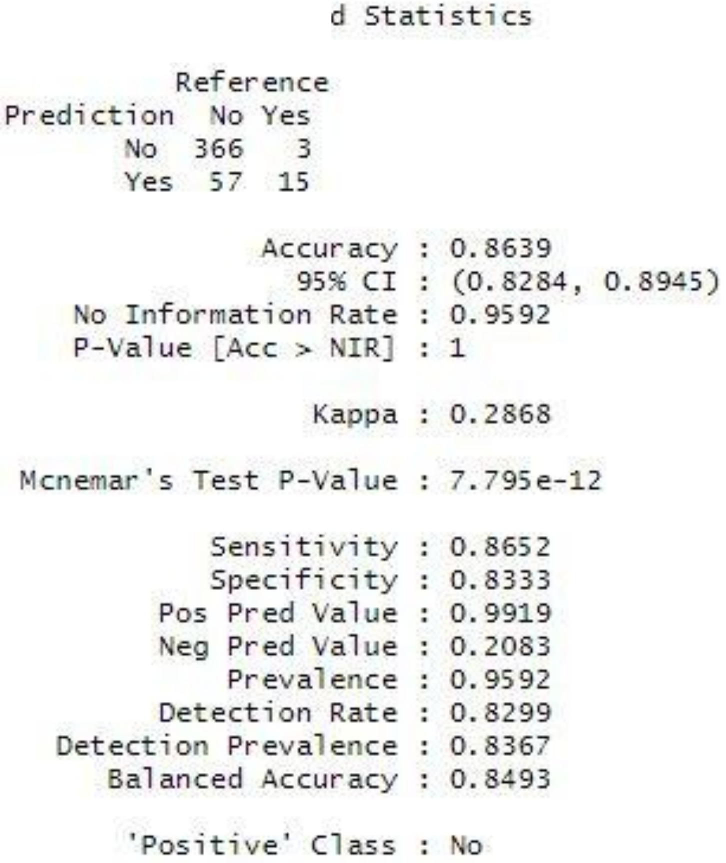
Specificity = 0.833

AUC = 0.8512



```
> m2_rf_tuned = randomForest(train_x,train_y,mtry = 5, ntree = 600 )
> m2_rf_tuned

Call:
 randomForest(x = train_x, y = train_y, ntree = 600, mtry = 5)
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 5

        OOB estimate of  error rate: 14.58%
Confusion matrix:
     No Yes class.error
No  858   6 0.006944444
Yes 144  21 0.872727273
```





