PART FOUR: WHY DO OUR EMPLOYEES LEAVE US?

The problem would be analysed using R. Load libraries and the data first:

```
library(plyr)
library(dplyr)            #filter()
library(corrplot)         #corrplot()
library(ggplot2)          #ggplot()
library(gridExtra)        #grid.arrange()
library(Rmisc)            #multiplot()
library(e1071)            #naiveBayes()
library(rpart)            #raprt()
library(pROC)             #roc()
library(rpart.plot)       #raprt.plot()
library(randomForest)     #randomForest()
library(caret)            #confusionMatrix()
library(DT)               #datatable()
hr <- read.csv("../input/HRSurveyData.csv", stringsAsFactors = FALSE)
```

4.1 Understand Data

```
str(hr)
```

Result:

```
'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 ...
```

The data set contains ten attributes and has 14999 records. These ten attributes will be used to analyse employee attrition and we would build models to predict the turnover in part five of the report.

4.2 Data Pre-processing

Rename variables

There is a typo 'montly' in the dataset and we need to correct it and then simplify variables names.

```
hr <- rename(hr,
   satisfaction = satisfaction_level, evaluation = last_evaluation,
   project = number_project, monthlyhour = average_montly_hours,
   serviceyear = time_spend_company, accident = Work_accident,
   promotion = promotion_last_5years, dept = sales)
```

Transfer the variable "left" into a factor variable

```
hr$left <- factor(hr$left, levels = c("0","1"))
```

Add a new variable salarynum

```
hr$salarynum[hr$salary == "low"] <- 1
hr$salarynum[hr$salary == "medium"] <- 2
hr$salarynum[hr$salary == "high"] <- 3


summary(hr)
```

Result:

```
 satisfaction_level last_evaluation  number_project  average_montly_hours time_spend_company
 Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0        Min.   : 2.000
 1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0        1st Qu.: 3.000
 Median :0.6400     Median :0.7200   Median :4.000   Median :200.0        Median : 3.000
 Mean   :0.6128     Mean   :0.7161   Mean   :3.803   Mean   :201.1        Mean   : 3.498
 3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0        3rd Qu.: 4.000
 Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0        Max.   :10.000

 Work_accident         left        promotion_last_5years         sales         salary
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000       sales      :4140   high  :1237
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000       technical  :2720   low   :7316
 Median :0.0000   Median :0.0000   Median :0.00000       support    :2229   medium:6446
 Mean   :0.1446   Mean   :0.2381   Mean   :0.02127       IT         :1227
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000       product_mng: 902
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000       marketing  : 858
 salarynum
 Min.   :1.000
 1st Qu.:1.000
 Median :2.000
 Mean   :1.595
 3rd Qu.:2.000
 Max.   :3.000
```

4.3 Variable Description

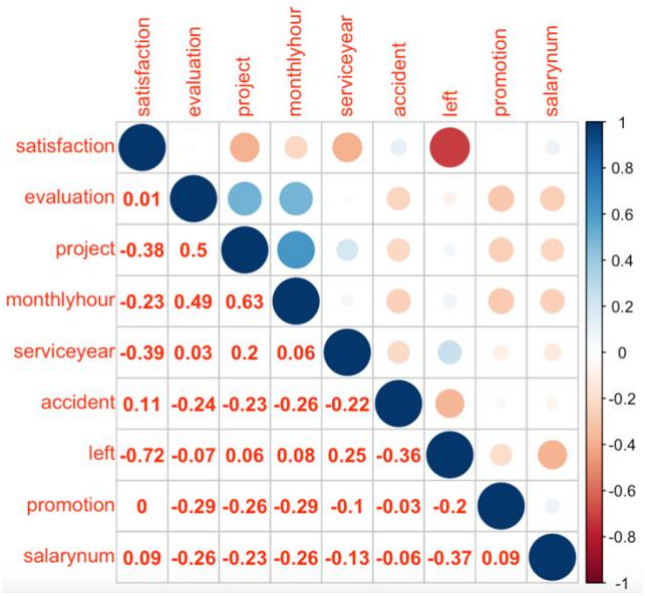
Based on the elementary analysis, we produce a table "Variable Description" as a data dictionary to summary the variables information:

Variable Type	Variable Name	Abbreviation	Definition	Range	Note
Dependent	left	left	Whether the employee has left	0 for has not left, 1 for left	23.81% employees has left
Independent	satisfaction_level	satisfaction	Employee's satisfaction level	0~1	0.6 on average
	last_evaluation	evaluation	Employee's last evaluation	0~1	0.7 on average
	number_project	project	Number of projects each employee have worked on	2~7	4 on average
	average_monthly_hours	monthlyhour	Average hours working each month	96~310	
	time_spend_company	serviceyear	Number of years each employee spent at the company	2~10	
	Work_accicdent	accident	Whether the employee have had an accident at work	0 for has not had an accident, 1 for has had an accident	14.46% has an accident on average
	promotion_last_5years	promotion	Whether the employee have had a promotion in the last 5 years	0 for not had a promotion, 1 for had a promotion	21.27% has promotion on average
	sales	dept	Employee's department	sales, hr, accounting etc.	
	salary	salary	Employee's salary	low, medium, high	48.78% for low, 8.25% for high
	salarynum	salarynum	Employee's salary	0, 1, 2	48.78% for low, 8.25% for high

