HR

Suchitra

3/21/2017

Questions:

Why are our best and most experienced employees leaving prematurely?

Which Valuable employee will leave next

```
#Code the missing values as NA
hr_data <- read.csv("HR_comma_sep.csv", header = T, na.strings = c(""))
sapply(hr_data, function(x) sum(is.na(x))) #No missing values present in the data
```

```
##    satisfaction_level       last_evaluation        number_project
##                     0                     0                     0
##  average_montly_hours    time_spend_company         Work_accident
##                     0                     0                     0
##                  left promotion_last_5years                 sales
##                     0                     0                     0
##                salary
##                     0
```

```
#Lets explore this dataset
names(hr_data)
```

```
##  [1] "satisfaction_level"    "last_evaluation"
##  [3] "number_project"        "average_montly_hours"
##  [5] "time_spend_company"    "Work_accident"
##  [7] "left"                  "promotion_last_5years"
##  [9] "sales"                 "salary"
```

```
#Structure of the dataset
str(hr_data)
```

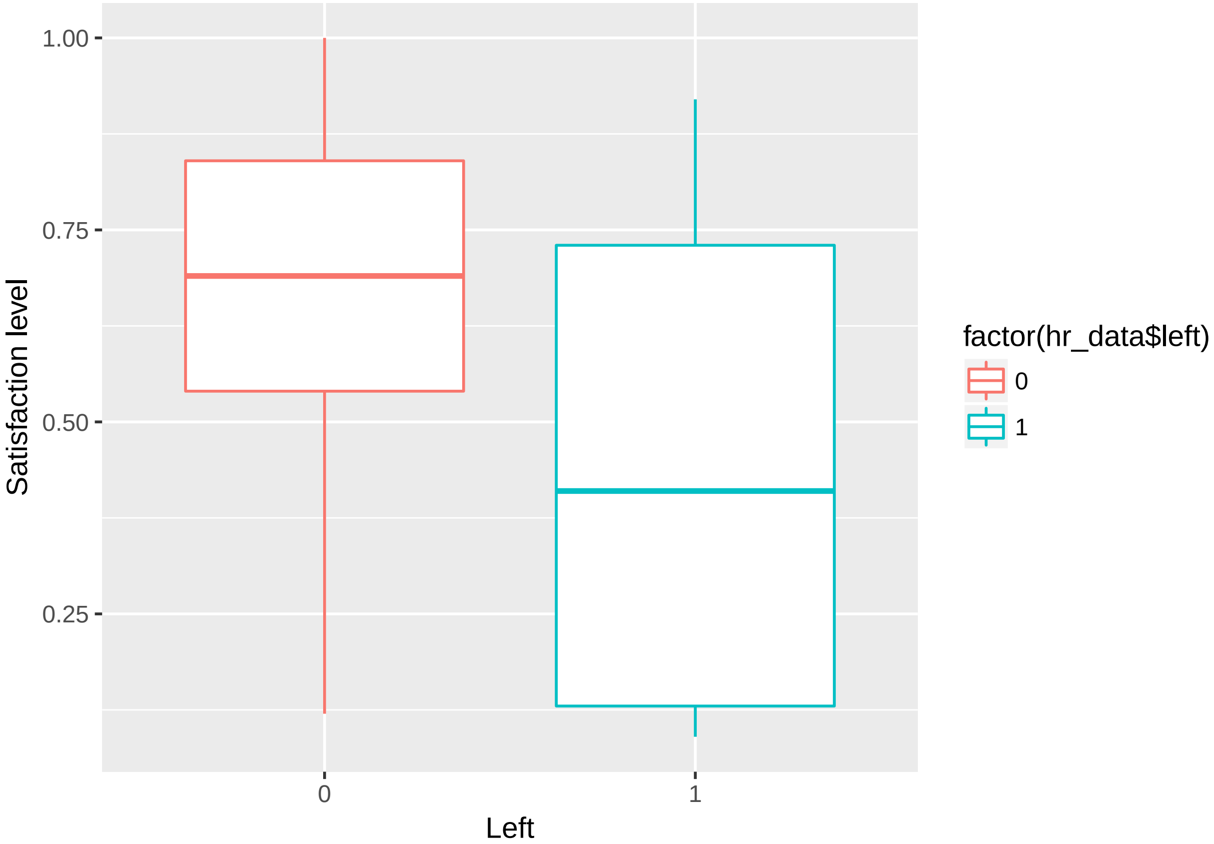
```
## 'data.frame':    14999 obs. of  10 variables:
##  $ satisfaction_level   : num  0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...
##  $ last_evaluation      : num  0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...
##  $ number_project       : int  2 5 7 5 2 2 6 5 5 2 ...
##  $ average_montly_hours : int  157 262 272 223 159 153 247 259 224 142 ...
##  $ time_spend_company   : int  3 6 4 5 3 3 4 5 5 3 ...
##  $ Work_accident        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ left                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ promotion_last_5years: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ sales                : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ salary               : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...
```

Finding the structure of the dataset gives us an information about the following: Type of dataset: Data Frame Number of variables and records Data Type of the variables: Num, int, factor Target variable : left

```
table(hr_data$left)
```

```
##
##     0     1
## 11428  3571
```

```
#Satisfaction level of people who left
ggplot(data=hr_data, aes(x=factor(hr_data$left),y=hr_data$satisfaction_level))+
  geom_boxplot(aes(color=factor(hr_data$left)))+
  xlab("Left")+
  ylab("Satisfaction level")
```



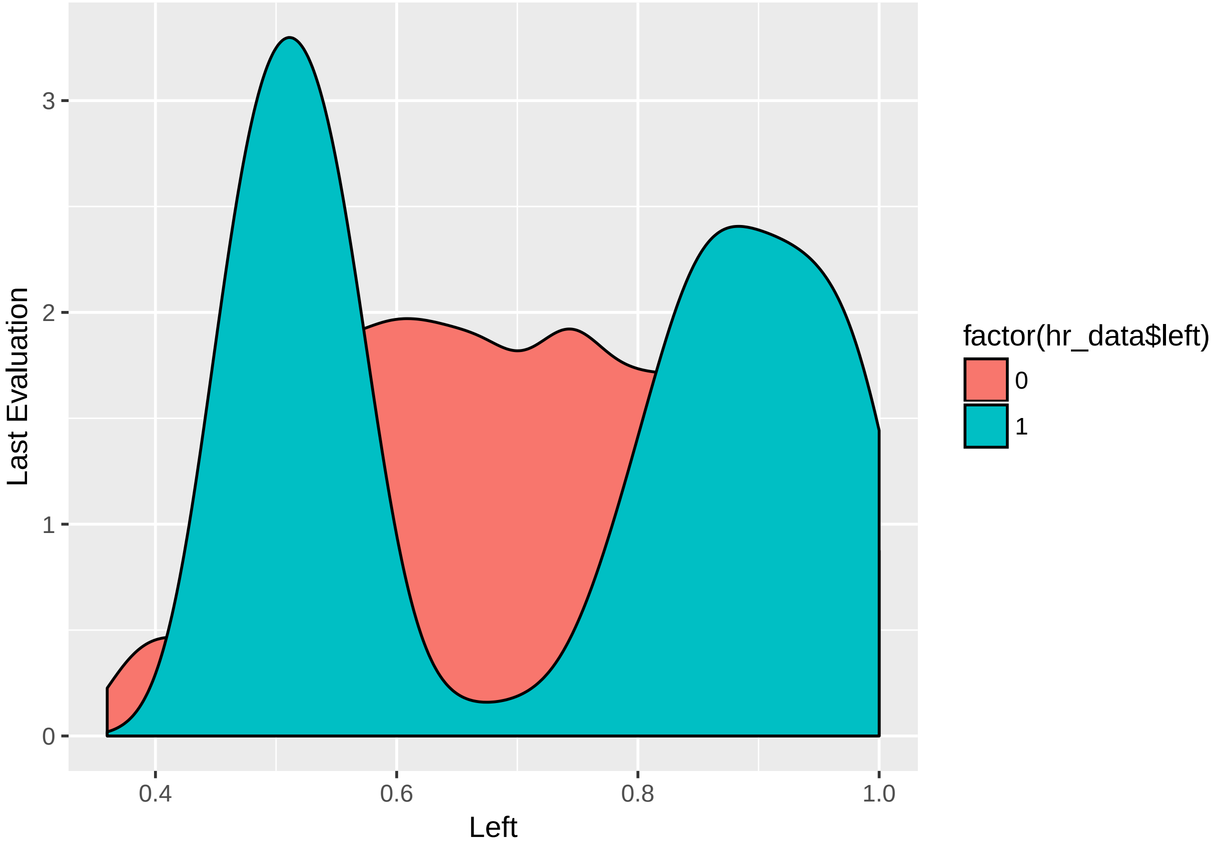
```
by(hr_data$satisfaction_level, hr_data$left, summary)
```

```
## hr_data$left: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.5400  0.6900  0.6668  0.8400  1.0000
## --------------------------------------------------------
## hr_data$left: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0900  0.1300  0.4100  0.4401  0.7300  0.9200
```

