Assignment - 1 (Exploratory Data Analytics and model building)

Problem Statement

A dataset has been given that contains the records of employees who have left and remained in a company along with various other metrics like job satisfaction, salary etc., We have to build a model that predicts whether a given employee leaves or not.

Understanding the Data Through EDA

The given dataset captures the following variables for each employee -

- Satisfaction Level
- Last evaluation
- Number of projects
- Average monthly hours
- Time spent at the company
- Whether they have had a work accident
- Whether they have had a promotion in the last 5 years
- Departments (column sales)
- Salary
- Whether the employee has left

Case Study on HR data:

The data set is given to predict if the employee leaves or stays back in the company based on information like salary, department, working hours, time spent, promotions etc.

Basic Information:

The data set has the following features:

The data set is read with the name of the data frame “hrdata”

The structure of the data frame is given using str(hrdata)

Total number of observations(rows): 14999

Total number of Variables (columns): 10

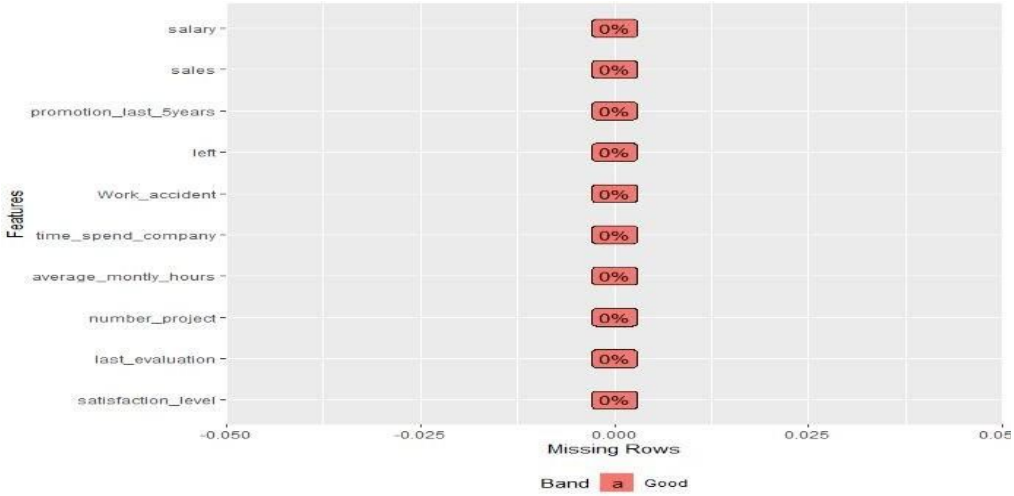
The data set looks like this

satisfaction_level	last_evaluation	number_project	average_montly	time_spend_	Work_accident	left	promotion_	sales	salary
0.38	0.53	2	157	3	0	1	0	sales	low
0.8	0.86	5	262	6	0	1	0	sales	medium
0.11	0.88	7	272	4	0	1	0	sales	medium
0.72	0.87	5	223	5	0	1	0	sales	low
0.37	0.52	2	159	3	0	1	0	sales	low
0.41	0.5	2	153	3	0	1	0	sales	low
0.1	0.77	6	247	4	0	1	0	sales	low

There are 2 categorical variables and 8 numeric variables.

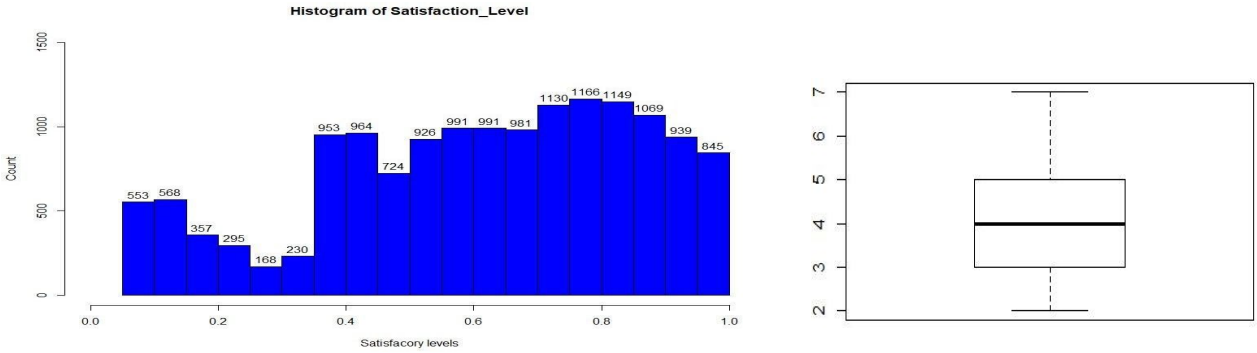
```
data.frame':   14999 obs. of  10 variables:
$ satisfaction_level   : num  0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...
```

The data set is free of null values. So the data is clean



Variable Descriptions:

1. Satisfactory level:
    - This is a numeric variable
    - The satisfactory level is between the values 0.09 to 1. Where 0.09 is the least satisfactory level and 1 is the highest satisfactory level.
    - Around 195 members' satisfaction level was 0.09 and 111 members' satisfactory level was 1.
    - The mean and median are almost the same and we can say that this follows a normal distribution.
    - The highest count of the satisfactory levels are between 0.7 to 0.8
    - The lowest count of the satisfactory levels is between 0.25 to 0.3
    - There are no outliers for this feature. The median is nearer to the second quartile. So the data is left-skewed.



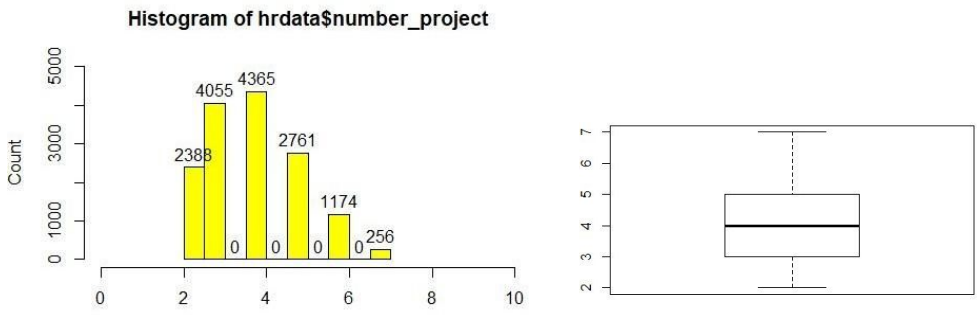
2. Number of projects:

It is a numeric variable with the values ranging from 2 to 7.

Nearly 4365 employees did 4 projects which have the highest employee count.

The box plot shows that there are no outliers. The least number of employees did 7 projects. The graph follows a normal distribution with right-skewed.

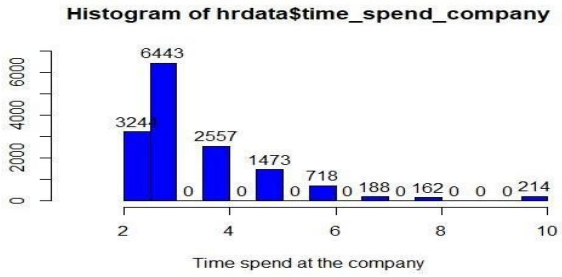
The median lies almost in the middle to the first and second quartile

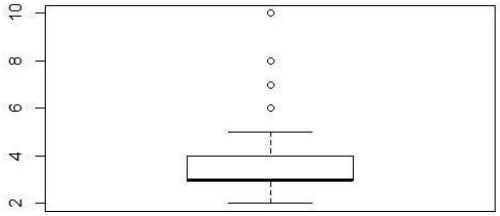


3. Time spent at the company.

This is a numeric variable with the values ranging from 2 to 10. The maximum number(6443) of employees spent 3 hours.