4.4 Data Analysis and Visualisation: Correlation Matrix

For any two variables X and Y, the correlation between them represents the degree of linearly relation of X and Y (Lu, 2019). Correlation matrix show the relationship between multiple variables intuitively in a table (Lu, 2019).

```
1. hr_cor <- hr %>%
2. select(satisfaction:promotion, salarynum)
3. corrmatrix <- cor(hr_cor)
4. corrplot.mixed(corrmatrix)
5. corrplot(cor(corrmatrix),
   type="upper",method = "circle",tl.pos = "tl",tl.offset = 0.05)
6. corrplot(cor(corrmatrix),
   add=T,type="lower",method = "number",col="red",diag=F, tl.pos ="n",cl.pos ="n")
```



The size of circle represents the size of the correlation and the colour represents two variables are positively or negatively correlated. As the correlation matrix showed, the factors related to employee attrition are satisfaction level (-0.72), salary (-0.37), work accident (-0.36), promotion (-0.2), last evaluation (-0.07), number of projects involved (0.06), monthly working hours (0.08), serve year (0.25). In general, employees choose to leave the company because low satisfaction level, low salary level, no promotion opportunity, long serve year and long monthly working hours.

4.5 Data Analysis and Visualisation: Exploratory Data Analysis

Transfer variable data type into factor data type

```
1. hr$accident <- as.factor(hr$accident)
2. hr$left <- as.factor(hr$left)
3. hr$promotion <- as.factor(hr$promotion)
4. hr$dept <- as.factor(hr$dept)
5. hr$salary <- as.factor(hr$salary)
```

Analysing categorical variables

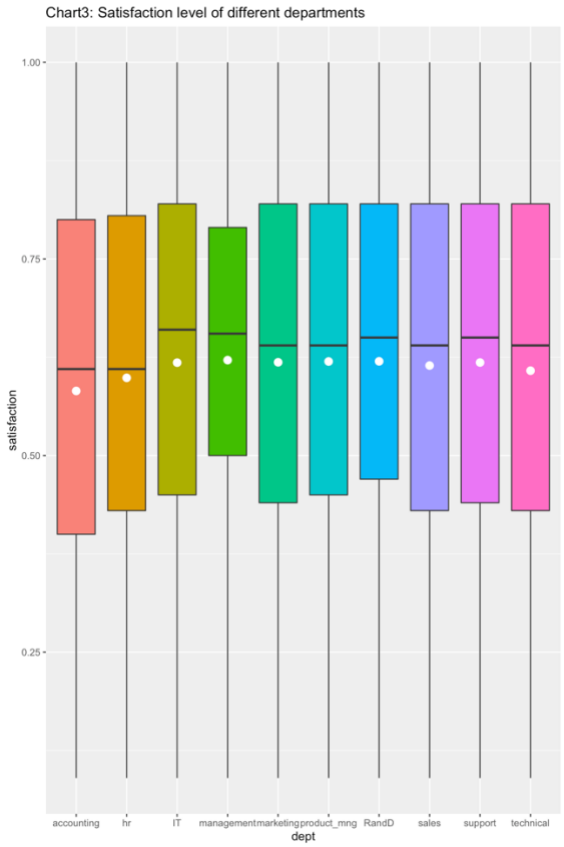
```
g1 <- ggplot(group_by(hr, dept), aes(x = dept, fill = dept)) + geom_bar(width = 1) +
      coord_polar(theta = "x") + ggtitle("Chart1: Number of employees in different departments")
g2 <- ggplot(group_by(hr, dept), aes(x = dept, fill = salary)) + geom_bar(width = 1) +
      coord_polar(theta = "x") + ggtitle("Chart2: Salary status in different departments")
multiplot(g1, g2, cols = 2)
```

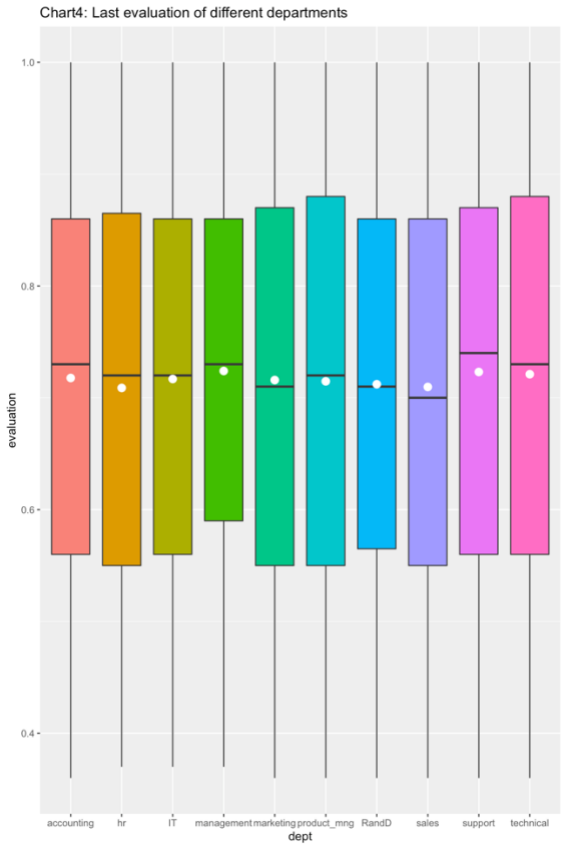
1) The employees in sales, support and technical departments are dominant in terms of population in this company.

