ITM883 Business Analytics Problem Solving Group Project

Employee Retention HR Data

Can you predict if an employee is going to leave the company?

Group 9

Barath Kurien Alapatt (alapattb@msu.edu)

Divisha Jain (jaindiv2@msu.edu)

Monika Sureshkumar (sureshk2@msu.edu)

Rohan Reddy Galipur (galipurr@msu.edu)

Sriya Kondabathula (kondabat@msu.edu)

Vamsi Gopala Krishna Kethepalli (kethepal@msu.edu)

Table of Contents

1. Problem Statement

Data Set Introduction:

Group 9 has chosen the HR Dataset, which contains data points on an employee's Satisfactory Level, Number of Projects, Average Monthly Hours, Time Spent in the Company, Promotion within the Last 5 years, and whether they have left the company in the past year.

Human Resource Company Data | Kaggle

Task:

The task is to analyze the HR dataset using descriptive and predictive analytics to determine factors affecting employee turnover. The purpose is to make informed HR decisions for improving employee engagement and retention, reducing turnover rates, and promoting a positive work culture.

Data description:

HR data set which contains 15000 rows of data and 10 potential attributes. Below Data Description table provides more information about each attribute:

Name	Type	Sample	Description
satisfaction_level	Quantitative	Continuous	Satisfaction level based on survey
last_evaluation	Quantitative	Continuous	Last evaluation score
number_project	Quantitative	Continuous	Number of projects they have worked on during the last 1 year
average_montly_hours	Quantitative	Continuous	Average monthly hours worked
time_spend_company	Quantitative	Continuous	Time spent in the company
Work_accident	Categorical / Dummy	0 or 1	Any accidents or mistakes done at work
left	Categorical / Dummy	0 or 1	Employee left or stayed in the company
promotion_last_5years	Categorical / Dummy	0 or 1	Promotion received in the last 5 years
Department	Categorical	Sales, HR, IT etc.	Department the employee belongs too
salary	Categorical	“low”,” medium” or “high”	Salary binned into low, medium and high brackets

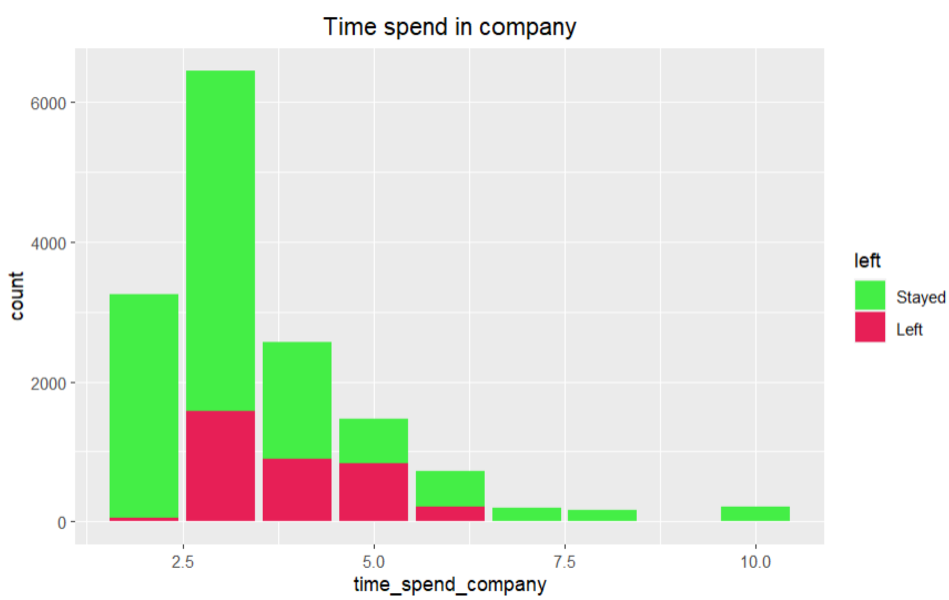
2. Exploratory Data Analysis

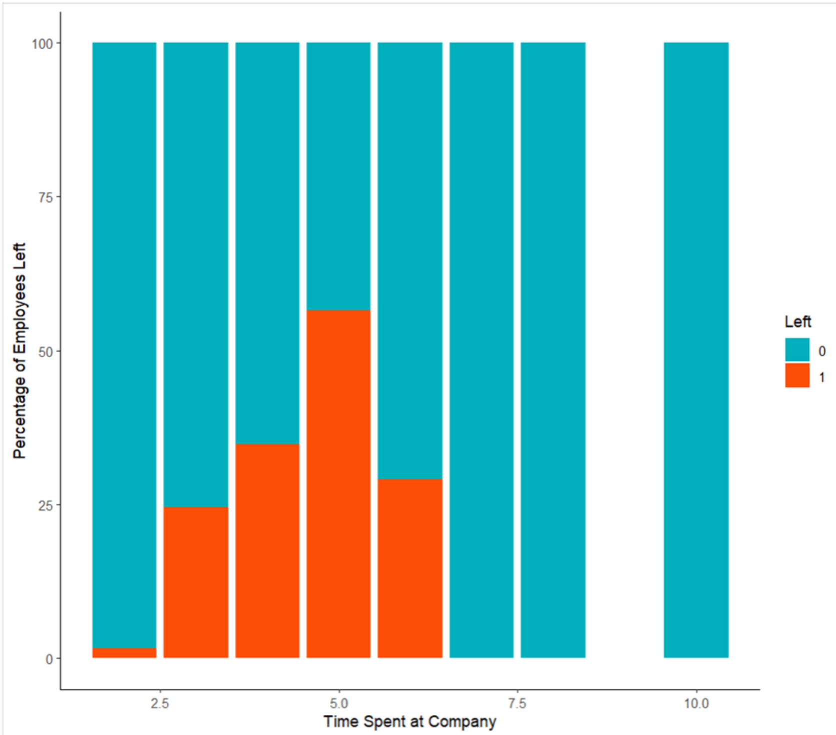
- In our data set we have a total of 9 independent variables out of which 4 are categorical and 5 are quantitative.
- We do not have missing values in our data set which is shown as below

```
> # Check for missing values
> sapply(HR_data, function(x) sum(is.na(x)))
   satisfaction_level       last_evaluation        number_project  average_montly_hours    time_spend_company
                    0                     0                     0                     0                     0
        Work_accident                  left promotion_last_5years            Department                salary
                    0                     0                     0                     0                     0
>
> # Check for duplicated rows and count them
> print(paste(sum(duplicated(HR_data)), "duplicates found."))
[1] "3008 duplicates found."
>
```

2.1 Visualizations

Graph 1: Comparing Time Spent by the employees and attrition rates



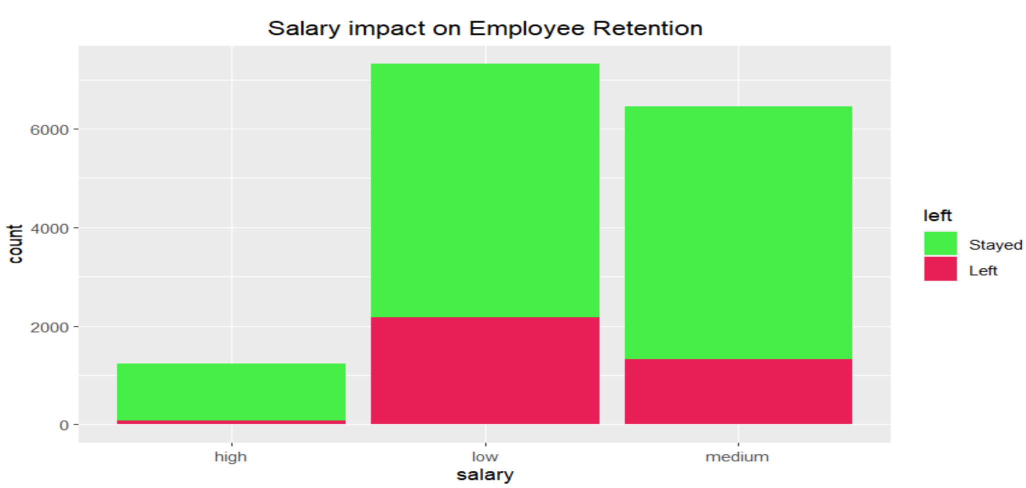


Inference

The graph above makes it evident that employees with experience of more than 6.5 years have remained with the company, while very few employees with experience of less than 2.5 years and most employees with experience of 2.5 to 6.5 years have left.