There are outliers present in this feature. The outliers are present to the top of the box and the longer part of the box is to the top of the median, it is slightly right-skewed.



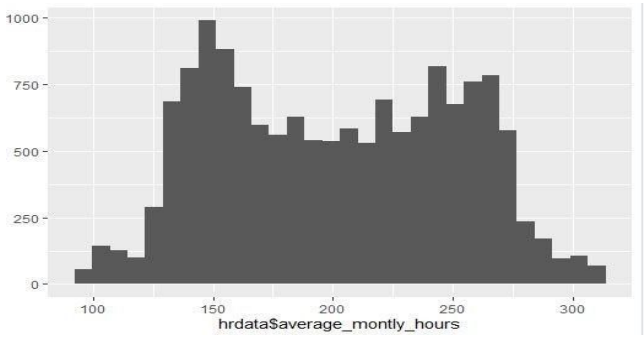


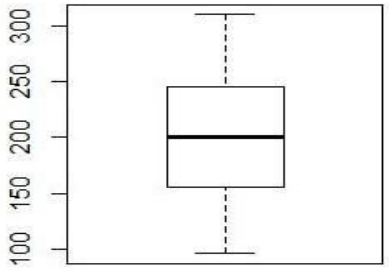
4. Average Monthly hours.

It is a numeric variable with values ranging from 96hours per month to 310 hours per month. Most numbers of employees are working for 150 to 250 hours per month on average.

The count of employees who are either working 100 hours per month or 300 hours per month is very low. There are no outliers.

The mean and median are same and the plot follows the normal distribution

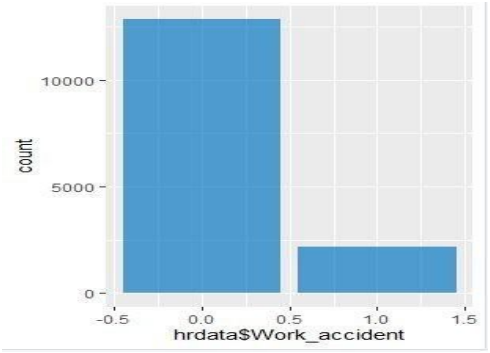




5. Work Accident:

The work accident is a numeric variable with only two values. 1 and 0. This variable can be changed to factor variable as there are two values

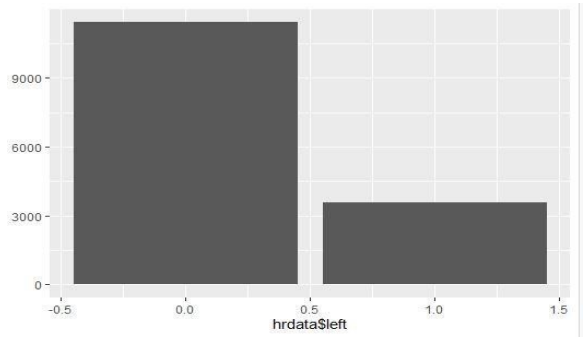
Out of 14999, only 14 percent of the people met with the accident.



6. Left:

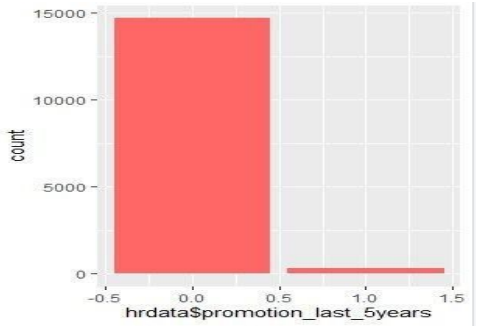
This is the numeric variable which tells that either the employee left or not. This is the dependent variable

Out of 14999, 23 percent of the employees left the company.



7. Promotion.

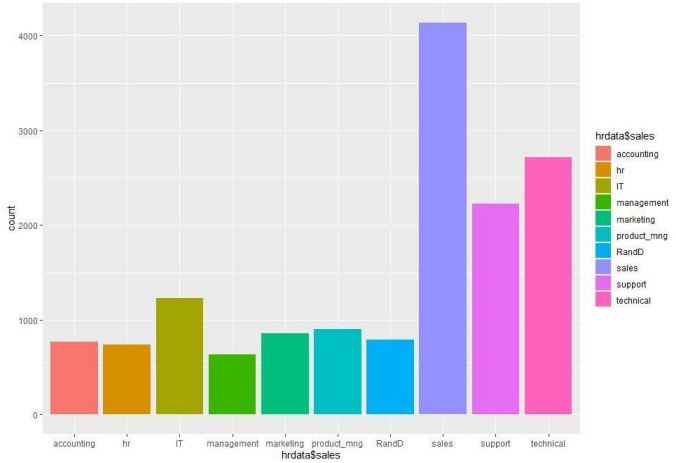
This is a numeric variable that tells either the employee got promoted or not. Only 2 percent of the employees got promoted



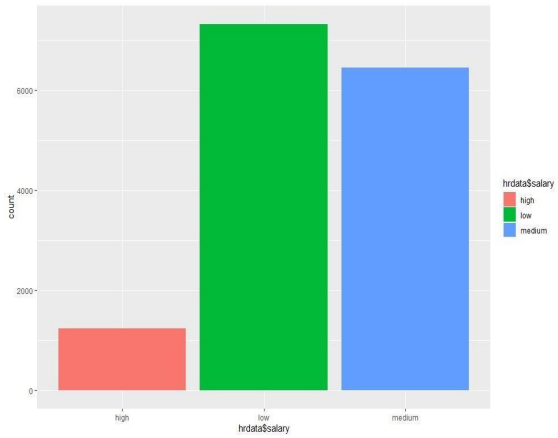
8. Sales and salary

Both are categorical variables.

Sales: The more number of employees are present in the sales department.The least number of employees are from management.