2) Except management, more employees working in other departments get medium- and high-level salary as the population in department increases.

Analysing categorical and numerical variables

```
g3 <- ggplot(hr, aes(x = dept, y = satisfaction, fill = dept)) + geom_boxplot() +
      ggtitle("Chart3: Satisfaction level of different departments") +
      stat_summary(fun.y = mean,size = 3, color = 'white', geom = "point") + theme(legend.position = "none")
g4 <- ggplot(hr, aes(x = dept, y = evaluation, fill = dept)) + geom_boxplot() +
      ggtitle("Chart4: Last evaluation of different departments") +
      stat_summary(fun.y = mean,size = 3, color = 'white', geom = "point") + theme(legend.position = "none")
grid.arrange(g3, g4, ncol = 2)
```



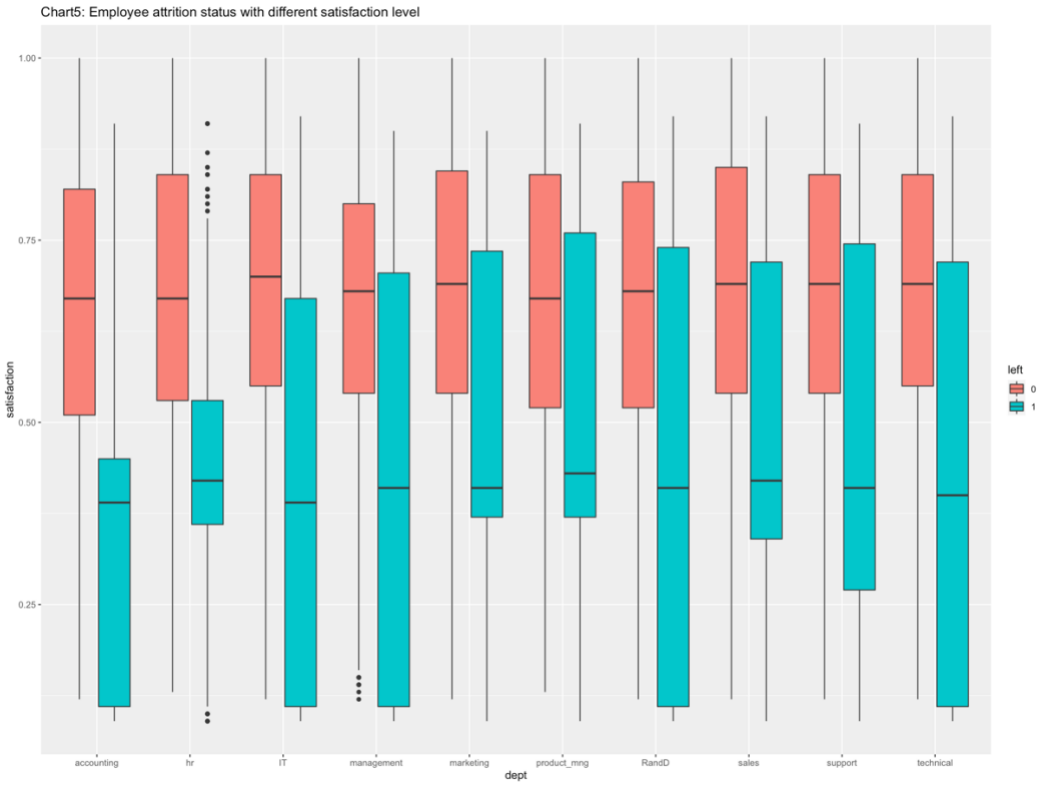


1) Employees in accounting department get lower satisfaction and satisfaction level is biggest employee turnover factor analysed in correlation matrix. This indicates that low satisfaction maybe the major reason for accounting employee attrition.

2) For last evaluation score, it is pretty similar among different departments.