From the above graph, most of those employees who left the company are from low salary range compared to the minority of the employees are from high salary range.

Graph 2: Comparing Salary of the employees and attrition rates.

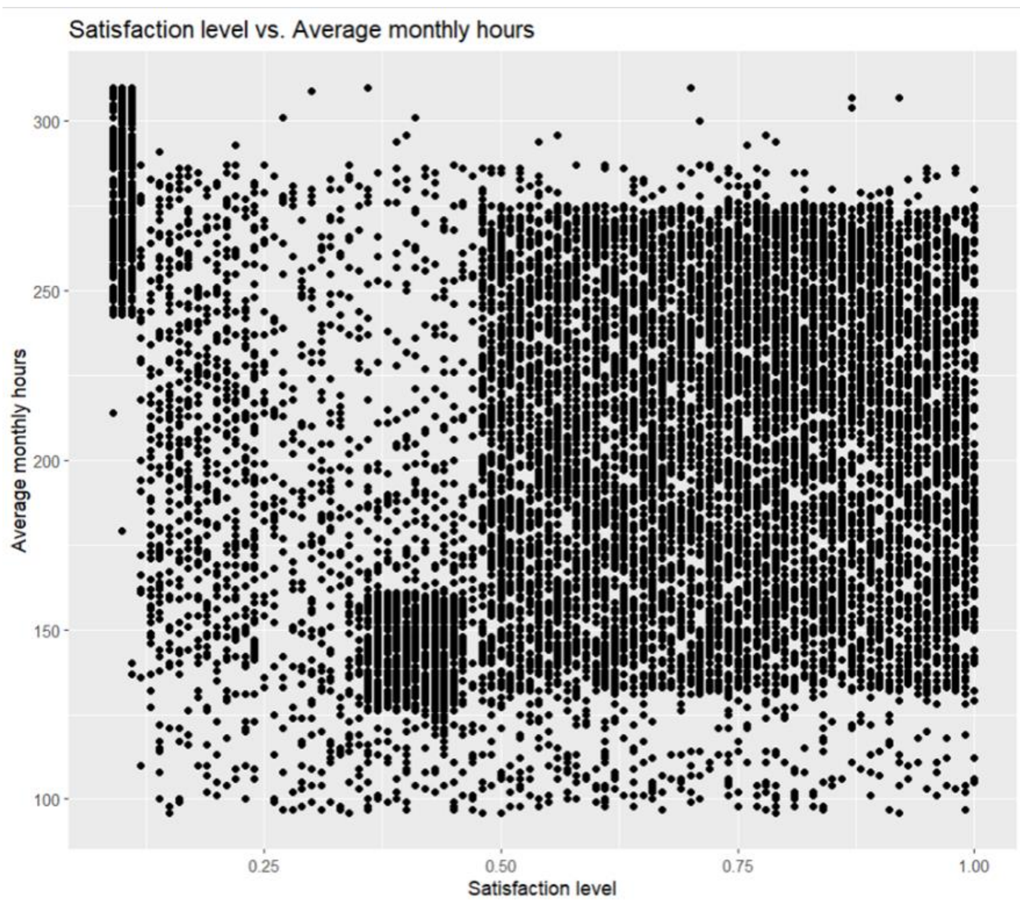


Inference

The percentage of employees leaving the company is quite low in the first two years, but it increases rapidly in the third year and then becomes constant. Employees who have spent 5 or more years at the company have a lower chance of leaving.

This suggests that employees who leave the company are more likely to do so after spending three years at the company. This may indicate that the company needs to focus on employee retention strategies for employees who have been at the company for three years or more.

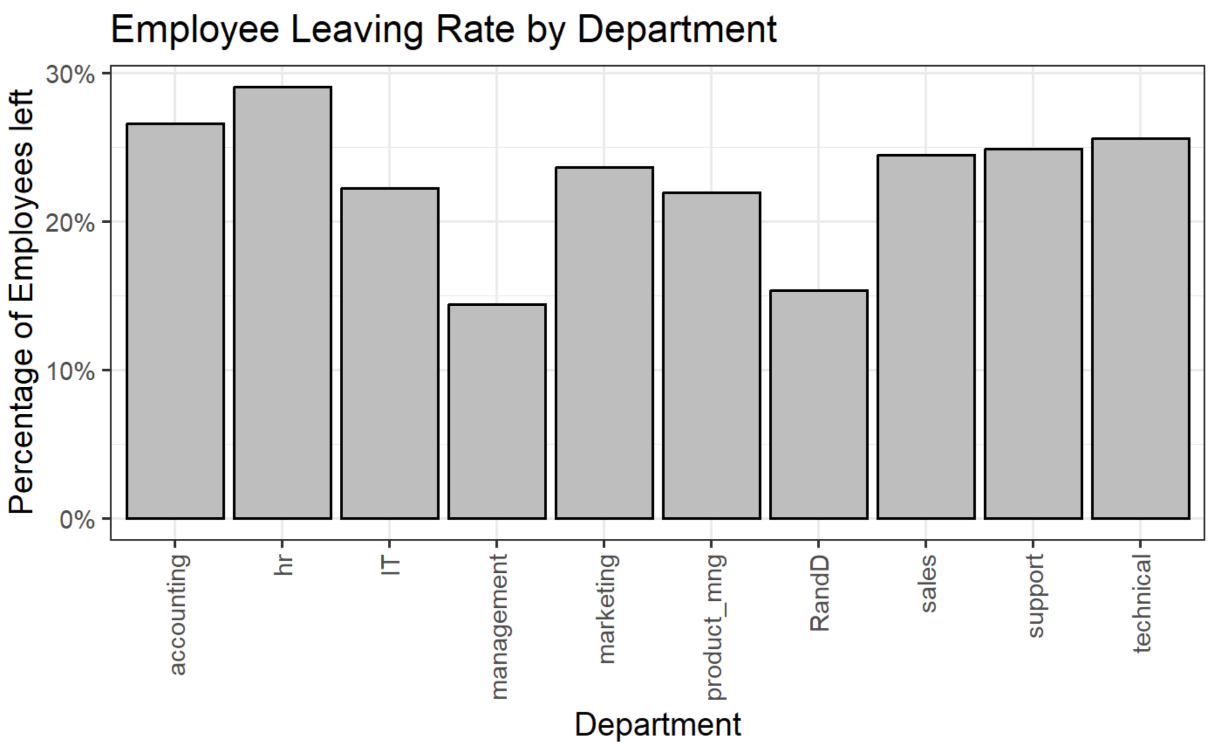
Graph 3: Checking if the employees satisfaction level is related with the average monthly hours spent by them in the company



Inference

The scatter plot shows that there is a slight positive correlation between satisfaction level and average monthly hours. Employees who work longer hours tend to have higher satisfaction levels.