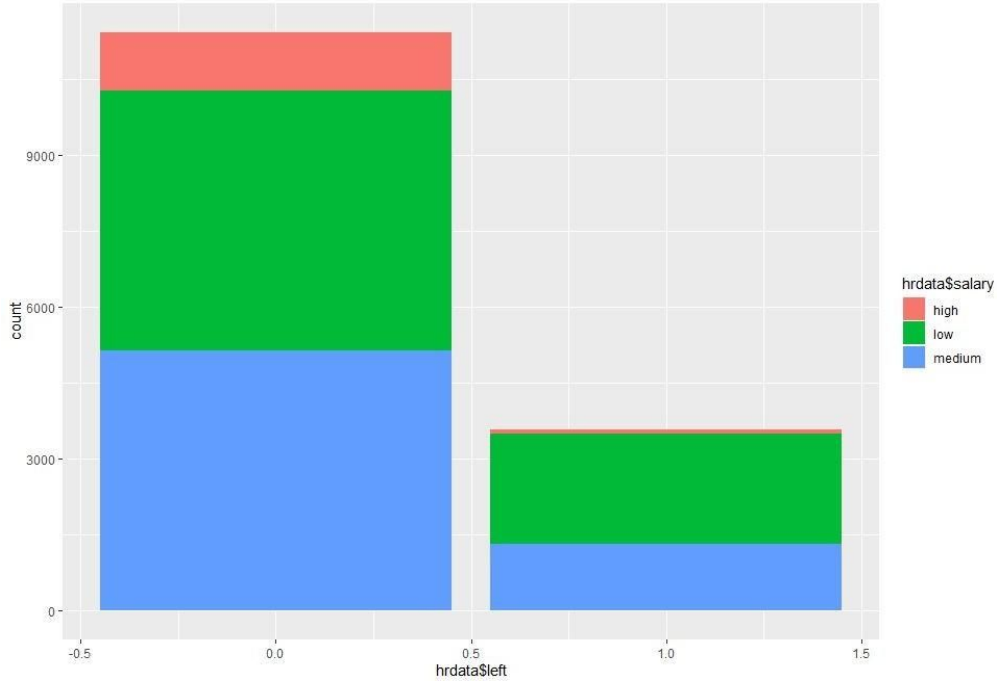


Salary: Most employees get a low salary



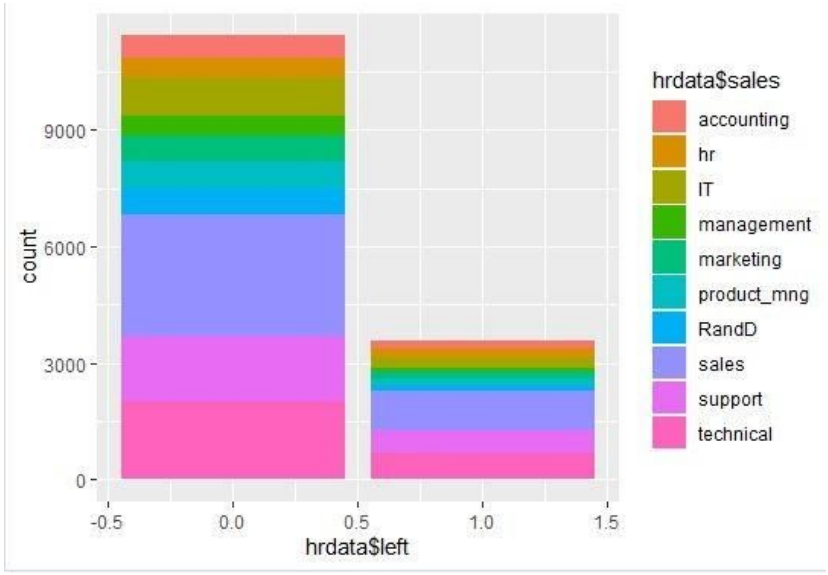
Insights:

1. The relation between salary and employees left: The employees whose salary is low, had left the company. The employees whose salary is high have a less chance of leaving the

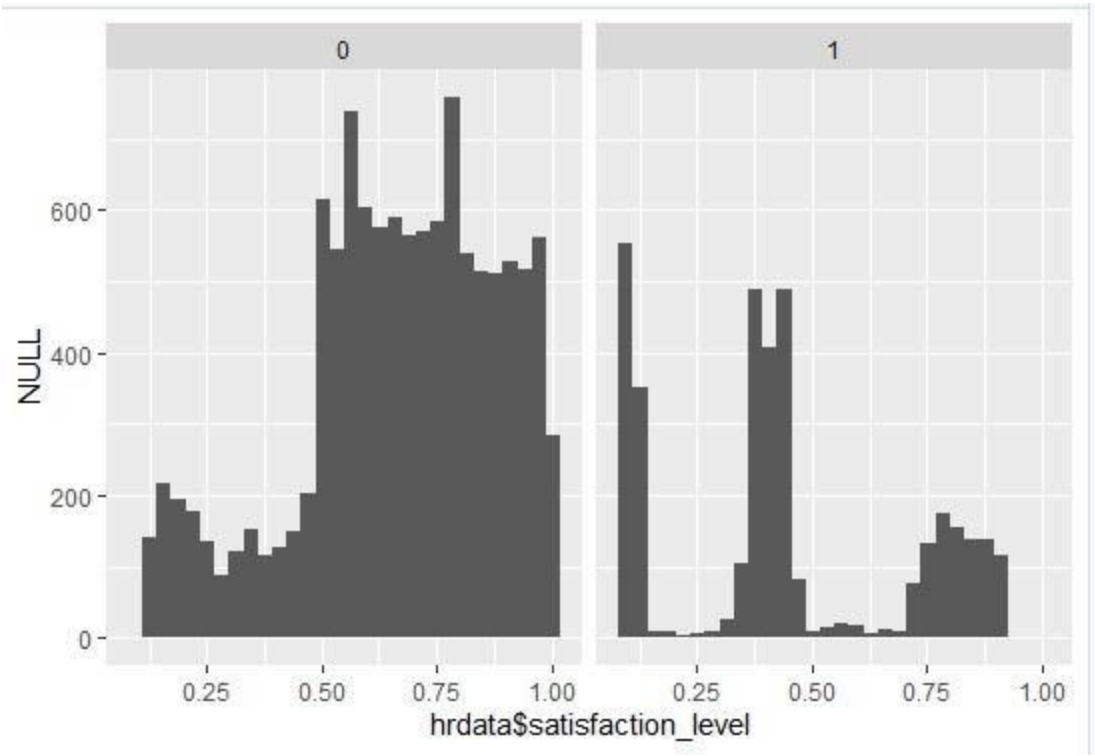


company

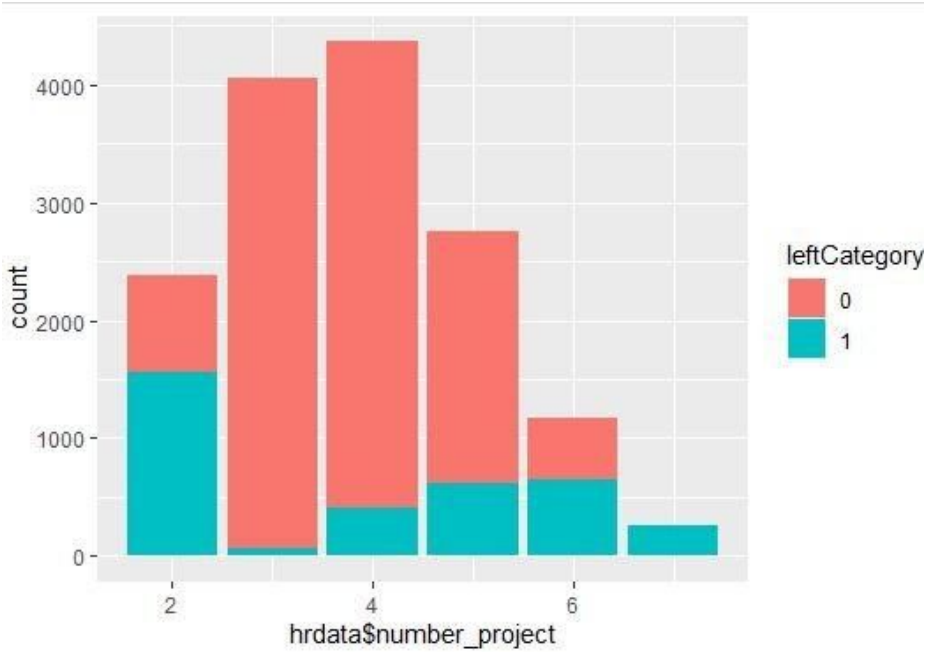
2. The relation between sales and employee left: The employee from the sales department has more chances of leaving. The employee from the accounting department has fewer chances of leaving



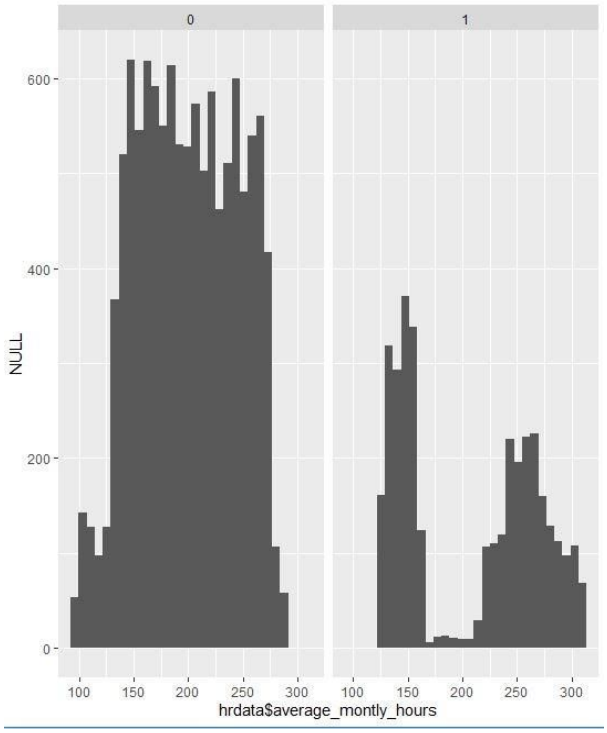
3. The relation between satisfactory levels and employees left: The employees who are less satisfied has left the company (less than 0.5)