Until now, 23.8% of the people have left the company.

The satisfaction level of employees who left the company(median= 0.44) is much lower than that of the employees who stayed(0.69). This may indicate that the employees are leaving the company due to dissatisfaction in their work.

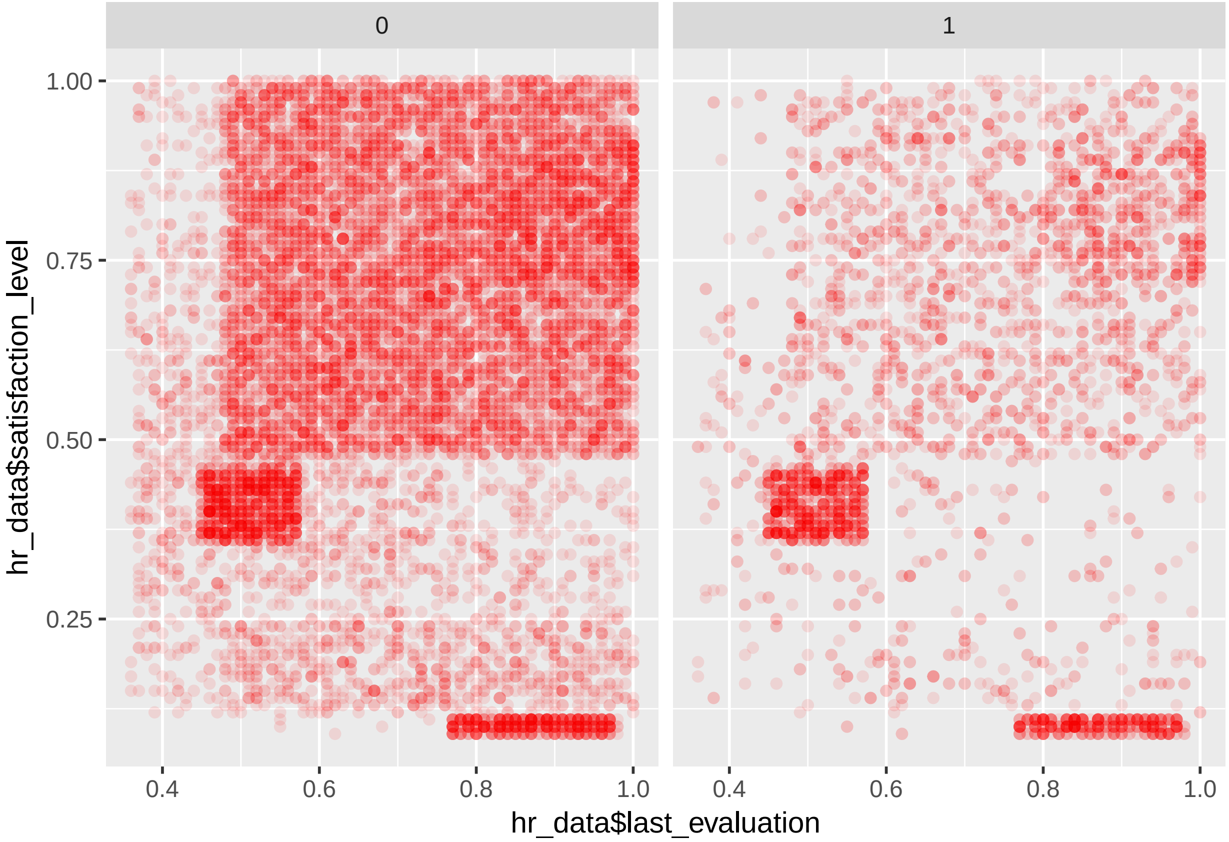
```r
#Evaluation
ggplot(data=hr_data, aes(hr_data$last_evaluation))+
  geom_density(aes(group= factor(hr_data$left),fill=factor(hr_data$left)))+
  xlab("Left")+
  ylab("Last Evaluation")
```



```r
by(hr_data$last_evaluation, hr_data$left, summary)
```

```
## hr_data$left: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3600  0.5800  0.7100  0.7155  0.8500  1.0000
## --------------------------------------------------------
## hr_data$left: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4500  0.5200  0.7900  0.7181  0.9000  1.0000
```

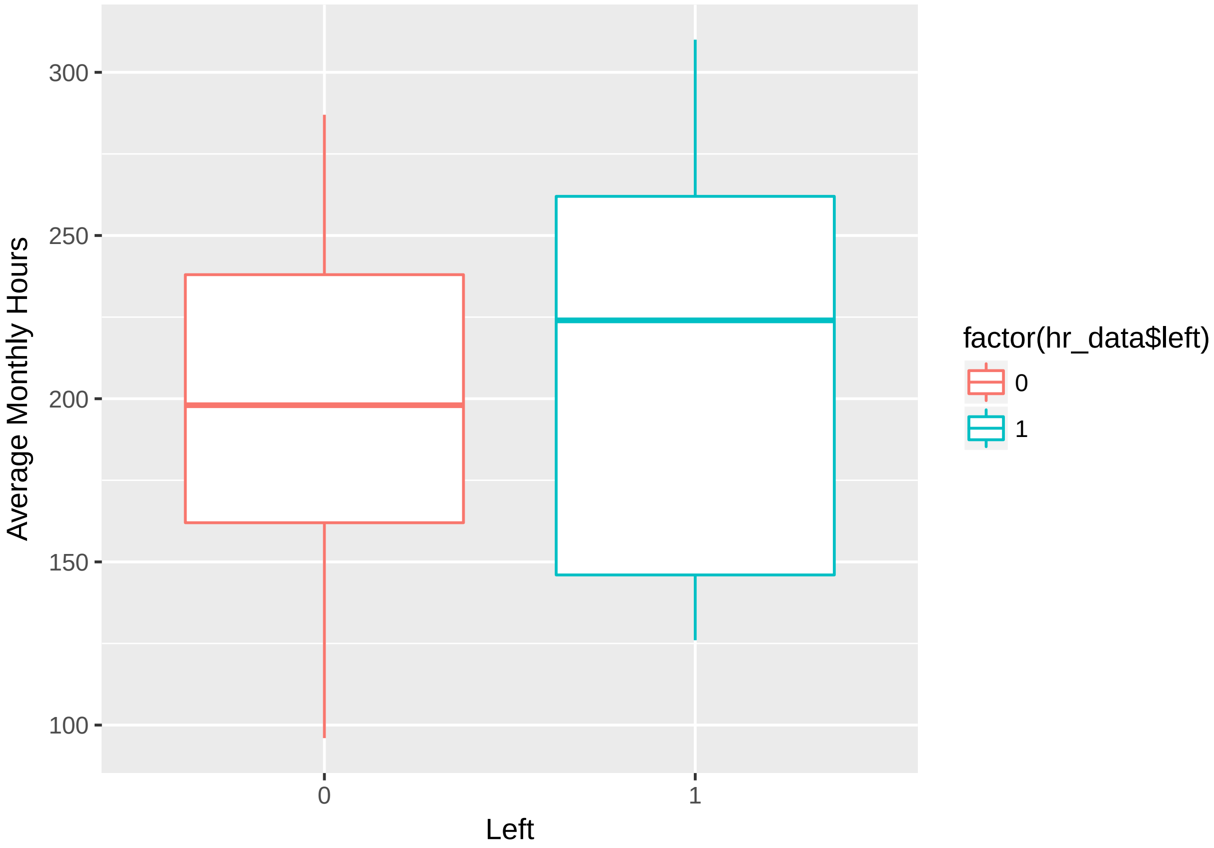
```r
#Relationship between satisfaction levels and last_evaluation.
ggplot(aes(hr_data$last_evaluation, hr_data$satisfaction_level), data=hr_data)+
  geom_point(alpha=1/10, col="red")+
  facet_wrap(~hr_data$left)
```