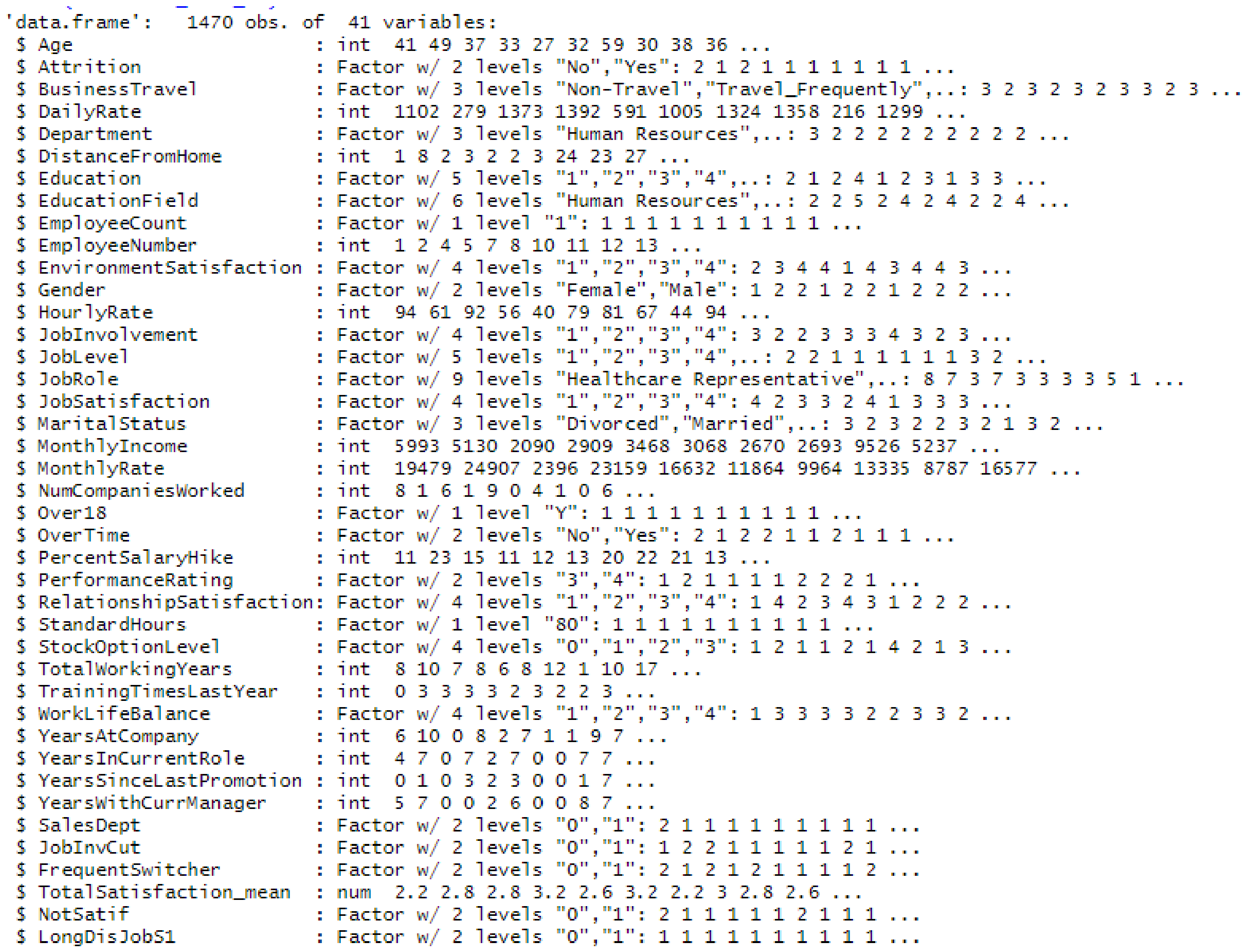
```
                              d Statistics

          Reference
Prediction  No Yes
       No  366   3
       Yes  57  15

               Accuracy : 0.8639
                 95% CI : (0.8284, 0.8945)
    No Information Rate : 0.9592
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2868

 Mcnemar's Test P-Value : 7.795e-12

            Sensitivity : 0.8652
            Specificity : 0.8333
         Pos Pred Value : 0.9919
         Neg Pred Value : 0.2083
             Prevalence : 0.9592
         Detection Rate : 0.8299
   Detection Prevalence : 0.8367
      Balanced Accuracy : 0.8493

       'Positive' Class : No
```