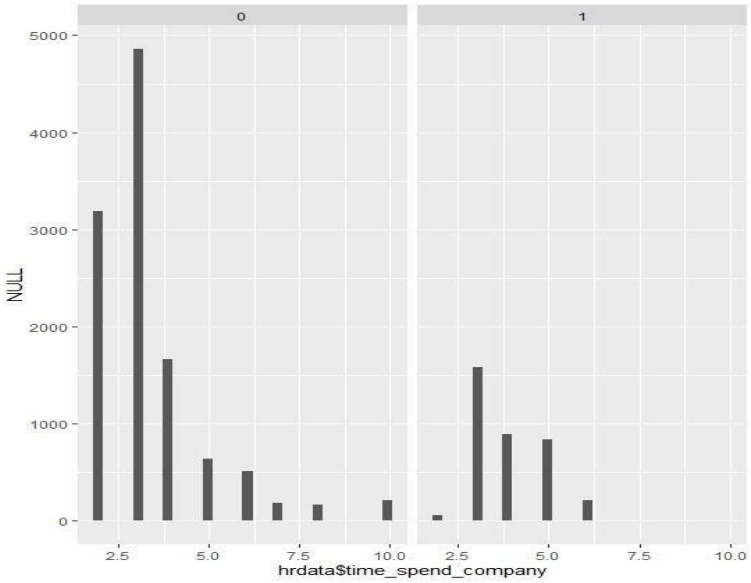
4. The relation between number of projects done and employees left: The employees who did less number of projects left.



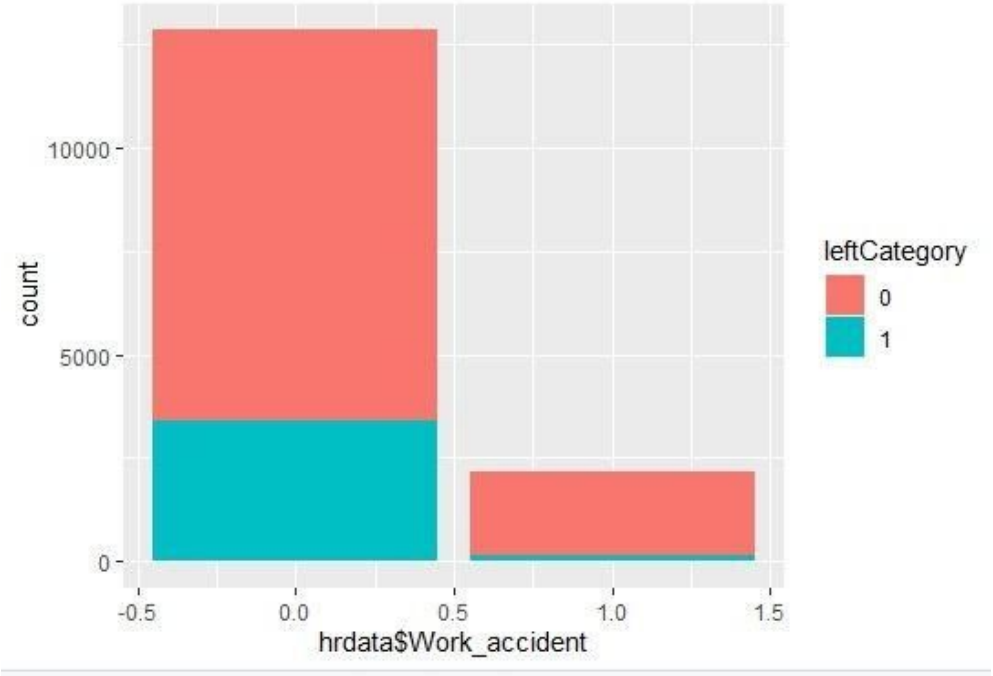
5. The relation between average monthly hours and employees left: The employees who worked either for fewer hours or more hours have high chance of leaving the company.



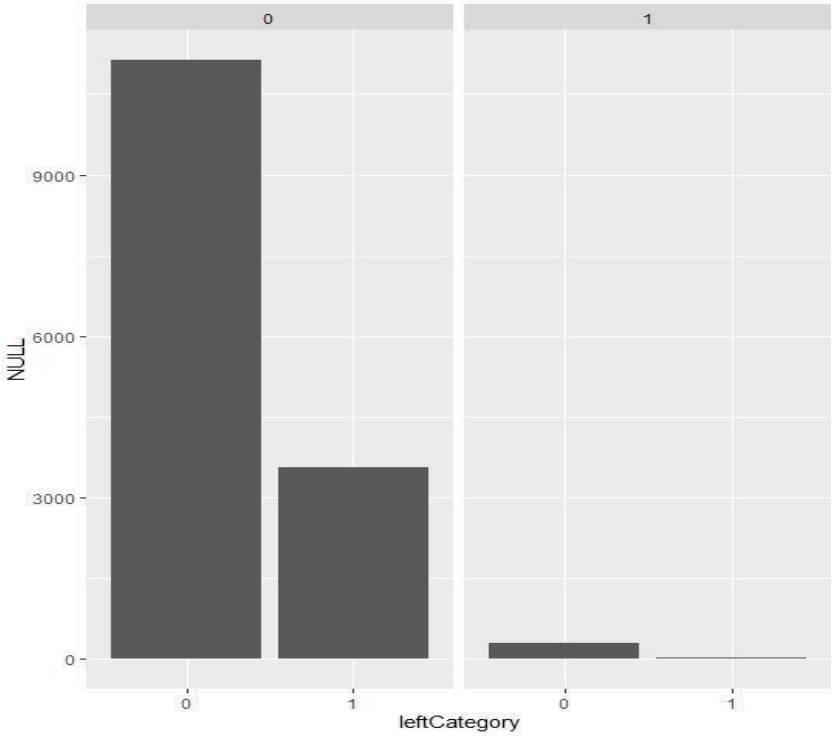
6. The relation between time spent in the company in the company and employee left: The employee who worked for fewer years, left the company. The employee who was working for more than 7 years did not leave the company.