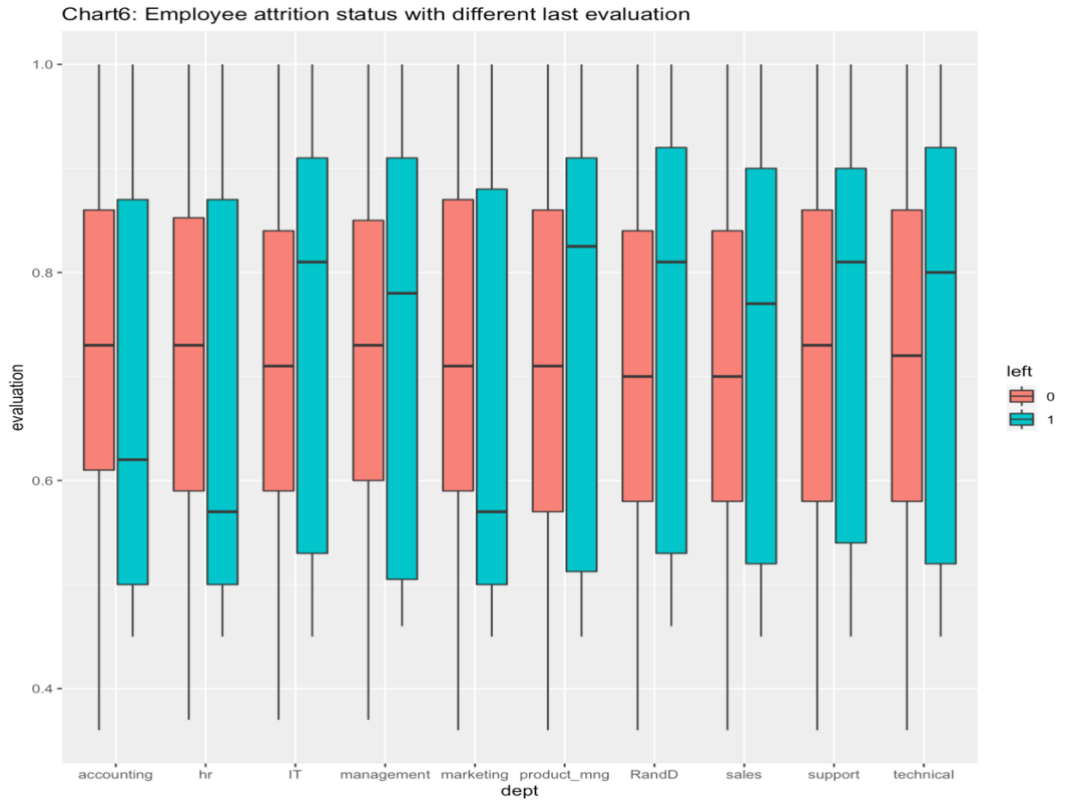
Analysing possible reasons of employee attrition

```
1. ggplot(hr, aes(x = dept, y = satisfaction, fill = left)) + geom_boxplot() +
2. ggtitle("Chart5: Employee attrition status with different satisfaction level")
```



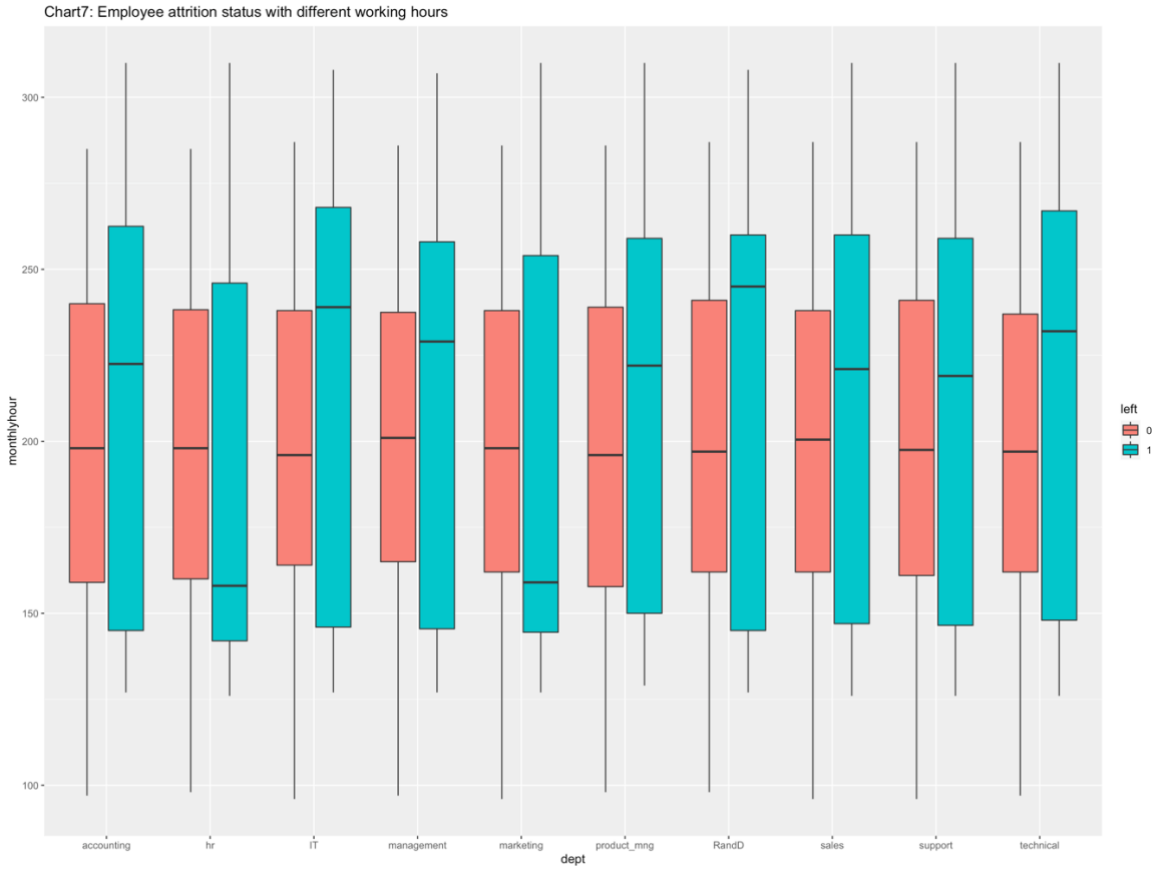
1) People who leave the company normally get low satisfaction level for all departments.

```
3. ggplot(hr, aes(x = dept, y = evaluation, fill = left)) + geom_boxplot() +
4. ggtitle("Chart6: Employee attrition status with different last evaluation")
```