We can see two peaks of evaluation scores for people who left and this indicates that most people who left are extremely high or extremely low performers.

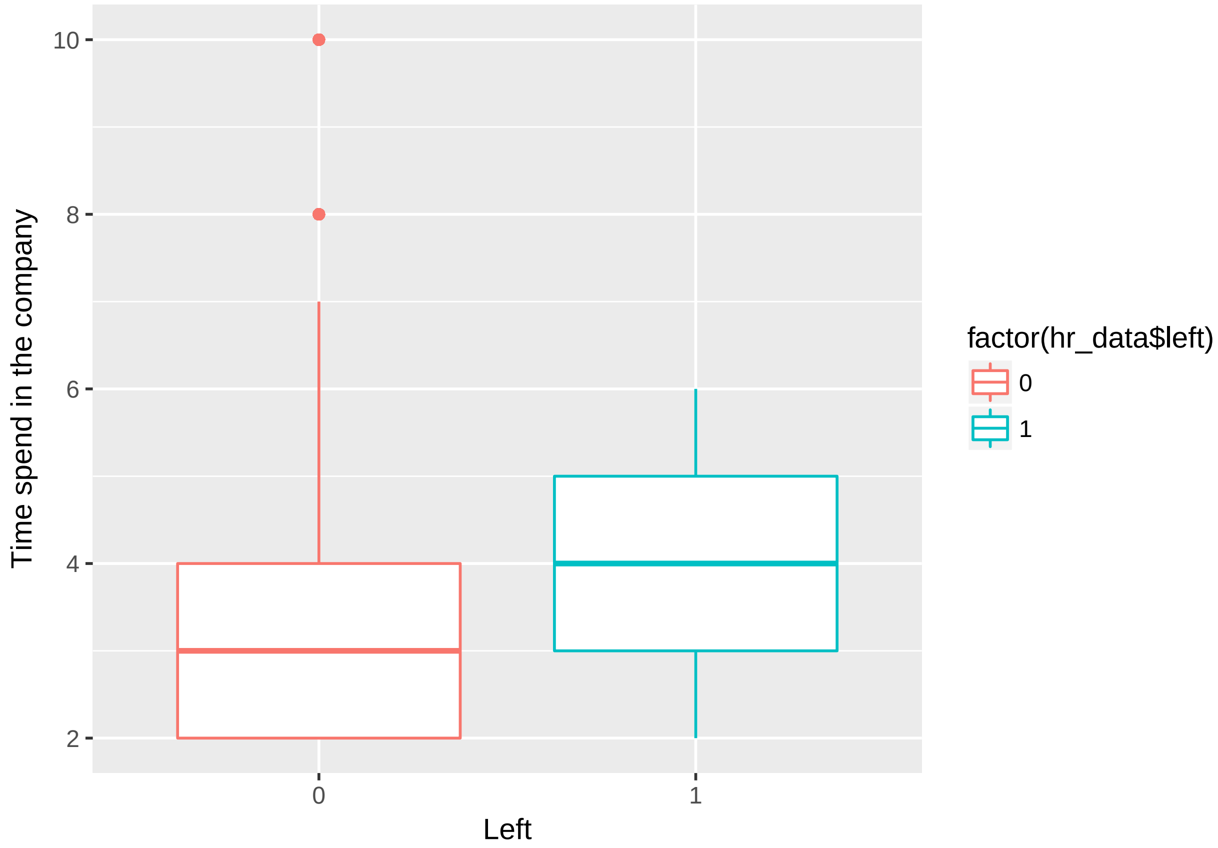
The plot for satisfaction levels and last evaluation is tells us that these both factors might be related. For the employees that left the company, satisfaction levels are lesser as compared to the ones staying back. We can see two distinct patterns for the employees who left the company, one where the evaluation is very high (high performers), but the satisfaction level is very less. Other where the satisfaction and evaluation are on the lower side.

```r
#Average_monthly_hours
ggplot(data=hr_data, aes(x=factor(hr_data$left),y=hr_data$average_montly_hours))+
  geom_boxplot(aes(color=factor(hr_data$left)))+
  xlab("Left")+
  ylab("Average Monthly Hours")
```



Average monthly hours of people who left is higher than that of people who stayed.

```
#Time spend in the company
ggplot(data=hr_data, aes(x=factor(hr_data$left),y=hr_data$time_spend_company))+
  geom_boxplot(aes(color=factor(hr_data$left)))+
  xlab("Left")+
  ylab("Time spend in the company")
```



People who left the company have a much higher tenure as compared to the ones who stayed.

```
#Salary
table(hr_data$salary)
```

```
##
##   high    low medium
##   1237   7316   6446
```

```
by(hr_data$salary, hr_data$left, table)
```

```
## hr_data$left: 0
##
##   high    low medium
##   1155   5144   5129
## --------------------------------------------------------
## hr_data$left: 1
##
##   high    low medium
##     82   2172   1317
```

6.6% of people from higher salary range left, 29.68% from low salary range left, 20.4% from medium salary range left. Thus, its clear that people from lower salary range tend to leave the company.

```
#Number of projects
by(hr_data$number_project,hr_data$left,table)
```

```
## hr_data$left: 0
##
##    2    3    4    5    6
##  821 3983 3956 2149  519
## ---------------------------------------------------------
## hr_data$left: 1
## 
##    2    3    4    5    6    7 
## 1567   72  409  612  655  256
```

Maximum number of people who did not leave, seem to work on 3 or 4 projects in the comapny.Maximum number of people who left seem to have worked in 2 projects or higher numbers like 6 or 7 in the comapny.

```
#Promotion in last 5 years
table(hr_data$promotion_last_5years)
```

```
## 
##     0     1 
## 14680   319
```

```
by(hr_data$promotion_last_5years,hr_data$left, table)
```