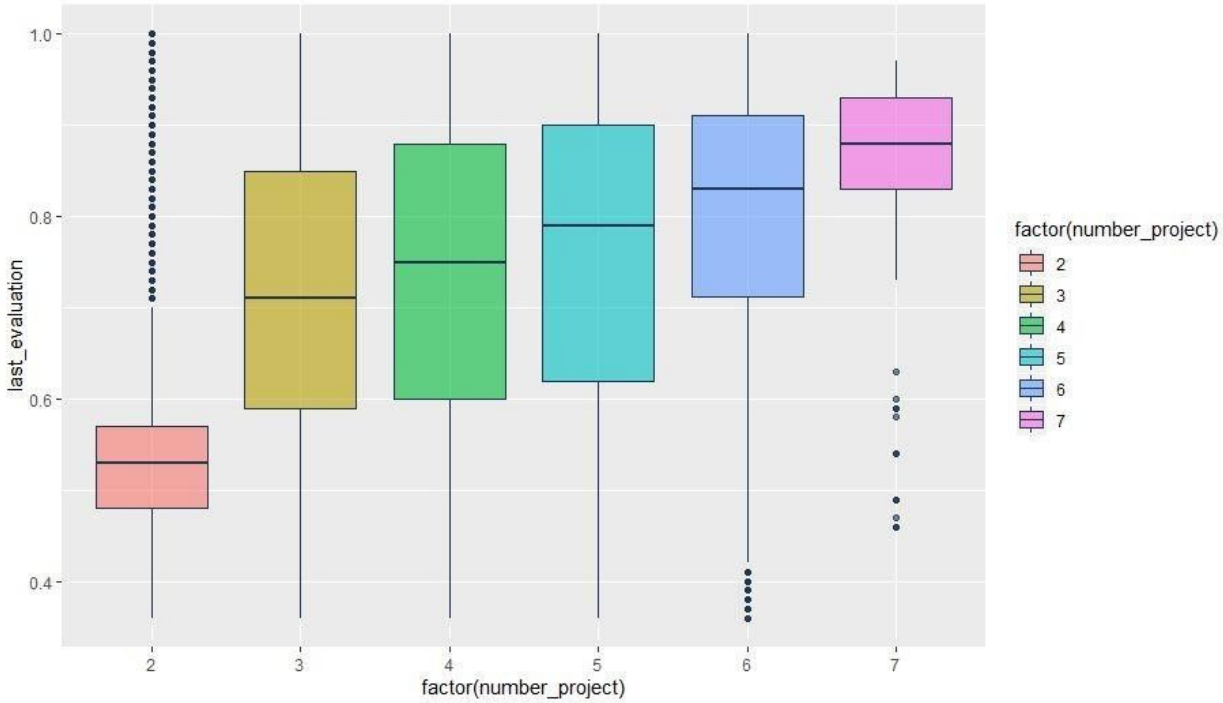


7. The relation between work accident and the employee left: The employee though met with accident, only a few left the company.

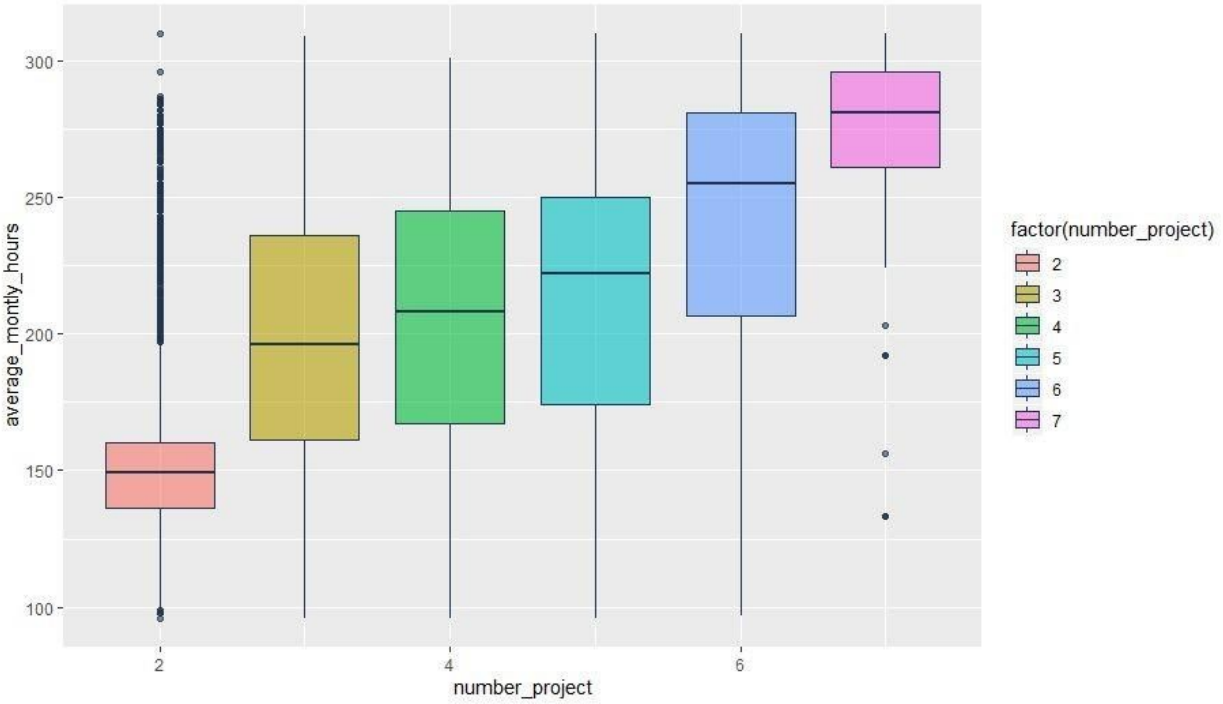


8. The relation between promotion and the employee left: Employee who got promoted never left the company





From the above plot, we can understand that people with more number of projects tend to get rated higher.



The above plot, as people with higher number of projects tend to work more hours at the office.

Model Building:

The data has a total of 10 attributes. We need to predict why the experience employees are leaving the company. For this, we need to consider **left** as a dependent variable (response). All the remaining 9 variables are the independent variables (regressors)

The dependent variable left constitutes only two values. Either 0 and 1. It defines if the employee has left the company or not

1 - employee left the company

0 - the employee did not leave the company.

As the response variable has two values, the logistic regression model has been built.

Model 1: Consider all the dependent variables as numerics. (To predict model)

```
#1. All varaibles - numeric
hrdata$sales = as.numeric(hrdata$sales)
hrdata$salary = as.numeric(hrdata$salary)

model = glm(left ~ ., data = hrdata, family = "binomial")
summary(model)

prob = predict(model, hrdata, type = "response")
pred_values = ifelse(prob>=0.5,1,0)

confusion = table(hrdata$left, pred_values)
confusion
Accuracy = sum(diag(confusion)/sum(confusion))
Accuracy
sensitivity(confusion)
specificity(confusion)

roc = roc(model$y, pred_values)
plot(roc)
auc(roc)
```

The following image shows the summary of the model. This model has considered all the independent variables as numeric.

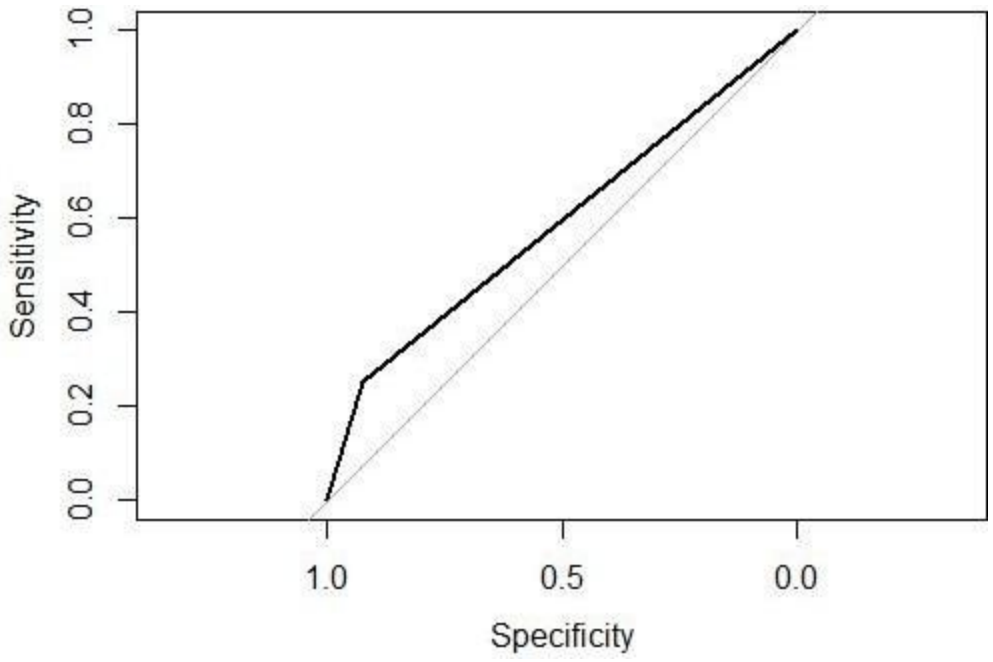
All independent variables are significant. The accuracy of the model is: 76.5584% The sensitivity of the model is: 79.8566% The specificity of the model is: 51.5723% The Area under curve is: 58.92%