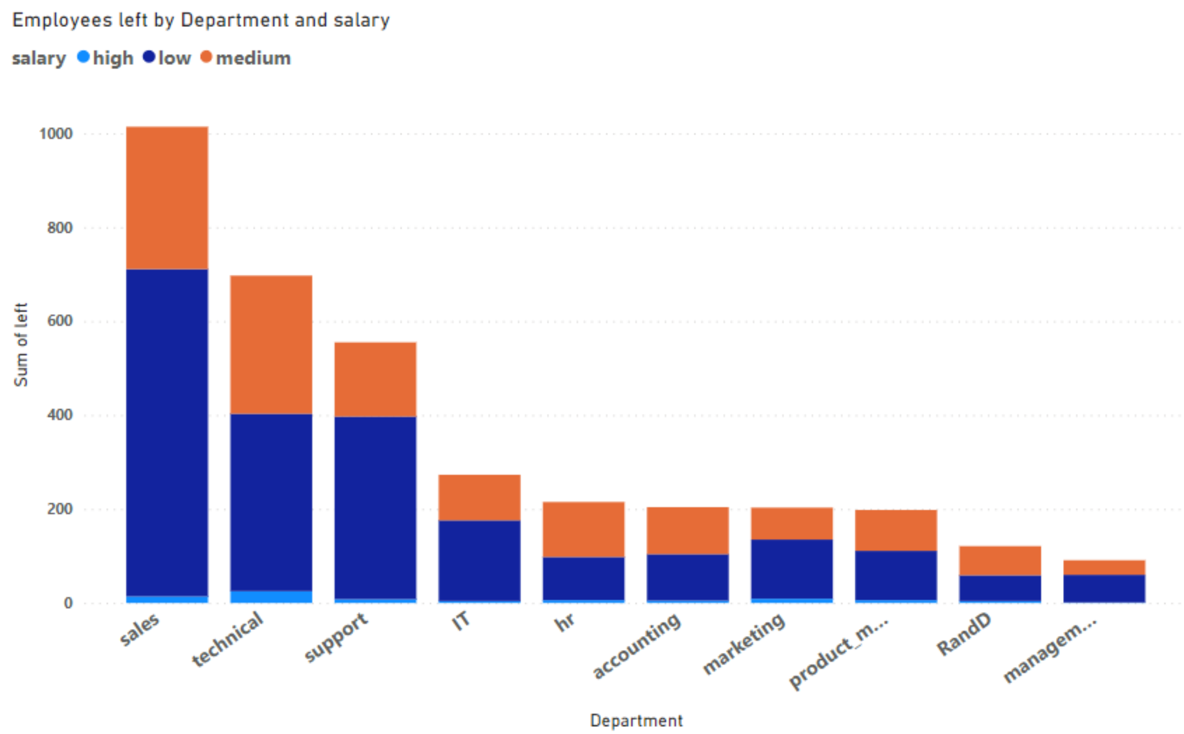
Graph 4: How do the attrition rates differ within various departments?



Inference

The above bar graph provides an overview of the employee leaving rate by the departments. The two departments with the highest employee leaving rate are HR and accounting whereas the lowest rate is shown in the management and R&D departments.

Graph 5: How do the attrition rates differ within various departments and salary levels?



Inference

From the above Stacked column chart, we can see which departments had the highest number of employees who left and in which salary levels. Here, in the sales department, a higher percentage of employees with low and medium salaries left compared to those with high salaries.

It also allows you to compare the number of employees who left in different departments and salary levels, which can help identify any patterns or trends which exist.

2.2 Summary of Dataset

```
summary(HR_data)
```

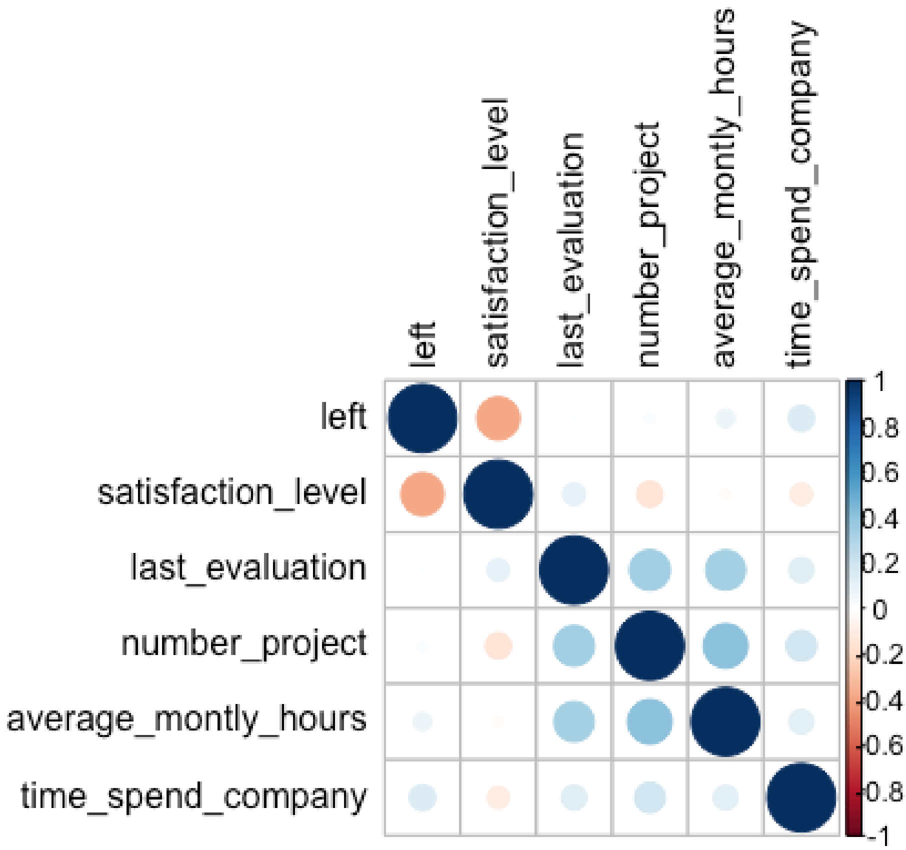
```
##  satisfaction_level last_evaluation  number_project  average_montly_hours
##  Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0
##  1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0
##  Median :0.6400     Median :0.7200   Median :4.000   Median :200.0
##  Mean   :0.6128     Mean   :0.7161   Mean   :3.803   Mean   :201.1
##  3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0
##  Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0


##  time_spend_company Work_accident        left        promotion_last_5years
##  Min.   : 2.000     Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 3.000     Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   : 3.498     Mean   :0.1446   Mean   :0.2381   Mean   :0.02127
##  3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :10.000     Max.   :1.0000   Max.   :1.0000   Max.   :1.00000


##   Department           salary
##  Length:14999       Length:14999
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

Here we get information about common statistical methods applied to our data in order to get an overall picture about data set.

2.3 Correlogram

Visual Correlogram

```
HR_records <- subset(HR_data, select = c(left, satisfaction_level, last_evaluation, number_project, avera
ge_montly_hours, time_spend_company))
corrplot(cor(HR_records), tl.col="black")
```



Matrix based Correlogram