Feature Engineering

In Feature Engineering NEW VERIABLE based on the main data are introduced

NEW VERIABLES ->

- SalesDept
- JobInvCut
- FrequentSwitcher
- TotalSatisfaction_mean
- NotSatif
- LongDisJobS1



```
'data.frame':   1470 obs. of  41 variables:
 $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
 $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
 $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 2 3 2 3 2 3 3 2 3 ...
 $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
 $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...
 $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
 $ Education               : Factor w/ 5 levels "1","2","3","4",..: 2 1 2 4 1 2 3 1 3 3 ...
 $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...
 $ EmployeeCount           : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...
 $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
 $ EnvironmentSatisfaction : Factor w/ 4 levels "1","2","3","4": 2 3 4 4 1 4 3 4 4 3 ...
 $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
 $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
 $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 3 2 2 3 3 3 4 3 2 3 ...
 $ JobLevel                : Factor w/ 5 levels "1","2","3","4",..: 2 2 1 1 1 1 1 1 3 2 ...
 $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1 ...
 $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4 1 3 3 3 ...
 $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...
 $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
 $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
 $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
 $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
 $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...
 $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
 $ PerformanceRating       : Factor w/ 2 levels "3","4": 1 2 1 1 1 1 2 2 2 1 ...
 $ RelationshipSatisfaction: Factor w/ 4 levels "1","2","3","4": 1 4 2 3 4 3 1 2 2 2 ...
 $ StandardHours           : Factor w/ 1 level "80": 1 1 1 1 1 1 1 1 1 1 ...
 $ StockOptionLevel        : Factor w/ 4 levels "0","1","2","3": 1 2 1 1 2 1 4 2 1 3 ...
 $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
 $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
 $ WorkLifeBalance         : Factor w/ 4 levels "1","2","3","4": 1 3 3 3 3 2 2 3 3 2 ...
 $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
 $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
 $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
 $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
 $ SalesDept               : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
 $ JobInvCut               : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 1 2 1 ...
 $ FrequentSwitcher        : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 1 1 1 2 ...
 $ TotalSatisfaction_mean  : num  2.2 2.8 2.8 3.2 2.6 3.2 2.2 3 2.8 2.6 ...
 $ NotSatif                : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 2 1 1 1 ...
 $ LongDisJobS1            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Model 7
Random Forest

(With New Variables)

NO. OF TREES = 500

Ntry = 6

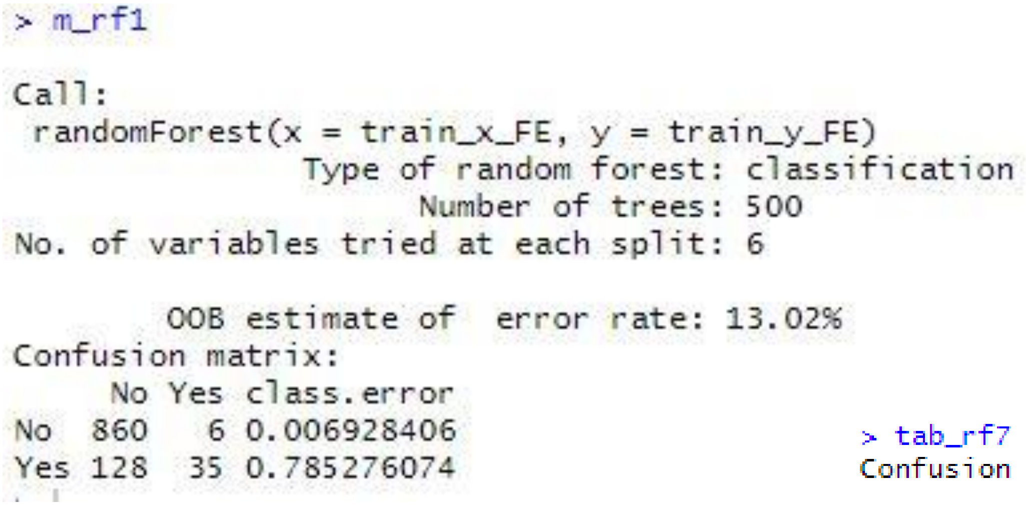
OOB error estimate = 13.02%

ACCURACY = 0.8762

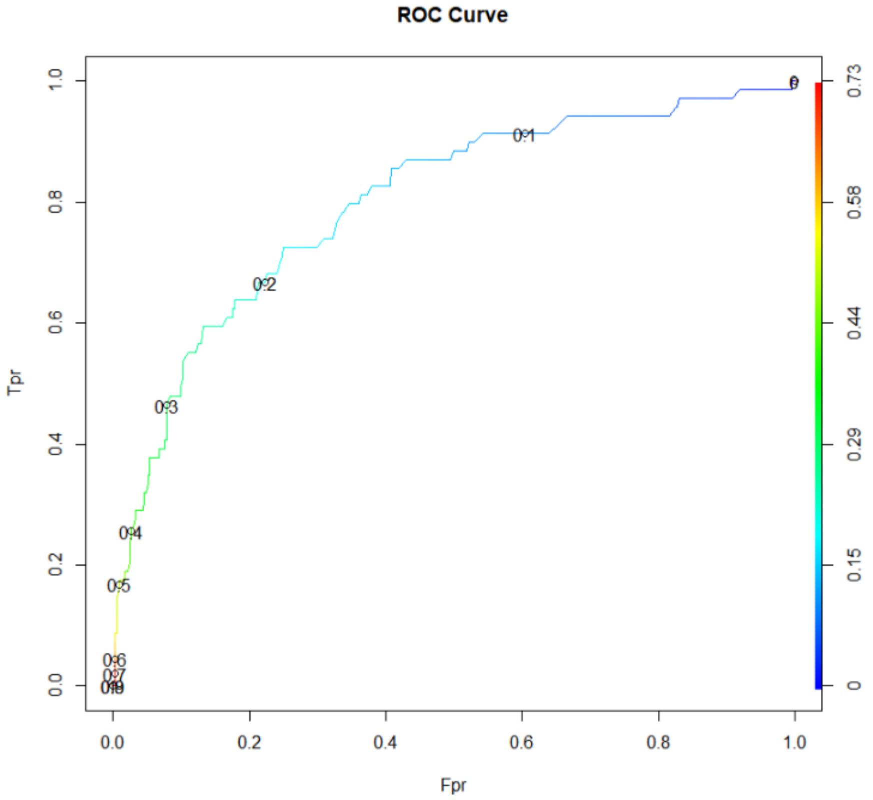
SENSITIVITY = 0.8762

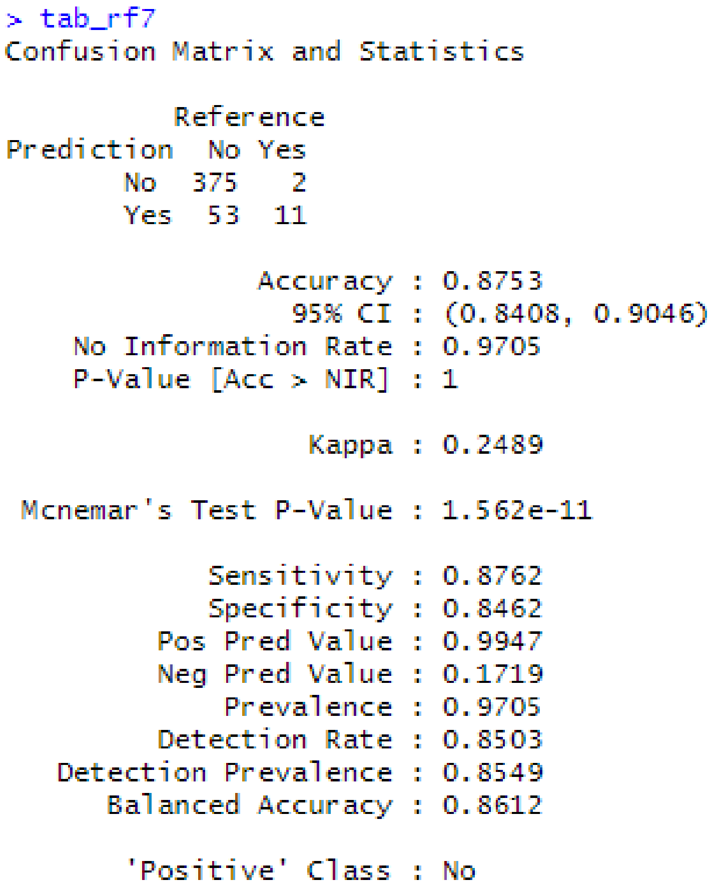
SPECIFICITY = 0.8462

AUC = 0.8291