The AIC of the model is: 13343

The model can be improved by increasing accuracy and decreasing the standard errors.
One of the assumptions of improvement is converting the numeric variables to factor variables

```
Call:
glm(formula = left ~ ., family = "binomial", data = hrdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3568  -0.6819  -0.4343  -0.1533   3.1068

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.054122   0.151993   0.356  0.72178
satisfaction_level     -4.129254   0.096584 -42.753  < 2e-16 ***
last_evaluation         0.762165   0.145708   5.231 1.69e-07 ***
number_project         -0.310068   0.020850 -14.872  < 2e-16 ***
average_montly_hours    0.004346   0.000504   8.624  < 2e-16 ***
time_spend_company      0.228638   0.014855  15.391  < 2e-16 ***
Work_accident1         -1.498575   0.088254 -16.980  < 2e-16 ***
promotion_last_5years1 -1.768024   0.255495  -6.920 4.52e-12 ***
sales                   0.020587   0.007854   2.621  0.00876 **
salary                  0.011953   0.035040   0.341  0.73300
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16465  on 14998  degrees of freedom
Residual deviance: 13323  on 14989  degrees of freedom
AIC: 13343

Number of Fisher Scoring iterations: 5
```

ROC curve:



Model 2: Consider all the dependent variables and convert the numeric variables to factor variables. (To improve accuracy)

```
hrdata$sales = as.factor(hrdata$sales)
hrdata$salary = as.factor(hrdata$salary)
hrdata$work_accident = as.factor(hrdata$work_accident)
hrdata$promotion_last_5years = as.factor(hrdata$promotion_last_5years)
str(hrdata)
model = glm(left ~ ., data = hrdata, family = "binomial")
summary(model)

prob = predict(model, hrdata, type = "response")
pred_values = ifelse(prob>=0.5,1,0)

confusion = table(hrdata$left, pred_values)
confusion
Accuracy = sum(diag(confusion)/sum(confusion))
Accuracy
sensitivity(confusion)
specificity(confusion)

roc = roc(model$y, pred_values)
plot(roc)
auc(roc)
```

The following image shows a summary of the model. This model has considered all the independent variables as factor variables

All independent variables are significant. The accuracy of the model is: 79.2319% The sensitivity of the model is: 82.1835% The specificity of the model is: 60.9405% The Area under curve is: 64.22%

The AIC of the model is: 12888

The accuracy of the model has been improved.

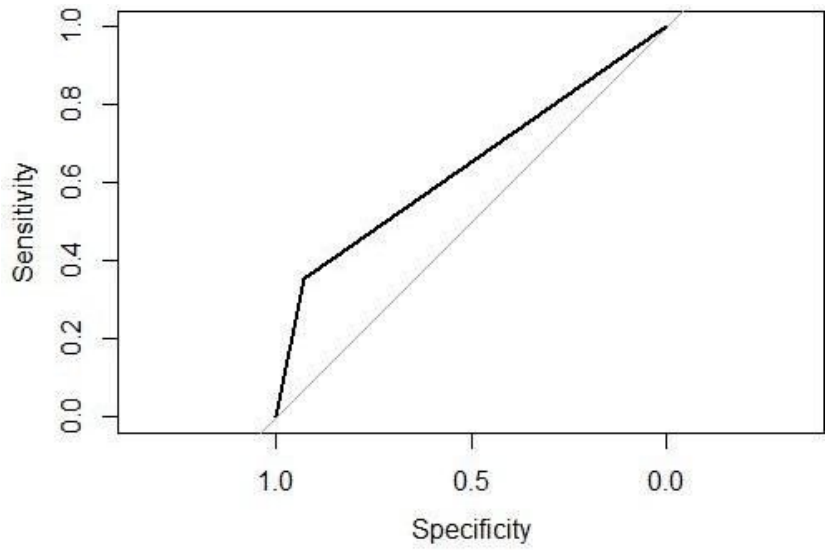
But there are insignificant variables present as the dependent variables. The standard errors also increased.

The area under the curve has been increased. So model 2 is better than model 1.

But as the variables increased, standard errors also increased. Though significant, the standard error of last evaluation and promotion for the last five years are more comparative to all the other variables. So removing these variables will decrease the standard errors.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2248  -0.6645  -0.4026  -0.1177   3.0688

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.4762862  0.1938373  -7.616 2.61e-14 ***
satisfaction_level      -4.1356889  0.0980538 -42.178  < 2e-16 ***
last_evaluation          0.7309032  0.1491787   4.900 9.61e-07 ***
number_project          -0.3150787  0.0213248 -14.775  < 2e-16 ***
average_montly_hours     0.0044603  0.0005161   8.643  < 2e-16 ***
time_spend_company       0.2677537  0.0155736  17.193  < 2e-16 ***
Work_accident1          -1.5298283  0.0895473 -17.084  < 2e-16 ***
promotion_last_5years1 -1.4301364  0.2574958  -5.554 2.79e-08 ***
saleshr                  0.2323779  0.1313084   1.770  0.07678 .
salesIT                 -0.1807179  0.1221276  -1.480  0.13894
salesmanagement         -0.4484236  0.1598254  -2.806  0.00502 **
salesmarketing          -0.0120882  0.1319304  -0.092  0.92700
salesproduct_mng        -0.1532529  0.1301538  -1.177  0.23901
salesRandD              -0.5823659  0.1448848  -4.020 5.83e-05 ***
salessales              -0.0387859  0.1024006  -0.379  0.70486
salessupport             0.0500251  0.1092834   0.458  0.64713
salestechnical           0.0701464  0.1065379   0.658  0.51027
salarylow                1.9440627  0.1286272  15.114  < 2e-16 ***
salarymedium             1.4132244  0.1293534  10.925  < 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: 16465  on 14998  degrees of freedom
Residual deviance: 12850  on 14980  degrees of freedom
AIC: 12888

Number of Fisher Scoring iterations: 5
```

ROC Curve:



Model 3: Removing last evaluation and promotion for last five years. (To decrease standard error)