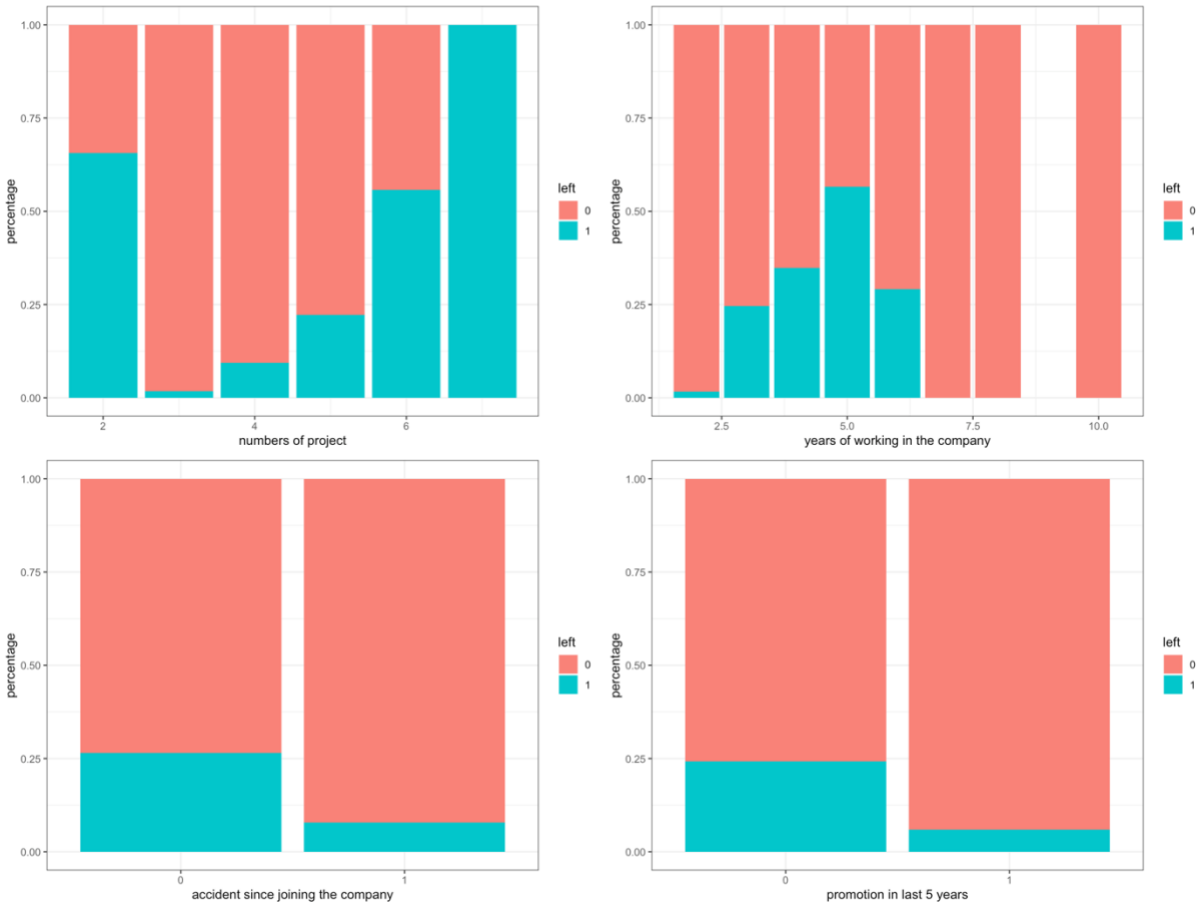
1) Employees in accounting, hr and marketing departments get low last evaluation score and they finally leave.

```
5.  ggplot(hr, aes(x = dept, y = monthlyhour, fill = left)) + geom_boxplot() +
6.  ggtitle("Chart7: Employee attrition status with different working hours")
```