```
> HR_records <- subset(hr_data, select = c(left, satisfaction_level, last_evaluation, number_project, average_montl
y_hours, time_spend_company))
> corr_matrix <- cor(HR_records)
> corr_matrix
```

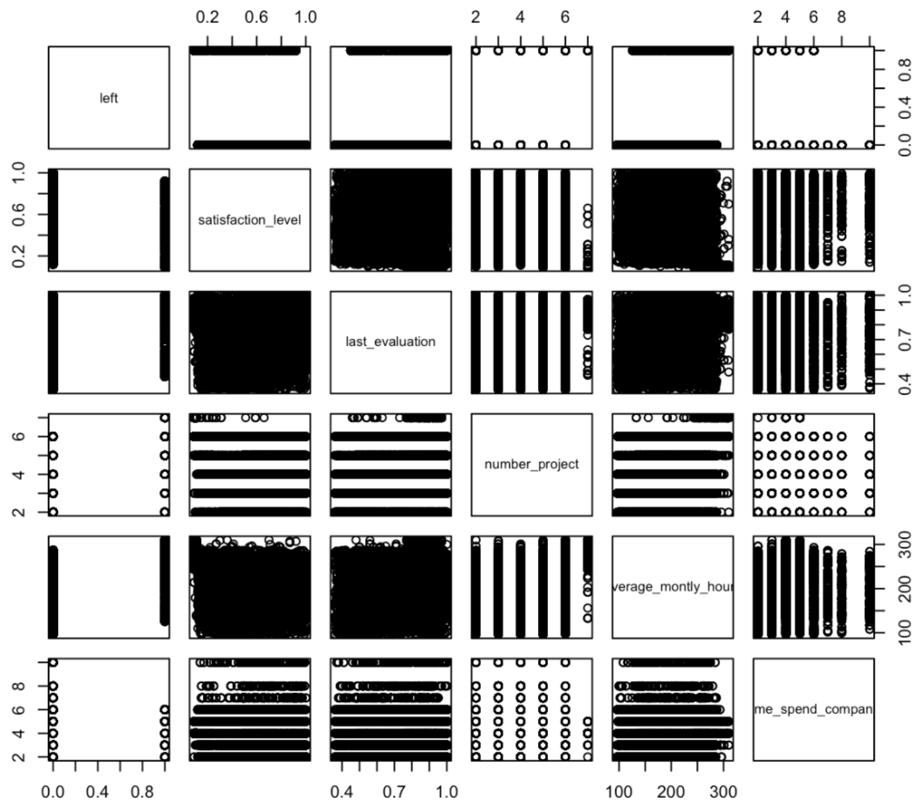
```
> corr_matrix
                            left satisfaction_level last_evaluation number_project
left                  1.00000000        -0.38837498      0.00656712     0.02378719
satisfaction_level   -0.38837498         1.00000000      0.10502121    -0.14296959
last_evaluation       0.00656712         0.10502121      1.00000000     0.34933259
number_project        0.02378719        -0.14296959      0.34933259     1.00000000
average_montly_hours  0.07128718        -0.02004811      0.33974180     0.41721063
time_spend_company    0.14482217        -0.10086607      0.13159072     0.19678589
                     average_montly_hours time_spend_company
left                           0.07128718          0.1448222
satisfaction_level            -0.02004811         -0.1008661
last_evaluation                0.33974180          0.1315907
number_project                 0.41721063          0.1967859
average_montly_hours           1.00000000          0.1277549
time_spend_company             0.12775491          1.0000000
```

Correlogram Interpretation

The correlogram represents the correlations for all pairs of quantitative variables. Positive correlations are displayed in blue and negative correlations in red. The intensity of the color is proportional to the correlation coefficient so the stronger the correlation (i.e., the closer to -1 or 1), the darker the boxes. So according to our correlogram we can see that highest positive correlation is between average_monthly_hours and number_project. And there is some medium positive correlation between last_evalution and number_project. There is strong negative strong correlation between satisfaction_level and left.

Scatter Plot Matrix

```
pairs(HR_records)
```

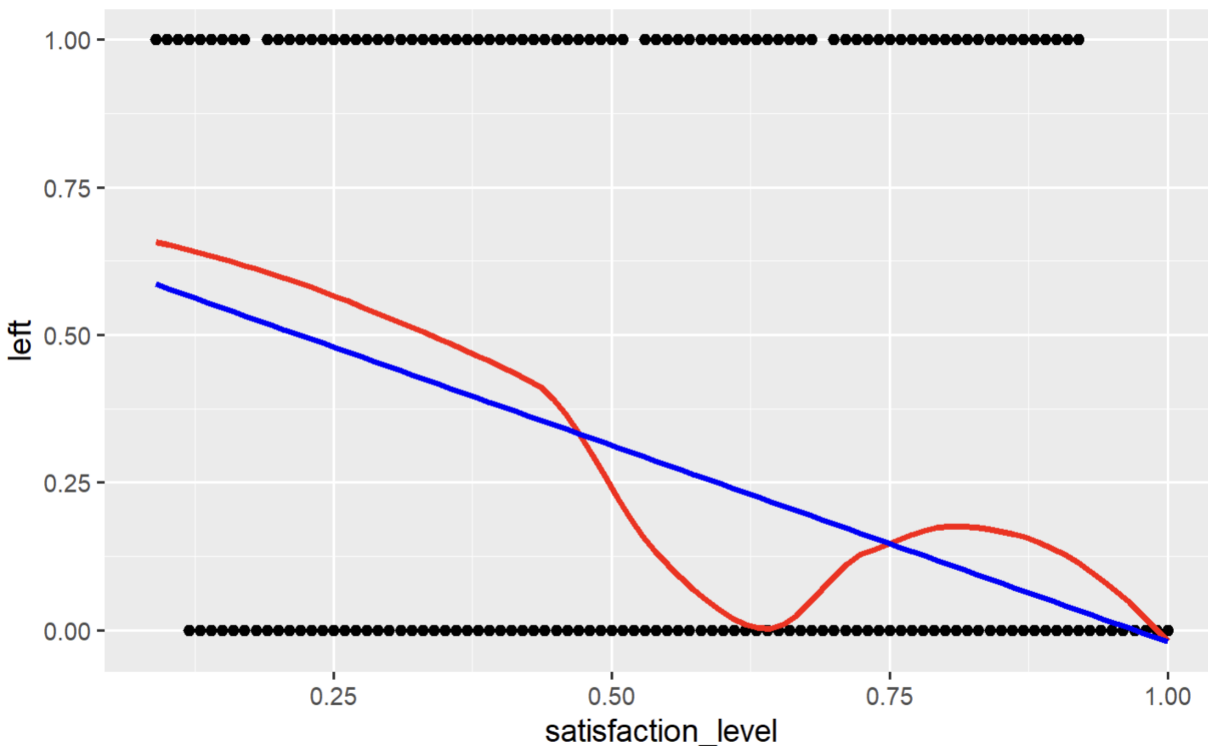


Scatter Plot Matrix Interpretation

Scatterplot matrices are a great way to roughly determine if you have a linear correlation between multiple variables. In our case we get very busy scatterplot matrices.

2.4 Drawing LOESS curves

```
library(ggplot2)

gplot(ggplot(data = hr_data) +
  geom_point(mapping = aes(x=  satisfaction_level, y= left)) +
  geom_smooth(mapping = aes(x=  satisfaction_level, y= left), method = "loess", se = FALSE, colo
r = "red") +
  geom_smooth(mapping = aes(x=  satisfaction_level, y= left), method = "lm", se = FALSE, color=
"blue")
```



We can observe that as the satisfaction level of the employee increases, the attrition rate decreases.

3. Selecting best features from the dataset

We aim to identify the optimal subset of predictor variables and obtain the best model using the "glmulti" package in R.

```r
install.packages("glmulti")

library(glmulti)

glmulti.logistic.out <-
    glmulti(left ~., data = data,
            level = 1,               # No interaction considered
            method = "h",            # Exhaustive approach
            crit = "aic",            # AIC as criteria
            confsetsize = 5,         # Keep 5 best models
            plotty = F, report = F,  # No plot or interim reports
            fitfunction = "glm",     # glm function
            family = binomial)       # binomial family for logistic regression

## Show 5 best models (Use @ instead of $ for an S4 object)
glmulti.logistic.out@formulas
summary(glmulti.logistic.out@objects[[1]])
MyROC <- roc(test$left ~ PredictedProb)
```

```
-2.3630  -0.6823  -0.4345  -0.1526   3.1097

Coefficients:
                      Estimate Std. Error z value             Pr(>|z|)
(Intercept)          0.0815369  0.1290068   0.632              0.52736
satisfaction_level  -4.1287921  0.0965692 -42.755 < 0.0000000000000002 ***
last_evaluation      0.7624413  0.1457099   5.233      0.0000001671442 ***
number_project      -0.3099587  0.0208455 -14.869 < 0.0000000000000002 ***
average_montly_hours 0.0043453  0.0005039   8.623 < 0.0000000000000002 ***
time_spend_company   0.2286246  0.0148556  15.390 < 0.0000000000000002 ***
Work_accident       -1.4987312  0.0882561 -16.982 < 0.0000000000000002 ***
promotion_last_5years -1.7694762 0.2555546  -6.924      0.00000000000439 ***
Department           0.0205877  0.0078539   2.621              0.00876 **
---
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 14990  degrees of freedom
AIC: 13341

Number of Fisher Scoring iterations: 5
```

We obtain the most important features as per the model above are satisfaction level, last evaluation, number of projects, average monthly hours, time spent in the company, no of work accidents, promotion in the last 5 years and departments.

4. Building Classification Models

Logistic Regression Model

MODEL 1:

Logistic regression model where only one independent variable salary is considered.