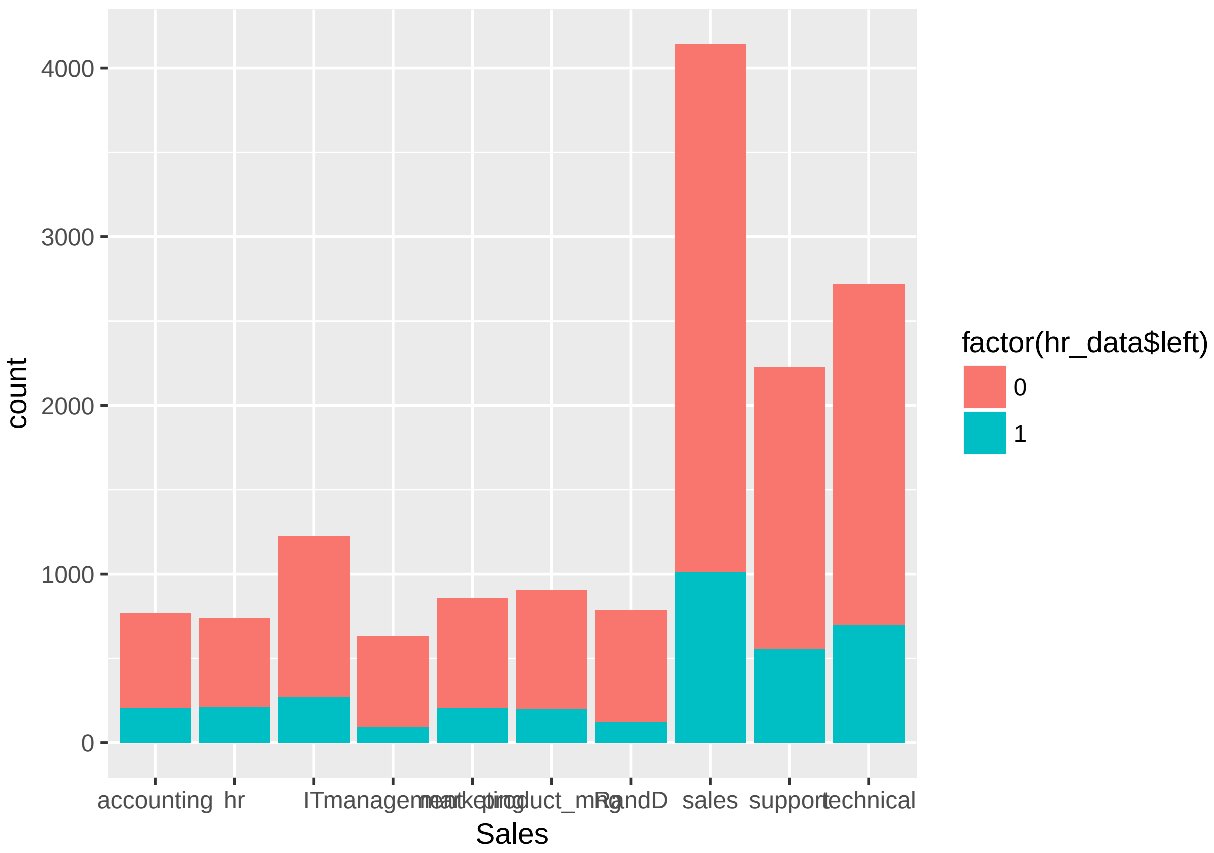
```
## hr_data$left: 0
## 
##     0     1 
## 11128   300 
## ---------------------------------------------------------
## hr_data$left: 1
## 
##    0    1 
## 3552   19
```

Only 2.2% of the people in the company were promoted in the last 5 years. 2.7% of people who stayed got the promotion, whereas only 0.5% of people who left had got a promotion.

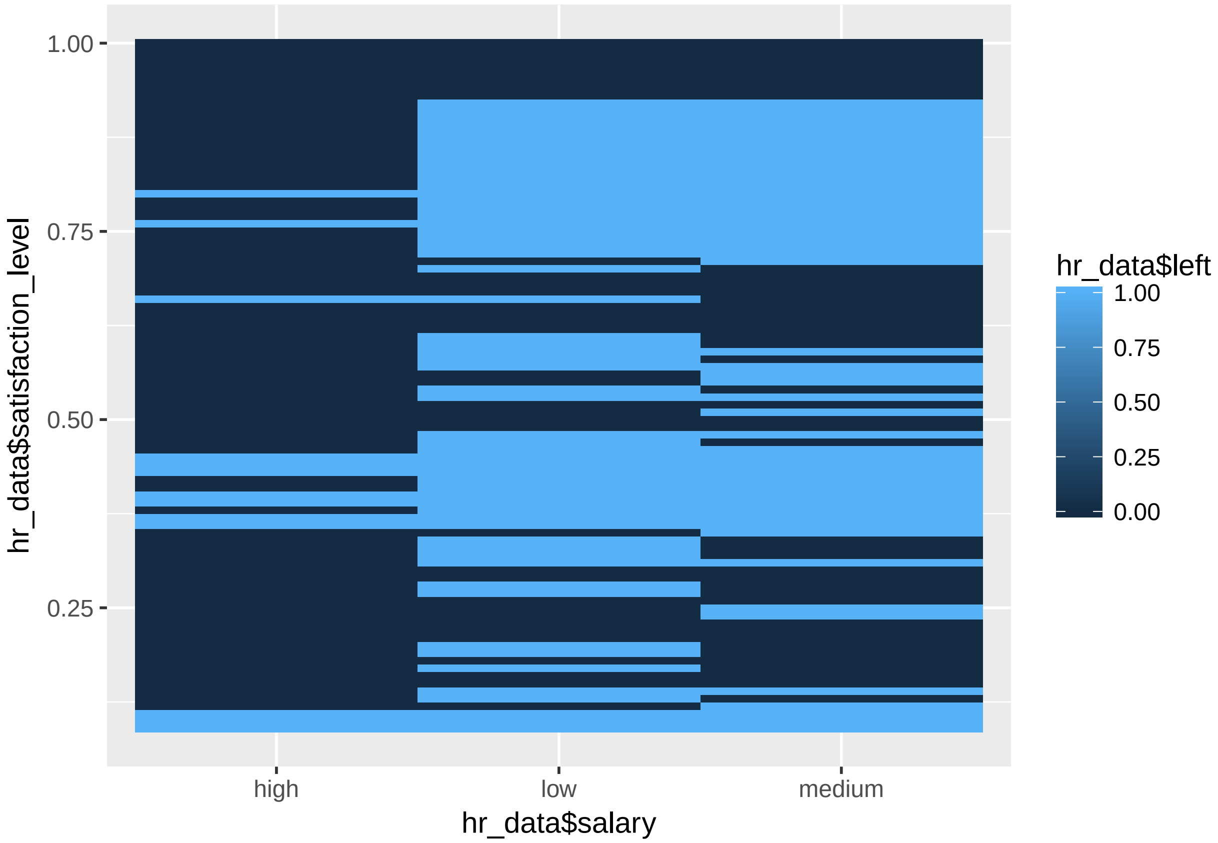
```
#Sales
x<- table(hr_data$sales, hr_data$left)
by(hr_data$sales, hr_data$left, table)
```

```
## hr_data$left: 0
## 
##  accounting          hr          IT  management   marketing product_mng 
##         563         524         954         539         655         704 
##       RandD       sales     support   technical 
##         666        3126        1674        2023 
## ---------------------------------------------------------
## hr_data$left: 1
## 
##  accounting          hr          IT  management   marketing product_mng 
##         204         215         273          91         203         198 
##       RandD       sales     support   technical 
##         121        1014         555         697
```

```
ggplot(aes(hr_data$sales), data=hr_data)+
  geom_bar(aes(fill=factor(hr_data$left)))+
  xlab("Sales")
```



```
#Satisfaction level vs salary
ggplot(aes(hr_data$salary,hr_data$satisfaction_level), data=hr_data)+
  geom_raster(aes(fill=hr_data$left))
```



Important observations/Insights:

People who left the company seem to be less satisfied as compared to the ones staying back. Higher working hours might be one of the reasons for the people to leave the company. People who left the company seem to have higher tenure. This may imply that they are looking for better opportunities or looking for a change in job. People having low salaries seem to have left the company in large numbers, this may be due to their dissatisfaction due to lower salaries or higher opportunities in the market for lower levels. People who left seem to have extremely high or low performance evaluation.This may mean that they are not happy in the job and are leaving or they are overqualified and are looking for better opportunities. Promotion might be an important factor in a person's decision to leave or stay back.