1) For people who still stay in the company, they work in a stable pace, with similar monthly working hours. However, for employees in hr and marketing departments, those who do not work hard finally leave.

```r
b1 <- ggplot(hr, aes(x = project, fill = left)) +
      geom_bar(position = 'fill') +
      theme_bw() + labs(x = 'numbers of project', y = 'percentage')
b2 <- ggplot(hr, aes(x = serviceyear, fill = left)) +
      geom_bar(position = 'fill') +
      theme_bw() + labs(x = 'years of working in the company', y = 'percentage'
)
b3 <- ggplot(hr, aes(x = accident, fill = left)) +
      geom_bar(position = 'fill') +
      theme_bw() + labs(x = 'accident since joining the company', y = 'percenta
ge')
b4 <- ggplot(hr, aes(x = promotion, fill = left)) +
      geom_bar(position = 'fill') +
      theme_bw() + labs(x = 'promotion in last 5 years', y = 'percentage')
grid.arrange(b1, b2, b3, b4, ncol = 2, nrow = 2)
```

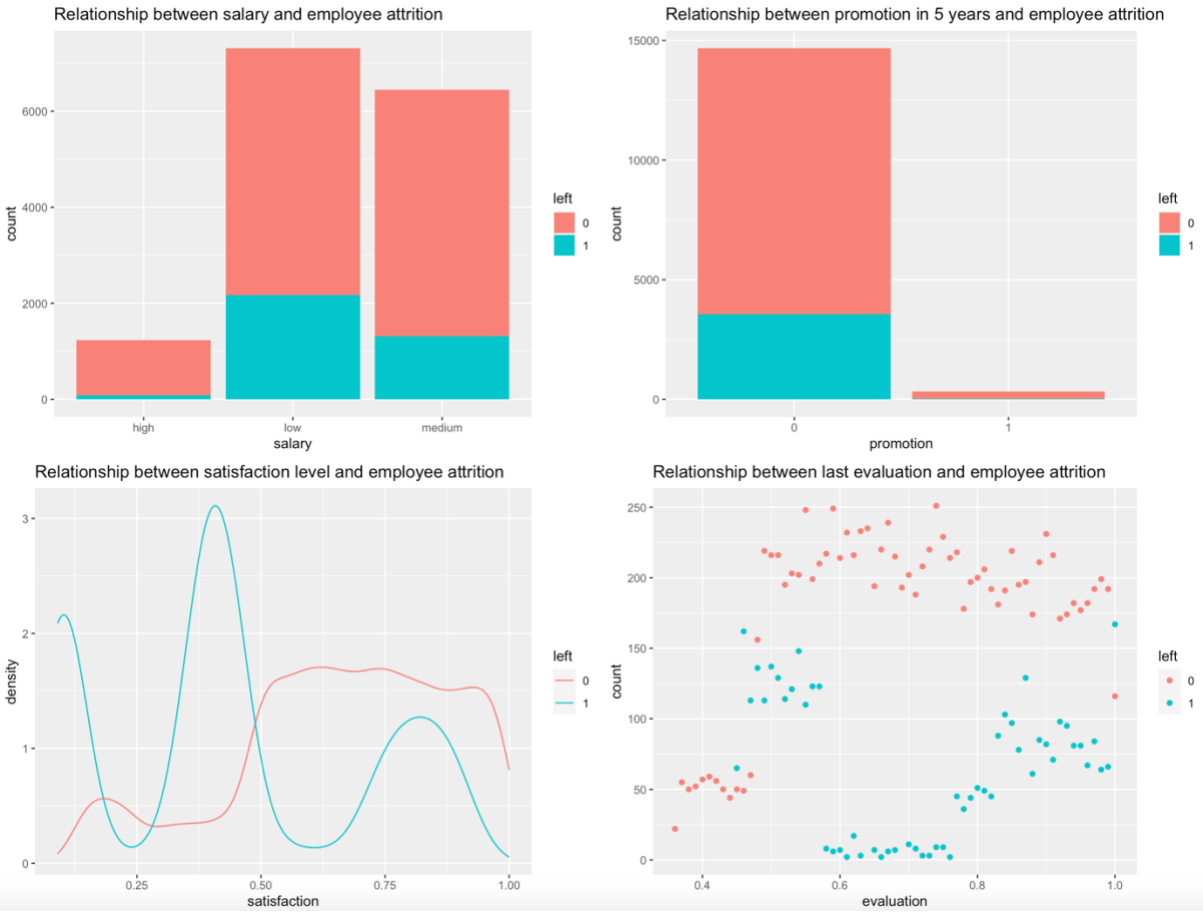


1) After involving three projects, more projects finished, more likely to leave.

2) People who work for three to six years in the company have highest turnover rate.