```
model2 <- glm(left ~ salary, data = train_hr, family = binomial(link = 'logit'))
summary(model2)
```

```
Call:
glm(formula = left ~ salary, family = binomial(link = "logit"),
    data = train_hr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8296  -0.8296  -0.6844  -0.3975   2.2706

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.4987     0.1281 -19.513   <2e-16 ***
salarylow      1.6090     0.1317  12.217   <2e-16 ***
salarymedium   1.1663     0.1332   8.756   <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: 11489  on 10468  degrees of freedom
Residual deviance: 11232  on 10466  degrees of freedom
AIC: 11238

Number of Fisher Scoring iterations: 5
```

Interpretations

1. Interpreting 1.6090: We are able to conclude from the smaller p-value (<0.05) for the coefficient that there is a statistically significant difference in the probability of an employee leaving company when comparing salary(low) and salary(high) as reference group. It can be quantified as, without accounting for any other variables, we can say that the odds of an employee leaving the company are 5 times higher for low salaried employees than high salaried employees.

2. Interpreting 1.1663: We are able to conclude from the smaller p-value (<0.05) for the coefficient that there is a statistically significant difference in the probability of an employee leaving company when comparing salary(medium) and salary(high) as reference group. It can be quantified as, without accounting for any other variables, we can say that the odds of an employee leaving the company are 3.21 times higher for medium salaried employees than high salaried employees.

Calculations:

Odds ratio = exp (1.6090) = 4.99

4.99 = (odds of salary(low))/ (odds of salary(high))

Odds ratio = exp (1.1663) = 3.21

3.21 = (odds of salary(medium))/ (odds of salary(high))

MODEL 2:

Logistic regression model predicting the probability of employee leaving (left) where the independent variables considered are satisfaction_level, last_evaluation, number_project, average_montly_hours, time_spend_company, Work_accident, promotion_last_5years, Department

```
#train-test split
set.seed(2)
train.index=sample(c(TRUE,FALSE),prob = c(0.7, 0.3), nrow(HR_data),replace=TRUE)
train_hr=HR_data[train.index,]
test_hr=HR_data[!train.index,]

model1 <- glm(left ~ satisfaction_level+last_evaluation+number_project
+average_montly_hours+time_spend_company+Work_accident+promotion_last_5years
+Department, data = train_hr, family = binomial(link = 'logit'))
summary(model1)

#prediction
log_odds <- predict(model1, test_hr)
odds <- exp(log_odds)
test_hr$prob <- odds/(1+odds)
test_hr_pred<- ifelse(test_hr$prob > 0.5, "1", "0")
confusion_matrix = table(actual = test_hr$left, predicted = test_hr$pred)
confusion_matrix

#ROC curve
library(pROC)
MyROC <- roc(test_hr$left ~ test_hr$prob)
plot(MyROC)
coords(MyROC, "best")
MyROC
```

```
Call:
glm(formula = left ~ satisfaction_level + last_evaluation + number_project +
    average_montly_hours + time_spend_company + Work_accident +
    promotion_last_5years + Department, family = binomial(link = "logit"),
    data = train_hr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2899  -0.6750  -0.4290  -0.1558   3.0969

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.1606815  0.1778024   0.904  0.36615
satisfaction_level    -4.1141835  0.1157915 -35.531  < 2e-16 ***
last_evaluation        0.7311437  0.1755319   4.165 3.11e-05 ***
number_project        -0.3287573  0.0250657 -13.116  < 2e-16 ***
average_montly_hours   0.0045936  0.0006049   7.594 3.10e-14 ***
time_spend_company     0.2577722  0.0182344  14.137  < 2e-16 ***
Work_accident1        -1.4189243  0.1036048 -13.696  < 2e-16 ***
promotion_last_5years1 -1.4308295  0.2813258  -5.086 3.66e-07 ***
Departmenthr           0.1337143  0.1597385   0.837  0.40255
DepartmentIT          -0.1380584  0.1458582  -0.947  0.34388
Departmentmanagement  -0.7579720  0.1864040  -4.066 4.78e-05 ***
Departmentmarketing   -0.0849917  0.1582863  -0.537  0.59130
Departmentproduct_mng -0.0261000  0.1536304  -0.170  0.86510
DepartmentRandD       -0.5643253  0.1723656  -3.274  0.00106 **
Departmentsales        0.0116562  0.1228552   0.095  0.92441
Departmentsupport      0.1394821  0.1306284   1.068  0.28562
Departmenttechnical    0.1076794  0.1273737   0.845  0.39790
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11488.6  on 10468  degrees of freedom
Residual deviance:  9256.9  on 10452  degrees of freedom
AIC: 9290.9

Number of Fisher Scoring iterations: 5
```

The coefficients for each predictor variable, along with their corresponding standard errors, z-scores, and p-values are displayed. A negative coefficient value for a predictor indicates that as the predictor increases, the probability of an employee leaving decreases, and vice versa. For example, a negative coefficient for satisfaction level (-4.114) suggests that as employee satisfaction level increases, the probability of leaving decreases.

The model was then used to make predictions on a test dataset, and the resulting confusion matrix (table of actual vs. predicted values) is given below:

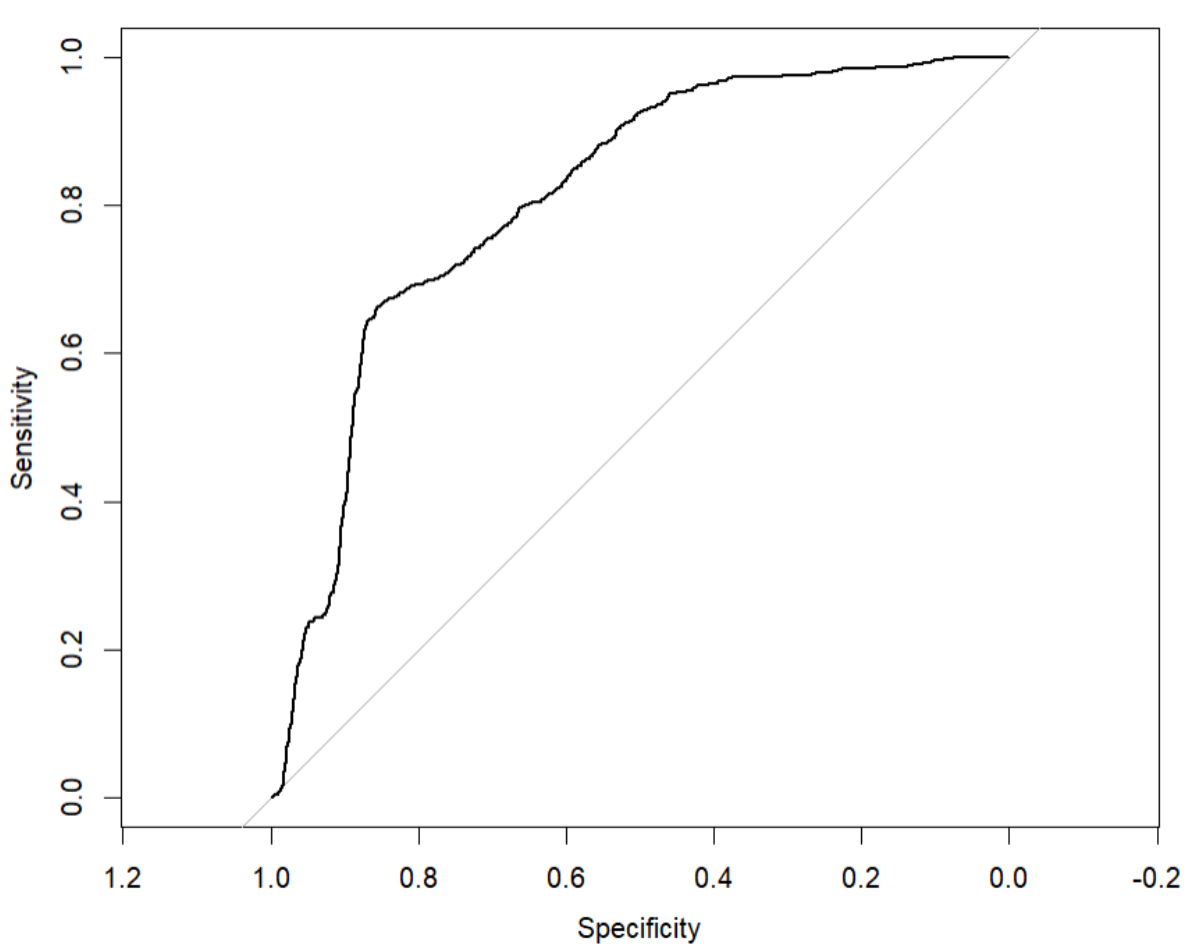
```
> confusion_matrix
       predicted
actual     0    1
      0 3201  249
      1  814  266
```