Let us find the bivariate relationship present in the data. First lets find the correlation between the output variable i.e left and all other variables.

```r
#Correlations are performed on numeric values and hence converting sales and salary to numeric value.
hr_data$sales <- as.numeric(hr_data$sales)
hr_data$salary <- as.numeric(hr_data$salary)
x <- cor(x=hr_data[,1:10], y= hr_data[,1:10])
```

We find the correlation between all the variables to examine the relationship between the variables themselves.Correlation shows how strongly two variables are related. A positive correlation shows that as 1 variable increases the other increases too, while a negative correlation shows that a one variable decreases the other decreases too.

Satisfaction level is the strongest correlated variable with left. Performance is correlated with average monthly hours and number of projects. Number of projects is correlated with average monthly hours.

Relationship between employees leaving and other factors

```
#Obtaining the train and test dataset
sample <- floor(0.7*nrow(hr_data))
set.seed(100)
hr_indices <- sample(seq_len(nrow(hr_data)), size=sample)

#Load the train and test data
hr_train <- hr_data[hr_indices,]
hr_test <- hr_data[-hr_indices,]

#Fitting a Binomial Logistic regression model for leaving the company
model <- glm(hr_data$left ~., family = binomial(link="logit"), data=hr_data)
summary(model)
```

```
##
## Call:
## glm(formula = hr_data$left ~ ., family = binomial(link = "logit"),
##     data = hr_data)
##
## 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_accident         -1.498575   0.088254 -16.980  < 2e-16 ***
## promotion_last_5years -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
```

The p value for all the variables are statistically significant. Satisfaction level, Number of projects, work accident, promotion and sales(considering all the coefficients for sales), these varaibles have a negative relationship with a person leaving the company.

Prediction

```
hr_predict <- predict(model,type = "response", hr_test)
hr_predict <-ifelse(hr_predict > 0.5,1,0)

Error <-mean(hr_predict != hr_test$left)
print(paste('Accuracy', 1-Error))
```

```
## [1] "Accuracy 0.769555555555556"
```

After performing out of sample validation using the test data, we get the the accuracy of this model to be 0.77 which is high. Thus, we can say that this model is a good fit to our data.

Performance of the logistic regression model

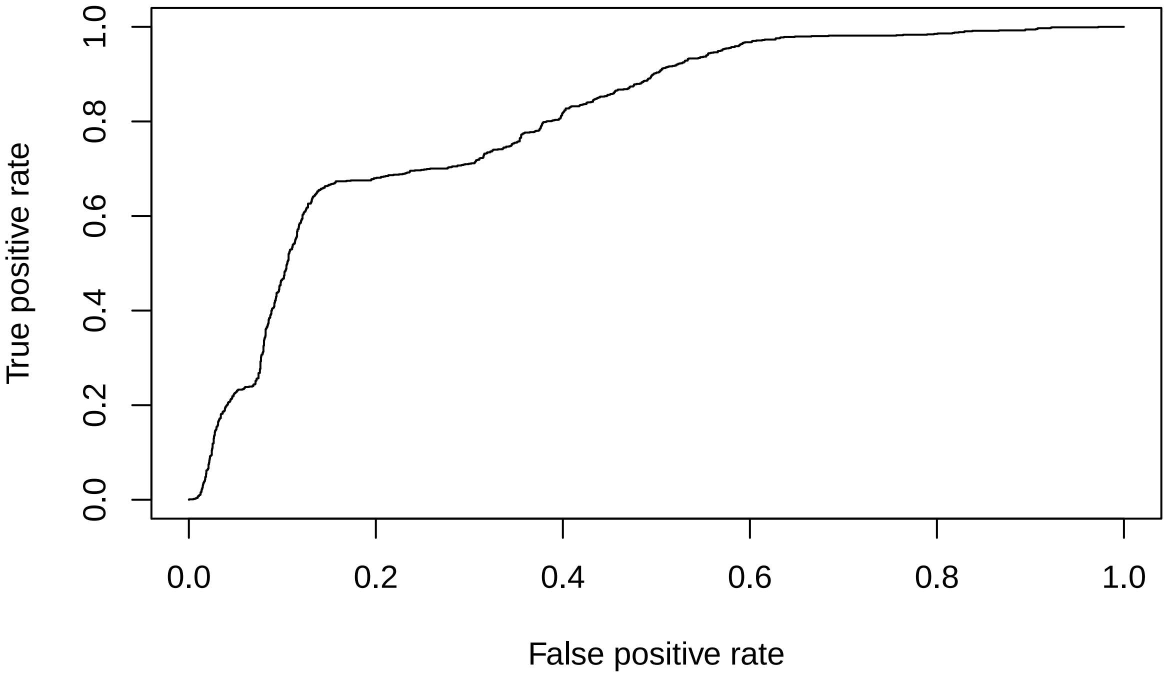
```
#install.packages("ROCR") Receiver operating characteristics.
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
hr_predict1 <- predict(model,type = "response", hr_test)
pr <- prediction(hr_predict1, hr_test$left)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```



```
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.8045439
```

We plot an ROC curve to get the Area under the curve(AUC), which is an indication of how well the model performs. Thue AUC comes out to be 0.8. Thus this tells us that there is scope of improvement to this model.

We try to model this data on a random forest algorithm, to compare it with the logistic regression model and see if this model has a better fit as compared to the previous.

Decision Trees

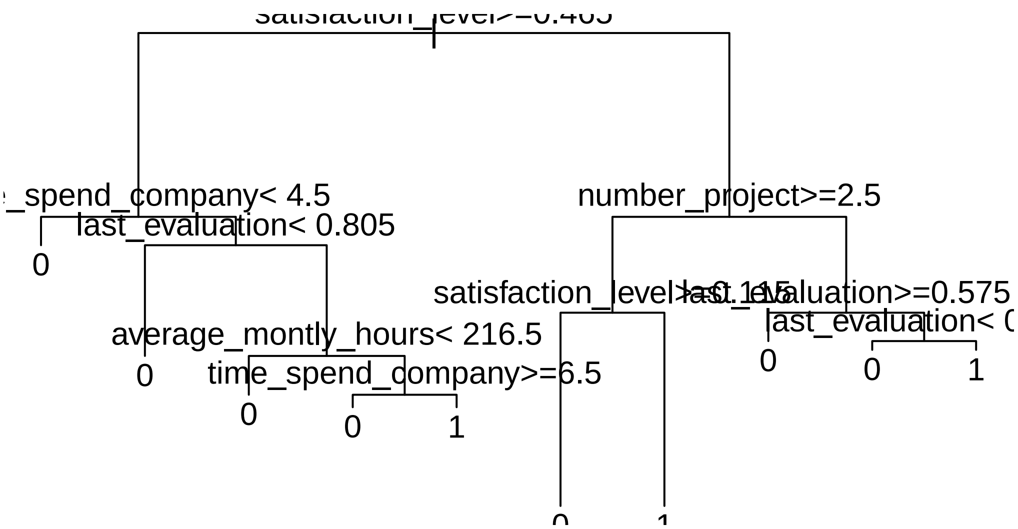
```
library(rpart)
tree_model <- rpart(hr_train$left~., hr_train, method = "class")
summary(tree_model)
```