```
model = glm(left ~ satisfaction_level +
              number_project + average_montly_hours +
              time_spend_company + work_accident+
              sales + salary, data = hrdata, family = "binomial")
summary(model)

prob = predict(model, hrdata, type = "response")
pred_values = ifelse(prob>=0.5,1,0)

confusion = table(hrdata$left, pred_values)
confusion
Accuracy = sum(diag(confusion)/sum(confusion))
Accuracy
Accuracy = sum(diag(confusion)/sum(confusion))
Accuracy
sensitivity(confusion)
specificity(confusion)

roc = roc(model$y, pred_values)
plot(roc)
auc(roc)
```

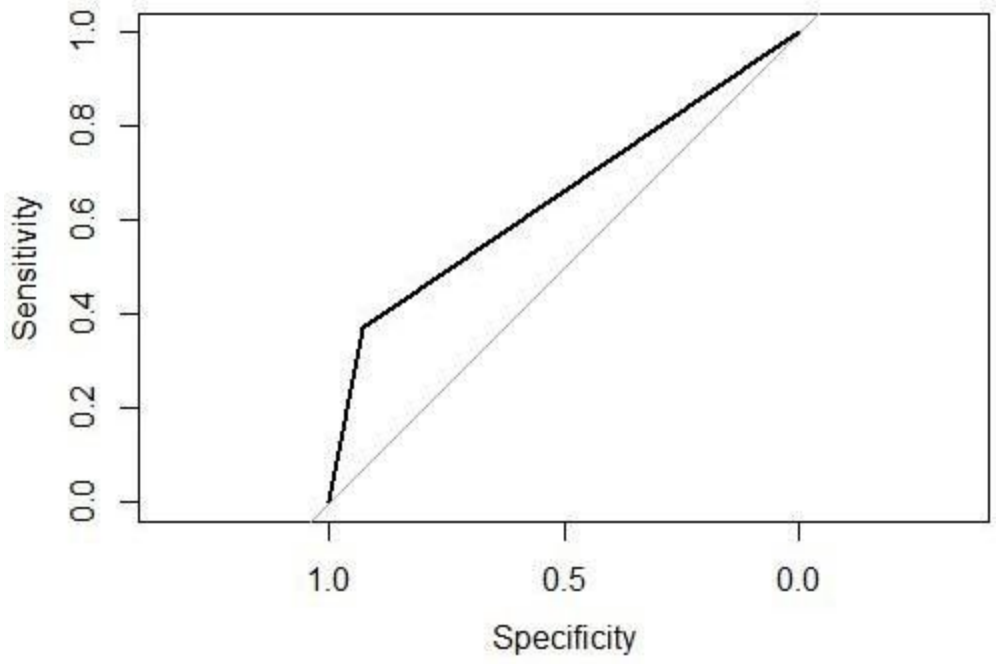
The following image shows a summary of the model. The two variables are removed as the independent variables. The standard error of the model has been decreased. Now, the standard error of all the coefficients is almost the same for the numeric variables. But the threshold of the dependent variable is not defined. We can apply the convenient threshold value to improve the accuracy and try to best fit the model

The accuracy of the model is: 79.5586% The sensitivity of the model is: 82.5217% The specificity of the model is: 61.7825% The Area under curve is: 64.95%

The AIC of the model is: 12951

```
glm(formula = left ~ satisfaction_level + number_project + average_montly_hours +
    time_spend_company + Work_accident + sales + salary, family = "binomial",
    data = hrdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2532  -0.6680  -0.4086  -0.1298   3.0409

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.3112396  0.1897405  -6.911 4.82e-12 ***
satisfaction_level   -4.0376050  0.0950288 -42.488  < 2e-16 ***
number_project       -0.2790360  0.0200221 -13.936  < 2e-16 ***
average_montly_hours  0.0050884  0.0004999  10.180  < 2e-16 ***
time_spend_company    0.2651214  0.0153721  17.247  < 2e-16 ***
Work_accident1       -1.5350368  0.0891983 -17.209  < 2e-16 ***
saleshr               0.2297243  0.1312927   1.750  0.08017 .
salesIT              -0.1658599  0.1220560  -1.359  0.17418
salesmanagement      -0.5042446  0.1590470  -3.170  0.00152 **
salesmarketing       -0.0402330  0.1315944  -0.306  0.75981
salesproduct_mng     -0.1366765  0.1301122  -1.050  0.29351
salesRandD           -0.5875253  0.1445897  -4.063 4.84e-05 ***
salessales           -0.0393777  0.1023921  -0.385  0.70055
salessupport          0.0615949  0.1092567   0.564  0.57292
salestechnical        0.0786773  0.1065376   0.738  0.46021
salarylow             1.9829278  0.1285479  15.426  < 2e-16 ***
salarymedium          1.4376334  0.1292272  11.125  < 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: 16465  on 14998  degrees of freedom
Residual deviance: 12917  on 14982  degrees of freedom
AIC: 12951

Number of Fisher Scoring iterations: 5
```

ROC Curve:



Model 4: Set the threshold value. (Improve the accuracy)

```
model = glm(left ~ satisfaction_level +
              number_project + average_montly_hours +
              time_spend_company + Work_accident+
              sales + salary, data = hrdata, family = "binomial")
summary(model)

prob = predict(model, hrdata, type = "response")
pred_values = ifelse(prob>=0.45,1,0)

confusion = table(hrdata$left, pred_values)
confusion
Accuracy = sum(diag(confusion)/sum(confusion))
Accuracy
Accuracy = sum(diag(confusion)/sum(confusion))
Accuracy
sensitivity(confusion)
specificity(confusion)

roc = roc(model$y, pred_values)
plot(roc)
auc(roc)
```

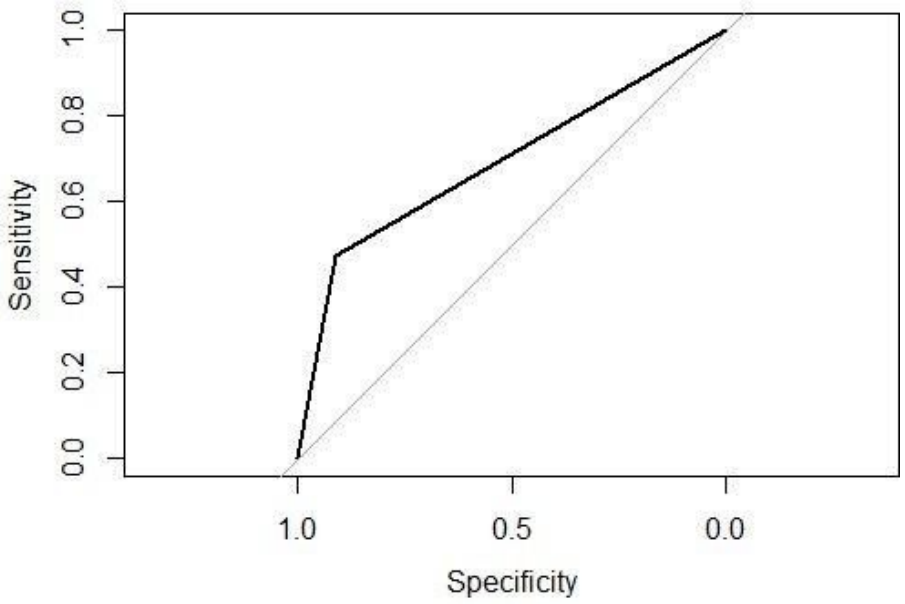
The following image shows a summary of the model.

The threshold value has been decreased. The accuracy and the area under the curve increased with the change in threshold value. The logistic function provides a sigmoid function, whose prediction values are between 0 and 1. The threshold value has been chosen as 0.5. The decrease in threshold value to 0.45 improved accuracy. Though threshold improved accuracy, the major change can overfit the model. So optimum threshold must be chosen

The accuracy of the model is: 80.5720% The
sensitivity of the model is: 84.6548% The
specificity of the model is: 62.0994% The
Area under the curve is: 69.1%

The AIC of the model is: 12951

```
glm(formula = left ~ satisfaction_level + number_project + average_montly_hours +
    time_spend_company + Work_accident + sales + salary, family = "binomial",
    data = hrdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2532  -0.6680  -0.4086  -0.1298   3.0409