```
> coords(MyROC, "best")
  threshold specificity sensitivity
1 0.3448859   0.8556522   0.662963
> MyROC

Call:
roc.formula(formula = test_hr$left ~ test_hr$prob)

Data: test_hr$prob in 3450 controls (test_hr$left 0) < 1080 cases (test_hr$left 1).
Area under the curve: 0.8101
```

AUC is 0.8101, which suggests a good predictive performance



MODEL 3:

Logistic regression model predicting the probability of employee leaving (left) where the independent variables considered are Departments and the amount of time spent by that employee in the company.

We want to interpret the odds of an employee leaving the employee against various departments and draw a comparison between them.

```
model4 <- glm(left ~ Department+time_spend_company, data = train, family = binomial)
summary(model4)
```

```
Call:
glm(formula = left ~ Department + time_spend_company, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4179  -0.7459  -0.6705  -0.4745   2.2172

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.90668    0.11566 -16.485  < 2e-16 ***
Departmenthr           0.07972    0.14376   0.555   0.5792
DepartmentIT          -0.22241    0.13146  -1.692   0.0907 .
Departmentmanagement  -0.94373    0.17244  -5.473 4.43e-08 ***
Departmentmarketing   -0.25123    0.14239  -1.764   0.0777 .
Departmentproduct_mng -0.17604    0.13933  -1.264   0.2064
DepartmentRandD       -0.64276    0.15671  -4.102 4.10e-05 ***
Departmentsales       -0.11160    0.11071  -1.008   0.3134
Departmentsupport     -0.01916    0.11765  -0.163   0.8707
Departmenttechnical    0.04661    0.11437   0.408   0.6836
time_spend_company     0.24097    0.01519  15.863  < 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: 11489  on 10468  degrees of freedom
Residual deviance: 11178  on 10458  degrees of freedom
AIC: 11200

Number of Fisher Scoring iterations: 4
```

Interpretations with respect to Department:

Comparing Individuals who have been in the company for the same number of years the

Odds of employee leaving the company in the below department

Note: Accounting Department is the reference Level

Looking at the P-values of all departments, we can infer that only the Management department and R&D department are statistically significant in comparison with the department of accounting.

The interpretations are as follows:

Departmentmanagement	is 94.3% less in Management than that of accounting
DepartmentRandD	Is 12.82 times higher in the **Accounting than in the R&D**

Interpretations with respect to Time spent in the company:

Odds of renewal = e^ (-1.90668) * (e^ (0.24097)) ^ (Number of Years spent), or

Odds of renewal = 0.189 * (1.272) ^ (Number of Years spent)

Comparing Individuals who are in the same Department the odds of an employee leaving the company increases 1.27 times for every additional year.

MODEL 4:

Logistic regression model where the independent variables considered are satisfaction level and salary

```
model1 <- glm(left ~ satisfaction_level + salary, data = train_hr, family = binomial(link =
'logit'))
summary(model1)
```

```
Call:
glm(formula = left ~ satisfaction_level + salary, family = binomial(link = "logit"),
    data = train_hr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5220  -0.7026  -0.4809  -0.2097   2.7803

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.4195     0.1413  -2.969  0.00299 **
satisfaction_level  -3.8049     0.1052 -36.170  < 2e-16 ***
salarylow            1.6575     0.1370  12.100  < 2e-16 ***
salarymedium         1.2121     0.1385   8.753  < 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: 11488.6  on 10468  degrees of freedom
Residual deviance:  9694.4  on 10465  degrees of freedom
AIC: 9702.4

Number of Fisher Scoring iterations: 5
```

Interpretation:

Comparing Individuals based on the salary range:

Odds of employees leaving the company based on the salary.

Note: High salary is the reference Level

Looking at the P-values of all salary ranges, we can infer that both low salary and medium salary are statistically significant in comparison with the high salary.

The interpretations are as follows:

Between the employees of same satisfaction level, the odds of employees leaving the company with lower salary are 5.2 times higher than an employee with high salary.

Between the employees of same satisfaction level, the odds of employees leaving the company with medium salary are 3.36 times higher than an employee with high salary.

Calculations:

exp (1.6575) = 5.2

exp (1.2121) = 3.36

Interpretations with respect to satisfaction level in the company:

For individuals in the same salary range, for every increase of 10% in the satisfaction level there is a decrease of 32% in employees leaving the company.

Calculation: exp (-0.38) = 0.68

MODEL 5:

Logistic regression model predicting the probability of employees leaving where the independent variables considered are Department and salary.

```
model1 <- glm(left ~ Department+salary, data = HR_data, family = binomial(link = 'logit'))
summary(model1)
```

Results:

```
Call:
glm(formula = left ~ Department + salary, family = binomial(link = "logit"),
    data = HR_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.9416  -0.8422  -0.6806  -0.3268   2.5320  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -2.43711    0.13987 -17.424  < 2e-16 ***
Departmenthr           0.10586    0.11671   0.907  0.36441    
DepartmentIT          -0.28004    0.10818  -2.589  0.00964 ** 
Departmentmanagement  -0.46627    0.14328  -3.254  0.00114 ** 
Departmentmarketing   -0.16438    0.11627  -1.414  0.15743    
Departmentproduct_mng -0.29388    0.11622  -2.529  0.01145 *  
DepartmentRandD       -0.72712    0.12963  -5.609 2.03e-08 ***
Departmentsales       -0.15941    0.09071  -1.757  0.07885 .  
Departmentsupport     -0.14202    0.09668  -1.469  0.14183    
Departmenttechnical   -0.08945    0.09418  -0.950  0.34219    
salarylow              1.74751    0.11818  14.786  < 2e-16 ***
salarymedium           1.25203    0.11933  10.492  < 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: 15965  on 14987  degrees of freedom
AIC: 15989

Number of Fisher Scoring iterations: 5
```

Interpretations:

Based on the p-values, we can conclude from the smaller p-value (<0.05) for the coefficient that there is a statistically significant difference in the probability of an employee leaving company when comparing salary(low), salary(medium) with the salary(high) as reference group. All salary levels are significant predictors of employees leaving as the p-values are small (<0.05), with both 'low' and 'medium' salaries having a positive effect on the odds of leaving increases for both ‘low’ and ‘medium’ salaries than the 'high' salaries.

1. Interpreting salarylow(1.747): It can be quantified as, when employees from same department are considered (Department constant), we can say that the odds of an employee leaving company are approximately 6 times higher for low salaried employees than high salaried employees.

2. Interpreting salarymedium(1.25): When employees from same department are considered (Department constant), we can say that the odds of an employee leaving company are 3.49 times higher for medium salaried employees than high salaried employees.
3. We can conclude from the smaller p-value (<0.05) for the coefficient that there is a statistically significant difference in the probability of an employee leaving company when comparing departments, 'IT', 'management', 'product_mng', and 'RandD' with the reference group(accounting).

MODEL 6:

Logistic regression model predicting the probability of left where only one independent variable promotion_last_5years is considered.