```
## Call:
## rpart(formula = hr_train$left ~ ., data = hr_train, method = "class")
##   n= 10499
##
##           CP nsplit rel error    xerror        xstd
## 1 0.24308063      0 1.0000000 1.0000000 0.017489310
## 2 0.19113518      1 0.7569194 0.7569194 0.015781261
## 3 0.07320497      3 0.3746490 0.3746490 0.011700911
## 4 0.05134376      5 0.2282391 0.2290413 0.009320789
## 5 0.03329322      6 0.1768953 0.1784998 0.008280436
## 6 0.01644605      7 0.1436021 0.1452066 0.007499162
## 7 0.01163257      8 0.1271560 0.1307661 0.007129139
## 8 0.01000000      9 0.1155235 0.1219414 0.006891824
##
## Variable importance
##   satisfaction_level       number_project average_montly_hours
##                   35                   17                   17
##      last_evaluation   time_spend_company
##                   17                   13
##
## Node number 1: 10499 observations,    complexity param=0.2430806
##   predicted class=0  expected loss=0.2374512  P(node) =1
##     class counts:  8006  2493
##    probabilities: 0.763 0.237
##   left son=2 (7533 obs) right son=3 (2966 obs)
##   Primary splits:
##       satisfaction_level   < 0.465 to the right, improve=1099.6850, (0 missing)
##       number_project       < 2.5   to the right, improve= 690.2286, (0 missing)
##       time_spend_company   < 2.5   to the left,  improve= 273.6989, (0 missing)
##       average_montly_hours < 274.5 to the left,  improve= 269.6209, (0 missing)
##       last_evaluation      < 0.575 to the right, improve= 154.5759, (0 missing)
##   Surrogate splits:
##       number_project       < 2.5   to the right, agree=0.790, adj=0.258, (0 split)
##       average_montly_hours < 275.5 to the left,  agree=0.750, adj=0.114, (0 split)
##       last_evaluation      < 0.485 to the right, agree=0.735, adj=0.063, (0 split)
```

```
##
## Node number 2: 7533 observations,    complexity param=0.07320497
##   predicted class=0  expected loss=0.09385371  P(node) =0.7174969
##     class counts:  6826   707
##    probabilities: 0.906 0.094
##   left son=4 (6153 obs) right son=5 (1380 obs)
##   Primary splits:
##       time_spend_company   < 4.5   to the left,  improve=432.09020, (0 missing)
##       last_evaluation      < 0.825 to the left,  improve=148.94940, (0 missing)
##       average_montly_hours < 216.5 to the left,  improve=115.43300, (0 missing)
##       number_project       < 4.5   to the left,  improve= 72.86789, (0 missing)
##       satisfaction_level   < 0.715 to the left,  improve= 57.57935, (0 missing)
##   Surrogate splits:
##       last_evaluation      < 0.995 to the left,  agree=0.823, adj=0.033, (0 split)
##       average_montly_hours < 285.5 to the left,  agree=0.817, adj=0.002, (0 split)
##
## Node number 3: 2966 observations,    complexity param=0.1911352
##   predicted class=1  expected loss=0.3978422  P(node) =0.2825031
##     class counts:  1180  1786
##    probabilities: 0.398 0.602
##   left son=6 (1742 obs) right son=7 (1224 obs)
##   Primary splits:
##       number_project       < 2.5   to the right, improve=314.0102, (0 missing)
##       satisfaction_level   < 0.115 to the right, improve=258.3409, (0 missing)
##       time_spend_company   < 4.5   to the right, improve=248.3809, (0 missing)
##       last_evaluation      < 0.575 to the right, improve=129.7737, (0 missing)
##       average_montly_hours < 160.5 to the right, improve=114.7603, (0 missing)
##   Surrogate splits:
##       satisfaction_level   < 0.355 to the left,  agree=0.887, adj=0.726, (0 split)
##       last_evaluation      < 0.575 to the right, agree=0.861, adj=0.663, (0 split)
##       average_montly_hours < 161.5 to the right, agree=0.856, adj=0.651, (0 split)
##       time_spend_company   < 3.5   to the right, agree=0.844, adj=0.622, (0 split)
##
## Node number 4: 6153 observations
##   predicted class=0  expected loss=0.01365188  P(node) =0.5860558
##     class counts:  6069    84
##    probabilities: 0.986 0.014
##
## Node number 5: 1380 observations,    complexity param=0.07320497
##   predicted class=0  expected loss=0.4514493  P(node) =0.1314411
##     class counts:   757   623
##    probabilities: 0.549 0.451
##   left son=10 (537 obs) right son=11 (843 obs)
##   Primary splits:
##       last_evaluation      < 0.805 to the left,  improve=304.3571, (0 missing)
##       average_montly_hours < 216.5 to the left,  improve=259.4762, (0 missing)
##       time_spend_company   < 6.5   to the right, improve=171.1974, (0 missing)
##       satisfaction_level   < 0.715 to the left,  improve=169.7424, (0 missing)
##       number_project       < 3.5   to the left,  improve=134.0749, (0 missing)
##   Surrogate splits:
##       average_montly_hours < 215.5 to the left,  agree=0.752, adj=0.363, (0 split)
##       number_project       < 3.5   to the left,  agree=0.712, adj=0.259, (0 split)
##       satisfaction_level   < 0.705 to the left,  agree=0.704, adj=0.240, (0 split)
##       time_spend_company   < 6.5   to the right, agree=0.678, adj=0.171, (0 split)
##       Work_accident        < 0.5   to the right, agree=0.649, adj=0.097, (0 split)
## 
## Node number 6: 1742 observations,    complexity param=0.1911352
##   predicted class=0  expected loss=0.4092997  P(node) =0.1659206
##     class counts:  1029   713
##    probabilities: 0.591 0.409 
##   left son=12 (1105 obs) right son=13 (637 obs)
##   Primary splits:
##       satisfaction_level   < 0.115 to the right, improve=700.7930, (0 missing)
##       average_montly_hours < 242.5 to the left,  improve=402.1237, (0 missing)
##       number_project       < 5.5   to the left,  improve=373.7189, (0 missing)
##       last_evaluation      < 0.765 to the left,  improve=280.5368, (0 missing)
##       time_spend_company   < 3.5   to the left,  improve=114.1181, (0 missing)
##   Surrogate splits:
##       average_montly_hours < 242.5 to the left,  agree=0.862, adj=0.622, (0 split)
##       number_project       < 5.5   to the left,  agree=0.839, adj=0.560, (0 split)
##       last_evaluation      < 0.765 to the left,  agree=0.777, adj=0.389, (0 split)
## 
## Node number 7: 1224 observations,    complexity param=0.03329322
##   predicted class=1  expected loss=0.123366  P(node) =0.1165825
##     class counts:   151  1073
##    probabilities: 0.123 0.877 
##   left son=14 (95 obs) right son=15 (1129 obs)
##   Primary splits:
##       last_evaluation      < 0.575 to the right, improve=136.31090, (0 missing)
##       average_montly_hours < 162   to the right, improve=120.56250, (0 missing)
##       satisfaction_level   < 0.355 to the left,  improve=111.78400, (0 missing)
##       time_spend_company   < 3.5   to the right, improve= 63.60429, (0 missing)
##       salary               < 1.5   to the left,  improve=  8.24520, (0 missing)
##   Surrogate splits:
##       average_montly_hours < 172.5 to the right, agree=0.946, adj=0.305, (0 split)
##       time_spend_company   < 3.5   to the right, agree=0.939, adj=0.211, (0 split)
##       satisfaction_level   < 0.355 to the left,  agree=0.936, adj=0.179, (0 split)
## 
## Node number 10: 537 observations
##   predicted class=0  expected loss=0.03538175  P(node) =0.05114773
##     class counts:   518    19
##    probabilities: 0.965 0.035 
## 
## Node number 11: 843 observations,    complexity param=0.05134376
##   predicted class=1  expected loss=0.2835113  P(node) =0.08029336
##     class counts:   239   604
##    probabilities: 0.284 0.716 
##   left son=22 (160 obs) right son=23 (683 obs)
##   Primary splits:
##       average_montly_hours < 216.5 to the left,  improve=150.10910, (0 missing)
##       time_spend_company   < 6.5   to the right, improve=136.72340, (0 missing)
##       satisfaction_level   < 0.715 to the left,  improve=116.17070, (0 missing)
##       number_project       < 3.5   to the left,  improve= 72.46999, (0 missing)
##       salary               < 1.5   to the left,  improve= 20.49752, (0 missing)
##   Surrogate splits:
##       time_spend_company < 6.5   to the right, agree=0.849, adj=0.206, (0 split)
##       satisfaction_level < 0.715 to the left,  agree=0.836, adj=0.138, (0 split)
##       number_project     < 3.5   to the left,  agree=0.827, adj=0.088, (0 split)
```