```
> m_rf1

Call:
 randomForest(x = train_x_FE, y = train_y_FE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 13.02%
Confusion matrix:
     No Yes class.error
No  860   6 0.006928406
Yes 128  35 0.785276074
```

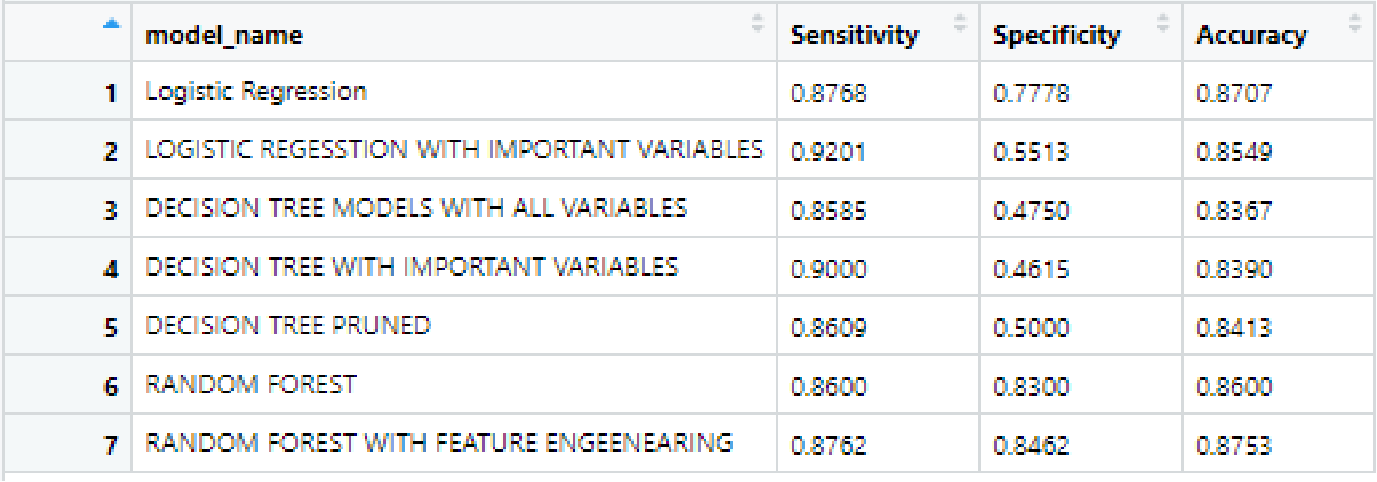




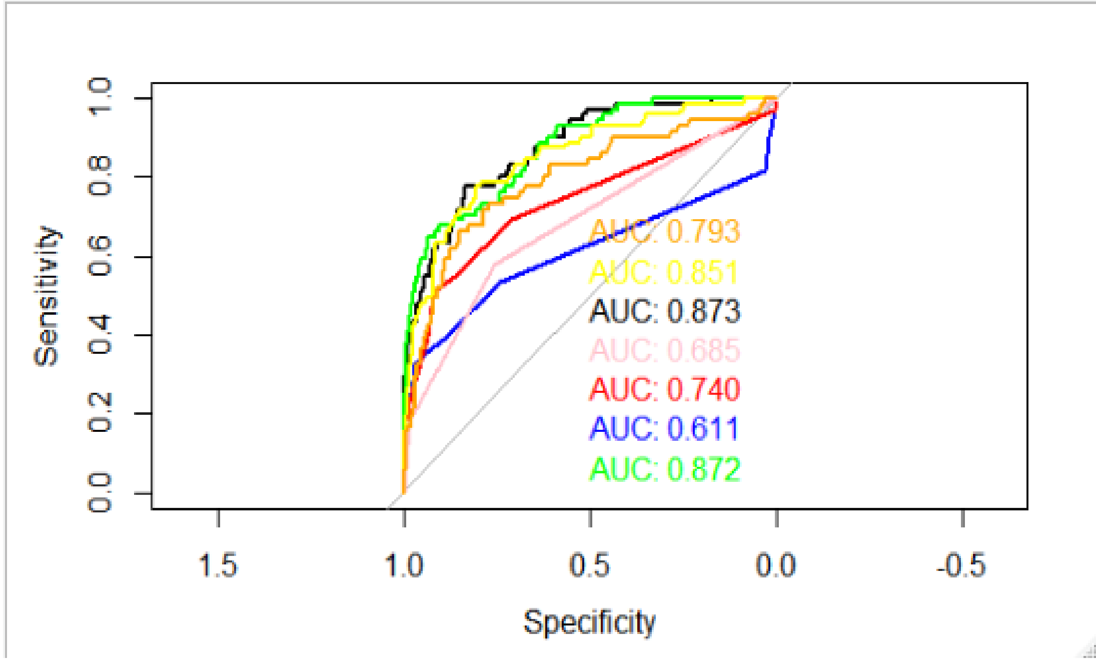
```
> tab_rf7
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  375   2
       Yes  53  11