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.3112396  0.1897405  -6.911 4.82e-12 ***
satisfaction_level   -4.0376050  0.0950288 -42.488  < 2e-16 ***
number_project       -0.2790360  0.0200221 -13.936  < 2e-16 ***
average_montly_hours  0.0050884  0.0004999  10.180  < 2e-16 ***
time_spend_company    0.2651214  0.0153721  17.247  < 2e-16 ***
Work_accident1       -1.5350368  0.0891983 -17.209  < 2e-16 ***
saleshr               0.2297243  0.1312927   1.750  0.08017 .
salesIT              -0.1658599  0.1220560  -1.359  0.17418
salesmanagement      -0.5042446  0.1590470  -3.170  0.00152 **
salesmarketing       -0.0402330  0.1315944  -0.306  0.75981
salesproduct_mng     -0.1366765  0.1301122  -1.050  0.29351
salesRandD           -0.5875253  0.1445897  -4.063 4.84e-05 ***
salessales           -0.0393777  0.1023921  -0.385  0.70055
salessupport          0.0615949  0.1092567   0.564  0.57292
salestechnical        0.0786773  0.1065376   0.738  0.46021
salarylow             1.9829278  0.1285479  15.426  < 2e-16 ***
salarymedium          1.4376334  0.1292272  11.125  < 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: 16465  on 14998  degrees of freedom
Residual deviance: 12917  on 14982  degrees of freedom
AIC: 12951

Number of Fisher Scoring iterations: 5
```



This model has good accuracy and less standard error but there are multiple levels in the factors which are least significant and are contributing to the errors.

In the sales factor, some of the levels like IT, marketing, product management, sales, support, technical have the least significance. Another model is built by not considering the sales factor.

Model 5: Remove sales feature (least significant level)

```
model = glm(left ~ satisfaction_level +
              number_project + average_montly_hours +
              time_spend_company + work_accident+ salary,
            data = hrdata, family = "binomial")
summary(model)

prob = predict(model, hrdata, type = "response")
pred_values = ifelse(prob>=0.45,1,0)

confusion = table(hrdata$left, pred_values)
confusion
Accuracy = sum(diag(confusion)/sum(confusion))
Accuracy
sensitivity(confusion)
specificity(confusion)
roc = roc(model$y, pred_values)
plot(roc)
auc(roc)
```

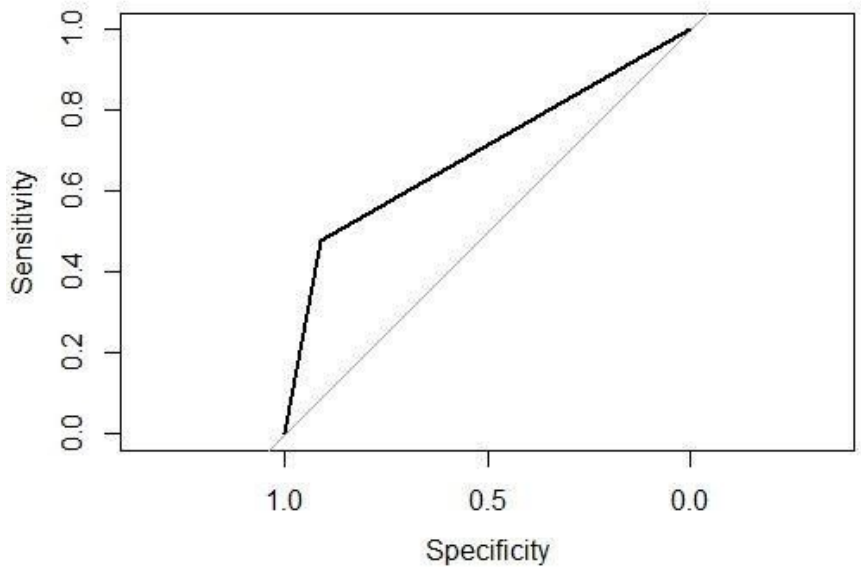
The following image shows a summary of the model.

The sales feature has been completely removed which improved accuracy, sensitivity, specificity, and the area under roc curve. The standard errors were also decreased as there are many insignificant levels in the sales factor.

The accuracy of the model is: 80.7453% The sensitivity of the model is: 84.8343%
The specificity of the model is:
62.4589% The Area under the curve is:
69.47% The AIC of the model is: 12994

```
Call:
glm(formula = left ~ satisfaction_level + number_project + average_montly_hours +
    time_spend_company + Work_accident + salary, family = "binomial",
    data = hrdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2311  -0.6676  -0.4136  -0.1322   3.0772

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.4003092  0.1671186  -8.379   <2e-16 ***
satisfaction_level   -4.0335672  0.0947571 -42.567   <2e-16 ***
number_project       -0.2788565  0.0199415 -13.984   <2e-16 ***
average_montly_hours  0.0050891  0.0004983  10.212   <2e-16 ***
time_spend_company    0.2571904  0.0151456  16.981   <2e-16 ***
Work_accident1       -1.5429590  0.0890600 -17.325   <2e-16 ***
salarylow             2.0621030  0.1272812  16.201   <2e-16 ***
salarymedium          1.5111803  0.1280808  11.799   <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: 16465  on 14998  degrees of freedom
Residual deviance: 12978  on 14991  degrees of freedom
AIC: 12994

Number of Fisher Scoring iterations: 5
```

ROC Curve:



Comparison Table:

- Model: Logistic regression model is used as the dependent variable left has two vales
- Specificity Score: It is called a true positive rate. It tells the proportion of actual true values that are correctly identified. A good model should have a high sensitivity score.
- Specificity Score: It is called as false positive rate. It tells the proportion of actual negative values that are correctly identified. A good model should have high specificity score
- Percentage of the area under ROC: Receiver Operator Characteristic curve helps in deciding the best threshold value. Higher the area under ROC, the better the model is.
- Accuracy: Defines the probability out all the correct predictions. Higher the accuracy, the better the model is.
- AIC: Akaike Information Criteria. It deals with both overfitting and underfitting.
- The low the value of AIC, the better the model is.

Model Number	Specificity Score	Sensitivity Score	Percentage of the area under curve ROC	Accuracy	AIC	Column details which you considered to build a model
1	51.5723%	79.8566%	58.92%	76.5584%	13343	All independent varaibles as numerics. (satisfaction_level + Last_evaluation+ number_project + average_montly_hours + time_spend_company + Promotion_last_5years + Work_accident + salary + sales)
2	60.9405%	82.1835%	64.22%	79.2319%	12888	All independent varaibles. Some are factors (satisfaction_level + Last_evaluation+ number_project + average_montly_hours + time_spend_company + Promotion_last_5years + Work_accident + salary + sales)
3	61.7825%	82.5217%	64.95%	79.5586%	12951	(satisfaction_level + number_project +

						average_montly_hours + time_spend_company + Work_accident + salary + sales)
4	62.0994%	84.6548%	69.1%	80.5720%	12951	(satisfaction_level + number_project + average_montly_hours + time_spend_company + Work_accident + salary + sales)
5	62.4589%	84.8343%	69.47%	80.7453%	12994	(satisfaction_level + number_project + average_montly_hours + time_spend_company + Work_accident + salary)

Model of my choice:

Considering all models, model 5 has high accuracy, specificity, and sensitivity. AIC is reduced to best fit the model. It has all the significant factors with the least standard error.

Interpretation:

- Out of the independent variables, satisfactory level, salary, time spent, hours spent, work accidents, a number of projects played a major role.
- In order to improve accuracy, the numerical variables are converted to factor variables.
- As the standard errors are high for variables like promotion and last evaluation, they are removed from the model for fewer errors.
- The threshold value is changed to get more ROC. This best fits the model.
- The AIC also decreased so that to ensure there is no overfit or underfit.
- Though the model has only one significant level from the sales variable, dropping the entries will lead to underfitting of data. So the overall model was built by not considering the sales factor.

Predictions:

A person with a low satisfaction level, undertaking less number of projects, with fewer hours spent on average, and whose salary is low is likely to leave a company.

```
new_observation = data.frame(satisfaction_level = 0.3, number_project = 2,
                             average_montly_hours = 100, time_spend_company = 5,
                             work_accident = 0, salary = "low")
p = predict(model, new_observation)
p
```

The probability of p is 0.68 which states that the person will leave the company.