```
model2 <- glm(left ~ promotion_last_5years, data = HR_data, family = binomial(link = 'logit'))
summary(model2)
```

Results:

```
Call:
glm(formula = left ~ promotion_last_5years, family = binomial(link = "logit"),
    data = HR_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.7443  -0.7443  -0.7443  -0.3504   2.3752  

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -1.14195    0.01927 -59.256  < 2e-16 ***
promotion_last_5yearsYes -1.61739    0.23735  -6.814 9.47e-12 ***
---
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: 16390  on 14997  degrees of freedom
AIC: 16394

Number of Fisher Scoring iterations: 5
```

Interpretations:

Interpreting -1.617: We can conclude from the smaller p-value (<0.05) for the coefficient that there is a statistically significant difference in the probability of an employee leaving company when comparing promoted_yes with promoted_no as reference group.

It can be quantified as the odds of an employee leaving the company are 5 times higher for not promoted employees than the employees who were promoted in the last 5 years.

Calculations:

Odds ratio = exp (-1.617) = 0.198

0.198 = (odds for promotion_last_5years_yes)/ (odds for promotion_last_5years_no)

We get,

5.037 = (odds for promotion_last_5years_no)/ (odds for promotion_last_5years_yes)

MODEL 7:

Logistic regression model predicting the probability of employee leaving (left) where all the independent variables are considered for modelling.

```
#training/test data
set.seed(2)
train2.index=sample(c(TRUE,FALSE),prob = c(0.7, 0.3), nrow(data),replace=TRUE)
train2=data[train2.index,]
test2=data[!train2.index,]

model4 <- glm(left ~ ., data = train2, family = binomial)
summary(model1)


#prediction
log_odds <- predict(model4, test2)
odds <- exp(log_odds)
PredictedProb <- odds/(1+odds)
PredictedProb
logisticprediction1 <- ifelse(PredictedProb > 0.5, "1", "0")
table(logisticprediction1, test2$left)

#obtain confusion matrix
confusion=table(actual=test2$left,predicted=logisticprediction1)
confusion
TP=confusion[2,2]
FN=confusion[2,1]
FP=confusion[1,2]
TN=confusion[1,1]
accuracy=(TP+TN)/nrow(test2)
sensitivity=TP/(TP+FN)
specificity=TN/(TN+FP)
c(accuracy,sensitivity,specificity)
```

```
> confusion
      predicted
actual    0    1
     0 3185  265
     1  690  390
> TP=confusion[2,2]
> FN=confusion[2,1]
> FP=confusion[1,2]
> TN=confusion[1,1]
> accuracy=(TP+TN)/nrow(test2)
> sensitivity=TP/(TP+FN)
> specificity=TN/(TN+FP)
> c(accuracy,sensitivity,specificity)
[1] 0.7891832 0.3611111 0.9231884
```

Interpretations

Accuracy	0.789
Sensitivity	0.361
Specificity	0.923

MODEL 8:

Logistic Regression Model predicting the probability of employee leaving (left) where all the independent variables are considered for modelling with threshold probability value.

```
#logisticprediction2 results
#predicting with updated threshold value
logisticprediction2 <- ifelse(PredictedProb > 0.321113, "1", "0")
table(logisticprediction2, test$left)
MyROC <- roc(test$left ~ PredictedProb)
plot(MyROC)
coords <- coords(MyROC, "best")
coords
MyROC
```

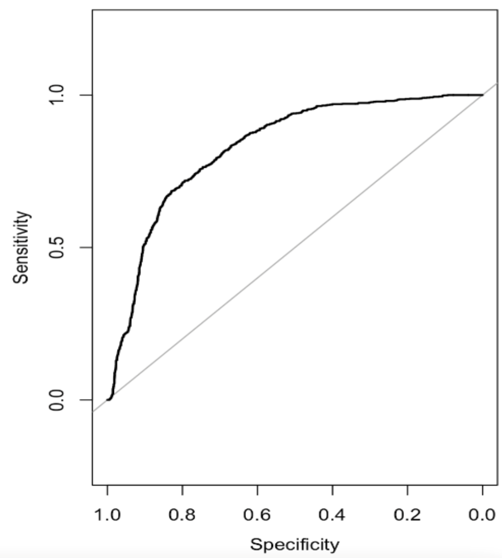
```
#obtain confusion matrix
confusion=table(actual=test2$left,predicted=logisticprediction2)
confusion
TP=confusion[2,2]
FN=confusion[2,1]
FP=confusion[1,2]
TN=confusion[1,1]
accuracy=(TP+TN)/nrow(test2)
sensitivity=TP/(TP+FN)
specificity=TN/(TN+FP)
c(accuracy,sensitivity,specificity)
```

```
> coords
  threshold specificity sensitivity
1  0.321113   0.8289855   0.6851852
> MyROC

Call:
roc.formula(formula = test$left ~ PredictedProb)

Data: PredictedProb in 3450 controls (test$left 0) < 1080 cases (test$left 1).
Area under the curve: 0.8278
```

```
> confusion
      predicted
actual    0    1
     0 2860  590
     1  340  740
> TP=confusion[2,2]
> FN=confusion[2,1]
> FP=confusion[1,2]
> TN=confusion[1,1]
> accuracy=(TP+TN)/nrow(test2)
> sensitivity=TP/(TP+FN)
> specificity=TN/(TN+FP)
> c(accuracy,sensitivity,specificity)
[1] 0.7947020 0.6851852 0.8289855
```



Interpretations

Accuracy	0.795
Sensitivity	0.685
Specificity	0.829

With the threshold p-value of 0.321, the accuracy and sum of sensitivity and specificity is improved in model 5 compared to model 4 with same independent variables.

Decision tree Model

MODEL 9:

Decision Trees Model predicting the probability of employee leaving (left) where the independent variables considered are satisfaction_level, number_project, time_spend_company, promotion_last_5years and Salary.