```
##
## Node number 12: 1105 observations
##   predicted class=0  expected loss=0.06877828  P(node) =0.1052481
##     class counts:  1029    76
##    probabilities: 0.931 0.069
##
## Node number 13: 637 observations
##   predicted class=1  expected loss=0  P(node) =0.06067244
##     class counts:     0   637
##    probabilities: 0.000 1.000
##
## Node number 14: 95 observations
##   predicted class=0  expected loss=0.06315789  P(node) =0.009048481
##     class counts:    89     6
##    probabilities: 0.937 0.063
##
## Node number 15: 1129 observations,    complexity param=0.01163257
##   predicted class=1  expected loss=0.05491585  P(node) =0.1075341
##     class counts:    62  1067
##    probabilities: 0.055 0.945
##   left son=30 (29 obs) right son=31 (1100 obs)
##   Primary splits:
##       last_evaluation      < 0.445 to the left,  improve=53.170430, (0 missing)
##       average_montly_hours < 163.5 to the right, improve=42.454020, (0 missing)
##       satisfaction_level   < 0.355 to the left,  improve=39.607650, (0 missing)
##       time_spend_company   < 2.5   to the left,  improve=23.493300, (0 missing)
##       salary               < 1.5   to the left,  improve= 1.089545, (0 missing)
##   Surrogate splits:
##       average_montly_hours < 191   to the right, agree=0.976, adj=0.069, (0 split)
##       satisfaction_level   < 0.22  to the left,  agree=0.975, adj=0.034, (0 split)
##
## Node number 22: 160 observations
##   predicted class=0  expected loss=0.1  P(node) =0.01523955
##     class counts:   144    16
##    probabilities: 0.900 0.100
##
## Node number 23: 683 observations,    complexity param=0.01644605
##   predicted class=1  expected loss=0.1390922  P(node) =0.06505381
##     class counts:    95   588
##    probabilities: 0.139 0.861
##   left son=46 (41 obs) right son=47 (642 obs)
##   Primary splits:
##       time_spend_company    < 6.5   to the right, improve=64.65659, (0 missing)
##       satisfaction_level    < 0.715 to the left,  improve=60.08955, (0 missing)
##       number_project        < 3.5   to the left,  improve=52.63534, (0 missing)
##       salary                < 1.5   to the left,  improve=16.12015, (0 missing)
##       promotion_last_5years < 0.5   to the right, improve=13.25727, (0 missing)
##   Surrogate splits:
##       promotion_last_5years < 0.5   to the right, agree=0.950, adj=0.171, (0 split)
##       satisfaction_level    < 0.59  to the left,  agree=0.944, adj=0.073, (0 split)
##
## Node number 30: 29 observations
##   predicted class=0  expected loss=0  P(node) =0.002762168
##     class counts:    29     0
##    probabilities: 1.000 0.000
##
## Node number 31: 1100 observations
##   predicted class=1  expected loss=0.03  P(node) =0.1047719
##     class counts:    33  1067
##    probabilities: 0.030 0.970
##
## Node number 46: 41 observations
##   predicted class=0  expected loss=0  P(node) =0.003905134
##     class counts:    41     0
##    probabilities: 1.000 0.000
##
## Node number 47: 642 observations
##   predicted class=1  expected loss=0.08411215  P(node) =0.06114868
##     class counts:    54   588
##    probabilities: 0.084 0.916
```

```
plot(tree_model)
text(tree_model)
```



```
tree_predict <- predict(tree_model, hr_test, type="class")


#Performance testing
#install.packages("caret")
#install.packages("e1071", dependencies = TRUE)
library(caret)
```

```
## Loading required package: lattice
```

```
confusionMatrix(tree_predict, hr_test$left)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 3379  101
##          1   43  977
##
```

```
##                Accuracy : 0.968
##                  95% CI : (0.9624, 0.9729)
##     No Information Rate : 0.7604
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9105
##  Mcnemar's Test P-Value : 2.034e-06
##
##             Sensitivity : 0.9874
##             Specificity : 0.9063
##          Pos Pred Value : 0.9710
##          Neg Pred Value : 0.9578
##              Prevalence : 0.7604
##          Detection Rate : 0.7509
##    Detection Prevalence : 0.7733
##       Balanced Accuracy : 0.9469
##
##        'Positive' Class : 0
##
```

The above data was modeled using decision trees with the help of th rpart package. The performance is tested using confusion matrix which gives a tabular summary of the actual test data labels vs the predicted labels. The confusion matrix gives an acciracy of 96.8%. Sensitivity which represents the true positive rate is 98.74%. i.e this is the percentage of times the model predicted that the an employee will leave the company and the employee actually left. Specificity which represents the true negative rate is 90.63% i.e this is the percentage of times the model predicted that an employee will not leave the company and the employee actually did not.

Random forest

```
library(randomForest)
```

```
## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
hr_rf <- randomForest(as.factor(hr_train$left)~.,hr_train, importance=TRUE, ntree=1000,method='class')

pred <- predict(hr_rf,hr_test)

confusionMatrix(pred, hr_test$left)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 3415   34
##          1    7 1044
##
##                Accuracy : 0.9909
##                  95% CI : (0.9877, 0.9935)
##     No Information Rate : 0.7604
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9748
##  Mcnemar's Test P-Value : 4.896e-05
##
##             Sensitivity : 0.9980
##             Specificity : 0.9685
##          Pos Pred Value : 0.9901
##          Neg Pred Value : 0.9933
##              Prevalence : 0.7604
##          Detection Rate : 0.7589
##    Detection Prevalence : 0.7664
##       Balanced Accuracy : 0.9832
##
##        'Positive' Class : 0
##
```

As we can see the random forest mode gives an accuracy of 99.07%, which is very higher than that given by the decision trees. This model fits our data much better than the logistic regression model and decision trees.

Extensive Logitic Regression:

```r
# We start the model with a single explanatory variable
var1 <- glm(hr_data$left~ hr_data$satisfaction_level, data=hr_data, family = binomial())
summary(var1)
```

```
##
## Call:
## glm(formula = hr_data$left ~ hr_data$satisfaction_level, family = binomial(),
##     data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4020  -0.6982  -0.5002  -0.3402   2.2922
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.97388    0.04935   19.73   <2e-16 ***
## hr_data$satisfaction_level -3.83248    0.08720  -43.95   <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: 14198  on 14997  degrees of freedom
## AIC: 14202
##
## Number of Fisher Scoring iterations: 4
```

```
# 2nd variable
var2 <- glm(hr_data$left ~ hr_data$satisfaction_level+hr_data$last_evaluation, data=hr_data, family = bi
summary(var2)
```

```
##
## Call:
## glm(formula = hr_data$left ~ hr_data$satisfaction_level + hr_data$last_evaluation,
##     family = binomial(), data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4619  -0.7050  -0.5015  -0.3359   2.2949
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.62697    0.09567   6.554 5.61e-11 ***
## hr_data$satisfaction_level -3.85391    0.08752 -44.034  < 2e-16 ***
## hr_data$last_evaluation     0.50871    0.12034   4.227 2.37e-05 ***
## ---
## 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: 14180  on 14996  degrees of freedom
## AIC: 14186
##
## Number of Fisher Scoring iterations: 4
```

```
var3 <- glm(hr_data$left ~ hr_data$satisfaction_level+hr_data$last_evaluation+ hr_data$number_project,
summary(var3)
```