3) Employees who do not have accident and do not get promotion in last 5 years have higher turnover rate.

```
1. b6 <- ggplot(hr, aes(x = salary, fill = left)) + geom_histogram(stat = "count")
       + ggtitle("Relationship between salary and employee attrition")
2. b7 <- ggplot(hr, aes(x = promotion, fill = left)) +
         geom_histogram(stat = "count") +
   ggtitle("Relationship between promotion in 5 years and employee attrition")
3. b8 <- ggplot(hr, aes(x = satisfaction, color = left)) +
         geom_line(stat = "density") +
   ggtitle("Relationship between satisfaction level and employee attrition")
4. b9 <- ggplot(hr, aes(x = evaluation, color = left)) +
         geom_point(stat = "count") +
   ggtitle("Relationship between last evaluation and employee attrition")
5. grid.arrange(b6, b7, b8, b9, ncol = 2, nrow = 2)
```



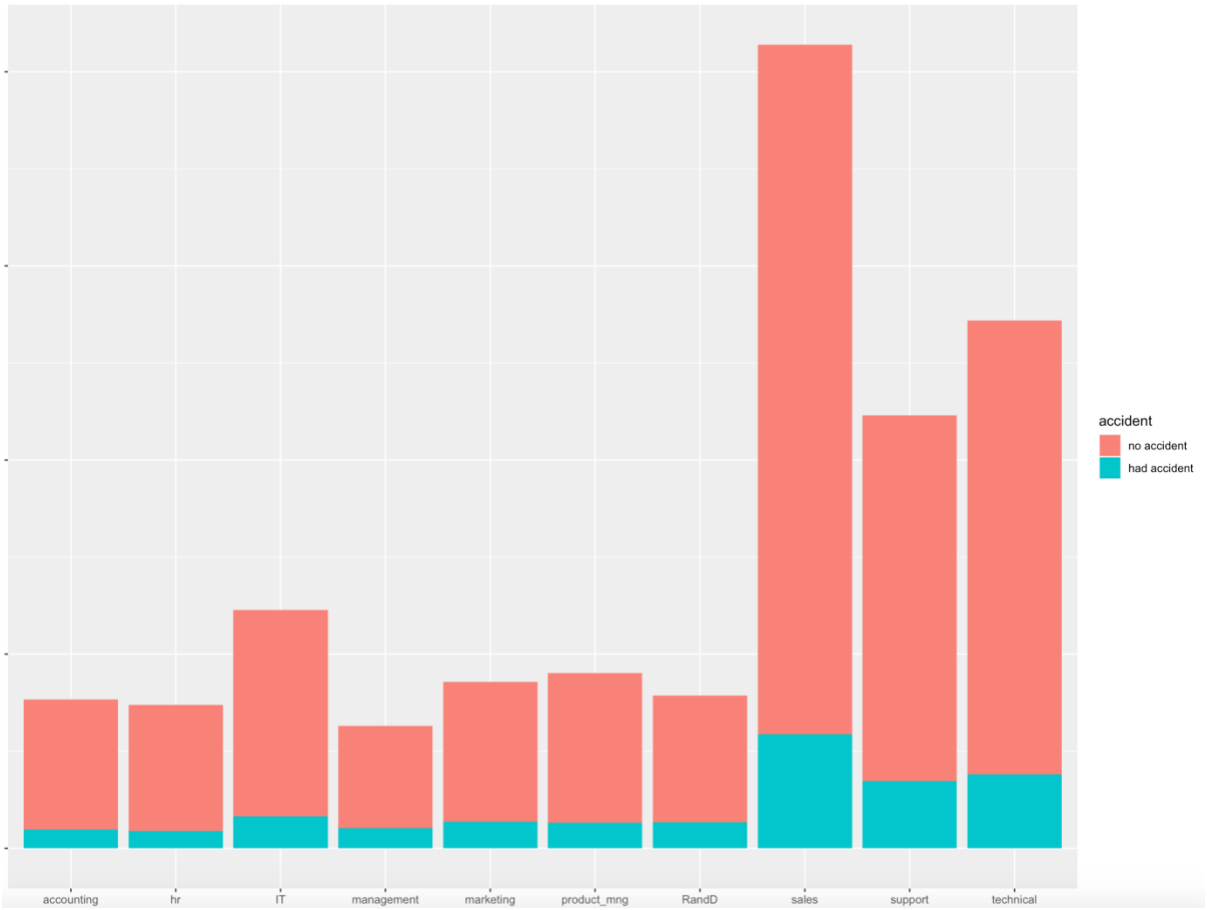
1) Employees who get high salary rarely leave.

2) If employees get promotion in the recent 5 years, most of them will not leave.

3) Retired employees rate satisfaction score in different regions, so we cannot say that employees who get low satisfaction level will have high probability of leaving.

4) A lot of retired employees will get high last evaluation score, so we cannot argue that those who get high evaluation level will probably not leave.

```
hr %>%
group_by(dept) %>%
ggplot(aes(y = dept,fill = accident)) + geom_bar() + coord_flip() +
theme(axis.text.y = element_blank(), axis.title.y = element_blank(),
axis.title.x = element_blank()) +
scale_fill_discrete(labels = c("no accident", "had accident"))
```



1) The number of people in each department who had accident during work is positive correlated with the population of the department. Therefore, accident is not a major reason for leaving the company.