```
```{r}


#Decision trees with only significant variables with very less p-value
library(rpart)
tree.fit <- rpart(left ~ satisfaction_level + number_project + time_spend_company + promotion_last_5years
+ salary,  method="class", data=train, cp = 0.01)

plot(tree.fit, uniform=TRUE, main="Classification Tree for Purple")
text(tree.fit, use.n=TRUE, all=TRUE, cex=.8)

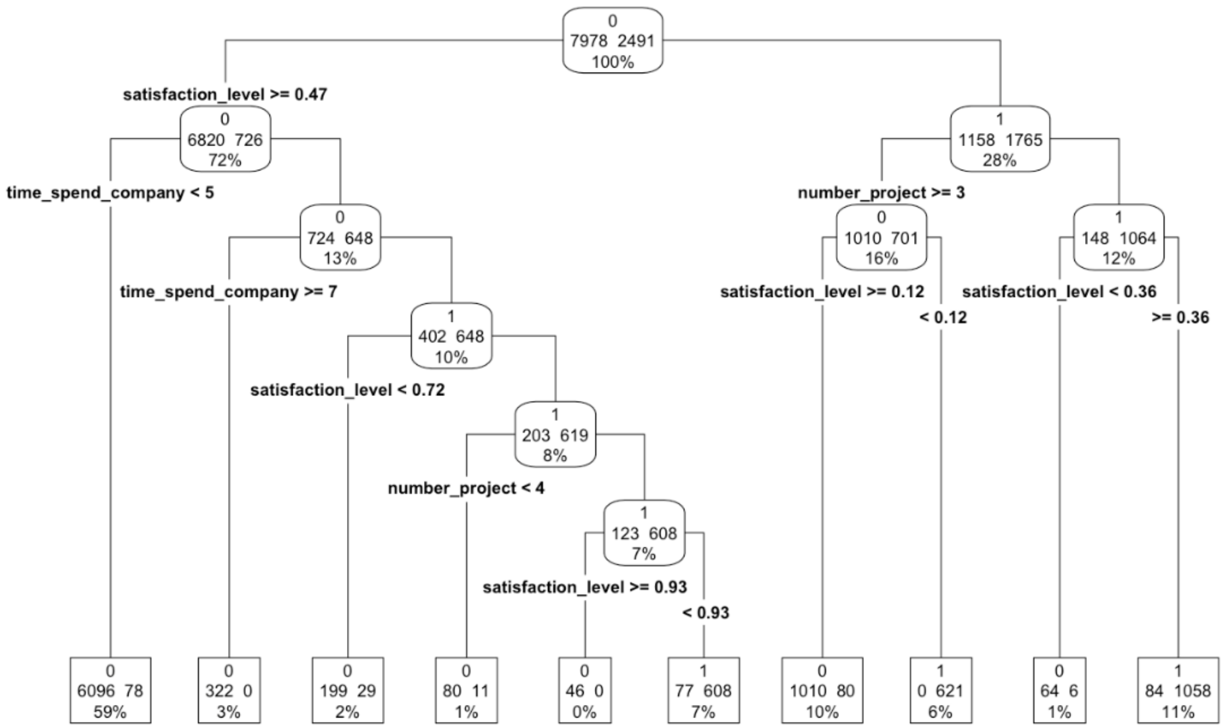
library(rpart.plot)
prp(tree.fit, type = 4, extra = 101, leaf.round = 0, fallen.leaves = TRUE, varlen = 0, tweak = 1.4)

predictions <- predict(tree.fit, newdata = test, type = "class")

#View(predictions)
# Create confusion matrix using actual class labels and predicted class labels
confusion_matrix <- table(test$left, predictions)

# Print confusion matrix
print(confusion_matrix)
acc = (confusion_matrix[2,2]+confusion_matrix[1,1])/(confusion_matrix[2,2]+confusion_matrix[1,2]+confusion_
matrix[2,1]+confusion_matrix[1,1])
print(acc)
library(pROC)

```
```




```
   predictions
       0    1
  0 3375   75
  1  103  977
[1] 0.9607064
```

MODEL 10:

Decision Trees Model predicting the probability of employee leaving (left) where all the independent variables are considered.

```
#training/test data
set.seed(2)
train.index=sample(c(TRUE,FALSE),prob = c(0.7, 0.3), nrow(HR_data),replace=TRUE)
train=HR_data[train.index,]
test=HR_data[!train.index,]

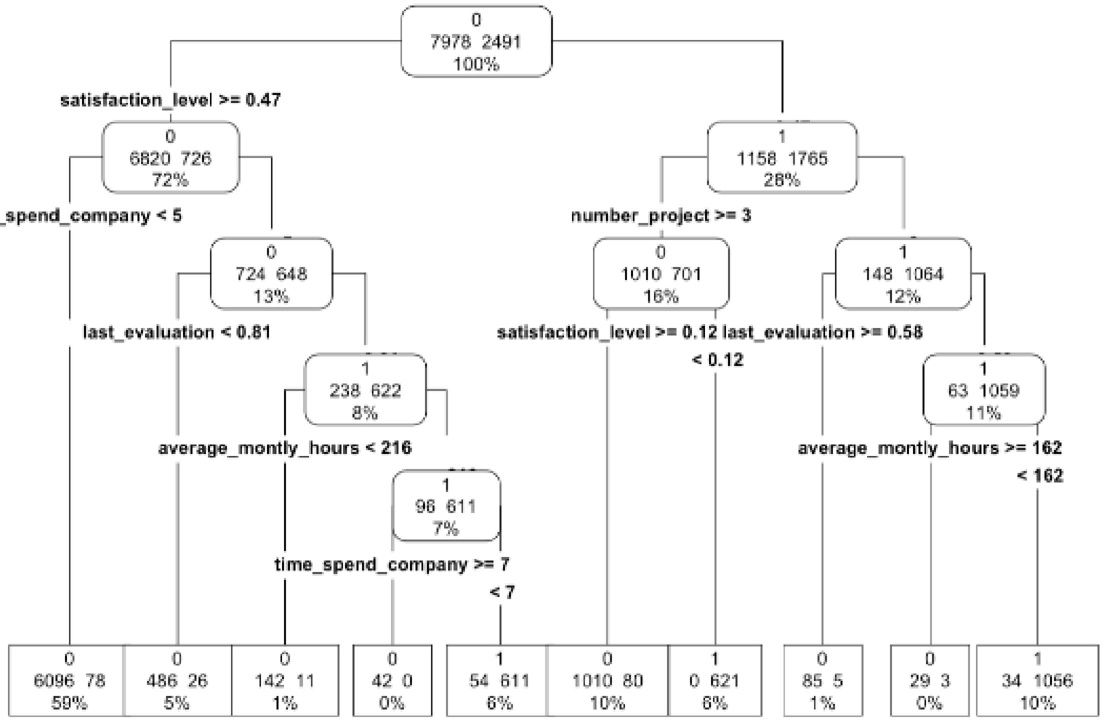
#Clasification tree
library(rpart)
library(rpart.plot)

tree.fit <- rpart(left ~ satisfaction_level + last_evaluation + number_project + average_montly_hours +
                    time_spend_company + Work_accident + promotion_last_5years + Department + salary, method = "class", data = train)

#Plot
prp(tree.fit, type = 4, extra = 101, leaf.round = 0, fallen.leaves = TRUE, varlen = 0, tweak = 1.4)

#Prediction with test data
Predictions <- rpart.predict(tree.fit, newdata = test, type="class")
#Predictions
table(Predictions, test$left)

#Confusion matrix
#obtain confusion matrix
confusion=table(actual=test$left,predicted=Predictions)
confusion
TP=confusion[2,2]
FN=confusion[2,1]
FP=confusion[1,2]
TN=confusion[1,1]
accuracy=(TP+TN)/nrow(test)
sensitivity=TP/(TP+FN)
specificity=TN/(TN+FP)
c(accuracy,sensitivity,specificity)
```




```
> #obtain confusion matrix
> confusion=table(actual=test$left,predicted=Predictions)
> confusion
      predicted
actual    0    1
     0 3410   40
     1  102  978
> TP=confusion[2,2]
> FN=confusion[2,1]
> FP=confusion[1,2]
> TN=confusion[1,1]
> accuracy=(TP+TN)/nrow(test)
> sensitivity=TP/(TP+FN)
> specificity=TN/(TN+FP)
> c(accuracy,sensitivity,specificity)
[1] 0.9686534 0.9055556 0.9884058
```

Conclusion:

Model	Accuracy
Model 7: Logistic regression (Full Model)	**78.9%**
Model 8: Logistic regression (Full Model with updated threshold)	**79.5%**
Model 9: Decision Tree (5 Variables)	**96.1%**
Model 10: Decision Tree (Full Model)	**96.8%**

The table above illustrates that the decision tree model with all variables included has the highest accuracy rate of **96.8%**.