```
##
## Call:
## glm(formula = hr_data$left ~ hr_data$satisfaction_level + hr_data$last_evaluation +
##     hr_data$number_project, family = binomial(), data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7031  -0.7059  -0.4837  -0.2859   2.4182
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.03173    0.10239  10.077   <2e-16 ***
## hr_data$satisfaction_level -4.16950    0.09429 -44.219   <2e-16 ***
## hr_data$last_evaluation     1.18345    0.13699   8.639   <2e-16 ***
## hr_data$number_project     -0.19176    0.01804 -10.631   <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: 14065  on 14995  degrees of freedom
## AIC: 14073
##
## Number of Fisher Scoring iterations: 4
```

```
var4 <- glm(hr_data$left ~ hr_data$satisfaction_level+hr_data$last_evaluation+ hr_data$number_project+ h
summary(var4)
```

```
##
## Call:
## glm(formula = hr_data$left ~ hr_data$satisfaction_level + hr_data$last_evaluation +
##     hr_data$number_project + hr_data$average_montly_hours, family = binomial(),
##     data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8019  -0.7040  -0.4820  -0.2669   2.5101
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  0.6452398  0.1112376   5.801 6.61e-09 ***
## hr_data$satisfaction_level   -4.1961067  0.0949322 -44.201  < 2e-16 ***
## hr_data$last_evaluation       0.8786325  0.1412950   6.218 5.02e-10 ***
## hr_data$number_project       -0.2646078  0.0200116 -13.223  < 2e-16 ***
## hr_data$average_montly_hours  0.0044340  0.0004884   9.079  < 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: 13981  on 14994  degrees of freedom
## AIC: 13991
##
## Number of Fisher Scoring iterations: 5
```

```
var5 <- glm(hr_data$left ~ hr_data$satisfaction_level+hr_data$last_evaluation+ hr_data$number_project+ h
summary(var5)
```

```
##
## Call:
## glm(formula = hr_data$left ~ hr_data$satisfaction_level + hr_data$last_evaluation +
##     hr_data$number_project + hr_data$average_montly_hours + hr_data$time_spend_company,
##     family = binomial(), data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1997  -0.6872  -0.4649  -0.2484   2.5728
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  0.1419922  0.1161974   1.222    0.222
## hr_data$satisfaction_level   -4.1345085  0.0951351 -43.459  < 2e-16 ***
## hr_data$last_evaluation       0.7621197  0.1426846   5.341 9.23e-08 ***
## hr_data$number_project       -0.3025850  0.0204281 -14.812  < 2e-16 ***
## hr_data$average_montly_hours  0.0043586  0.0004929   8.842  < 2e-16 ***
## hr_data$time_spend_company    0.1971188  0.0141593  13.922  < 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: 13794  on 14993  degrees of freedom
## AIC: 13806
##
## Number of Fisher Scoring iterations: 5
```

```
#
var6 <- glm(hr_data$left ~ hr_data$satisfaction_level+hr_data$last_evaluation+ hr_data$number_project+
summary(var6)
```

```
##
## Call:
## glm(formula = hr_data$left ~ hr_data$satisfaction_level + hr_data$last_evaluation +
##     hr_data$number_project + hr_data$average_montly_hours + hr_data$time_spend_company +
##     hr_data$Work_accident, family = binomial(), data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3008  -0.6839  -0.4391  -0.1619   2.9760
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.2327351  0.1168396   1.992   0.0464 *
## hr_data$satisfaction_level   -4.1332297  0.0963863 -42.882  < 2e-16 ***
## hr_data$last_evaluation       0.7849940  0.1453857   5.399 6.69e-08 ***
## hr_data$number_project       -0.3058886  0.0207663 -14.730  < 2e-16 ***
## hr_data$average_montly_hours  0.0043530  0.0005023   8.666  < 2e-16 ***
## hr_data$time_spend_company    0.2119469  0.0146232  14.494  < 2e-16 ***
## hr_data$Work_accident        -1.5063657  0.0879336 -17.131  < 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: 13403  on 14992  degrees of freedom
## AIC: 13417
##
## Number of Fisher Scoring iterations: 5
```

```
var7 <- glm(hr_data$left ~ hr_data$satisfaction_level+hr_data$last_evaluation+ hr_data$number_project+
summary(var7)
```

```
##
## Call:
## glm(formula = hr_data$left ~ hr_data$satisfaction_level + hr_data$last_evaluation +
##     hr_data$number_project + hr_data$average_montly_hours + hr_data$time_spend_company +
##     hr_data$Work_accident + hr_data$promotion_last_5years, family = binomial(),
##     data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3478  -0.6812  -0.4343  -0.1518   3.1237
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    0.2240079  0.1169674   1.915   0.0555 .
## hr_data$satisfaction_level    -4.1233688  0.0964963 -42.731  < 2e-16 ***
## hr_data$last_evaluation        0.7626360  0.1456844   5.235 1.65e-07 ***
## hr_data$number_project        -0.3085030  0.0208339 -14.808  < 2e-16 ***
## hr_data$average_montly_hours   0.0043376  0.0005037   8.611  < 2e-16 ***
## hr_data$time_spend_company     0.2268197  0.0148291  15.296  < 2e-16 ***
## hr_data$Work_accident         -1.4951671  0.0882135 -16.949  < 2e-16 ***
## hr_data$promotion_last_5years -1.7944627  0.2557227  -7.017 2.26e-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: 13330  on 14991  degrees of freedom
## AIC: 13346
##
## Number of Fisher Scoring iterations: 5
```

```
var8 <- glm(hr_data$left ~ hr_data$satisfaction_level+hr_data$last_evaluation+ hr_data$number_project+
summary(var8)
```

```
##
## Call:
## glm(formula = hr_data$left ~ hr_data$satisfaction_level + hr_data$last_evaluation +
##     hr_data$number_project + hr_data$average_montly_hours + hr_data$time_spend_company +
##     hr_data$Work_accident + hr_data$promotion_last_5years + hr_data$sales,
##     family = binomial(), data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3630  -0.6823  -0.4345  -0.1526   3.1097
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    0.0815369  0.1290068   0.632  0.52736
## hr_data$satisfaction_level    -4.1287921  0.0965692 -42.755  < 2e-16 ***
## hr_data$last_evaluation        0.7624413  0.1457099   5.233 1.67e-07 ***
## hr_data$number_project        -0.3099587  0.0208455 -14.869  < 2e-16 ***
## hr_data$average_montly_hours   0.0043453  0.0005039   8.623  < 2e-16 ***
## hr_data$time_spend_company     0.2286246  0.0148556  15.390  < 2e-16 ***
## hr_data$Work_accident         -1.4987312  0.0882561 -16.982  < 2e-16 ***
## hr_data$promotion_last_5years -1.7694762  0.2555546  -6.924 4.39e-12 ***
## hr_data$sales                  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
```

```
var8 <- glm(hr_data$left ~ hr_data$satisfaction_level+hr_data$last_evaluation+ hr_data$number_project+
summary(var8)
```

```
##
## Call:
## glm(formula = hr_data$left ~ hr_data$satisfaction_level + hr_data$last_evaluation +
##     hr_data$number_project + hr_data$average_montly_hours + hr_data$time_spend_company +
##     hr_data$Work_accident + hr_data$promotion_last_5years + hr_data$sales +
##     hr_data$salary, family = binomial(), data = hr_data)
##
## 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
## hr_data$satisfaction_level    -4.129254   0.096584 -42.753  < 2e-16 ***
## hr_data$last_evaluation        0.762165   0.145708   5.231 1.69e-07 ***
## hr_data$number_project        -0.310068   0.020850 -14.872  < 2e-16 ***
## hr_data$average_montly_hours   0.004346   0.000504   8.624  < 2e-16 ***
## hr_data$time_spend_company     0.228638   0.014855  15.391  < 2e-16 ***
## hr_data$Work_accident         -1.498575   0.088254 -16.980  < 2e-16 ***
## hr_data$promotion_last_5years -1.768024   0.255495  -6.920 4.52e-12 ***
## hr_data$sales                  0.020587   0.007854   2.621  0.00876 **
## hr_data$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
```