               Accuracy : 0.8753
                 95% CI : (0.8408, 0.9046)
    No Information Rate : 0.9705
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2489

 Mcnemar's Test P-Value : 1.562e-11

            Sensitivity : 0.8762
            Specificity : 0.8462
         Pos Pred Value : 0.9947
         Neg Pred Value : 0.1719
             Prevalence : 0.9705
         Detection Rate : 0.8503
   Detection Prevalence : 0.8549
      Balanced Accuracy : 0.8612

       'Positive' Class : No
```

Model Selection

- ü SELECTION OF MODEL IS DEPENDENT PROCESS ON THE BUSINESS CONTEXT
- ü SIMPLEST MODEL MUST BE CHOSEN
- ü WHILE SELECTING MODEL STABILITY IN OUTPUT MUST CONSIDERED
- ü AVOID UNDERFITTING, OVERFITTING
- ü MAKE TRAINING ERROR SMALL
- ü MAKE GAP BETWEEN TRAINING ERROR AND TEST ERROR SMALL



	model_name	Sensitivity	Specificity	Accuracy
1	Logistic Regression	0.8768	0.7778	0.8707
2	LOGISTIC REGESSTION WITH IMPORTANT VARIABLES	0.9201	0.5513	0.8549
3	DECISION TREE MODELS WITH ALL VARIABLES	0.8585	0.4750	0.8367
4	DECISION TREE WITH IMPORTANT VARIABLES	0.9000	0.4615	0.8390
5	DECISION TREE PRUNED	0.8609	0.5000	0.8413
6	RANDOM FOREST	0.8600	0.8300	0.8600
7	RANDOM FOREST WITH FEATURE ENGEENEARING	0.8762	0.8462	0.8753



- Selecting Best model is nothing but selecting with high Prediction, Accuracy , specificity, Sensitivity but also consideration of model simplicity and stability
- **Random Forest with New Variables** out performed in terms of **Accuracy**
- From above cases it can be seen that **<u>Logistic Regression</u> or <u>Random Forest</u>** can be implemented as it is **easier to implement and Performing good** at **Prediction**

Thank You!!!