4.6 Data Analysis Summary

The major reasons of employee attrition are listed below:

1) Low salary: High salary's employees rarely leave.
2) Overload work time: Overload monthly working hours, but for employees in hr and marketing departments, they will leave because of too little working hours.
3) Strong working ability: After participating in 2 project, ore projects participated afterward, more likely to leave. About half of those whose last evaluation score is high leave. Those who stay in the company for three to six years are likely to leave.
4) No promotion in last 5 years: If getting promotion, very few employees leave. If not, over 2/7 $\approx$ 28.6% of employees leave.

PART FIVE: MODELING TO PREDICT EMPLOYEE ATTRITION

Predicting employee turnover is a classification problem, with two classes of leave or not leave. Three classification models would be built and evaluated under AUC and confusion matrix. Then, we would choose the best model to build an interactive table for managers to check which employee is predicted to leave.

5.1 Naive Bayes Model

Naive Bayes is based on Bayes' theorem, under the assumption that there is an independent relationship between predictor variables and other variables (Bhartiya et al., 2019). The model is suitable for large datasets and it utilise the probabilities of attributes to make predictions (Bhartiya et al., 2019).

Splitting the data into training and test set

```
1. set.seed(1234)
2. n <- nrow(hr)
3. rnd <- sample(n, n*.70)
4. train <- hr[rnd,]
5. test <- hr[-rnd,]
```

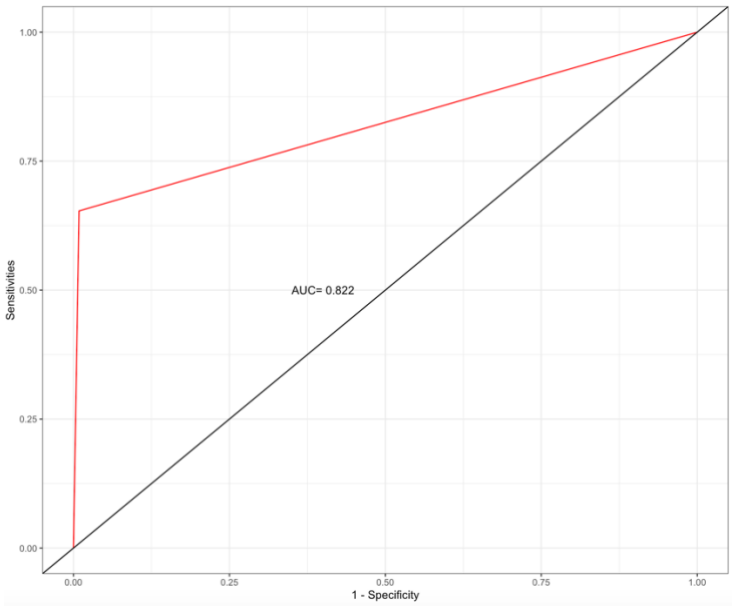
Build Naive Bayes Model

```
1. train_control <- trainControl(method = 'cv', number = 5)
2. nbmodel <- train(left ~ ., data = train,
   trControl = train_control, method = 'nb')
3. pred_nb <- predict(nbmodel, test[-7])
```

Build ROC Curve and check the model's AUC

Receiver operating characteristics (ROC) curve is a graph to general evaluate the prediction accuracy of machine learning models (Fawcett, 2006). The area under ROC curve is named as Area under the ROC curve (AUC) and the higher AUC means the higher accuracy of prediction (Fawcett, 2006). AUC will be used as one criterion to evaluate the general performance of our models.

```
pred_nb <- as.numeric(as.character(pred_nb))
roc_nb <- roc(test$left, pred_nb)
Specificity <- roc_nb$specificities
Sensitivity <- roc_nb$sensitivities
ggplot(data = NULL, aes(x = 1- Specificity, y = Sensitivity)) +
          geom_line(colour = 'red') + geom_abline() +
          annotate('text', x = 0.4, y = 0.5, label = paste('AUC=',
          round(roc_nb$auc, 3))) + theme_bw() +
          labs(x = '1 - Specificity', y = 'Sensitivities')
```



The Naive Bayes Model has AUC of 0.822, which is higher than 0.75 and this indicates the model is a good model to do prediction.

Use confusion matrix to check the model's accuracy

Confusion matrix is our second criterion to evaluate classification models. According to Rohit Hebbar et al. (2018), confusion matrix is commonly used to find the best performing machine learning model and it is formatted like the graph showed below. The corresponding illustration of terminologies is also listed. In our case, TP and TN represent the model correctly predict situations that employees who would not leave and would leave the company. Therefore, TP and TN are important indicators of our model's detailed accuracy in predicting the two situations.

Actual/Predicted	0	1
0	TN	FP
1	FN	TP

	Meaning
True Positive (TP)	the observation is predicted to be positive when the actual observation is positive
False Negative (FN)	the observation is predicted to be negative when the actual observation is positive
False Positive (FP)	the observation is predicted to be positive when the actual observation is negative
True Negative (TN)	the observation is predicted to be negative when the actual observation is negative

```
1.  prop.table(table(pred_nb, test$left, dnn = c("Actual", "Predicted")), 1)
```

Result:

```
1.        Predicted
2.  Actual          0          1
3.       0 0.75505668 0.08246277
4.       1 0.00689042 0.15559013
```

With the accuracy rate of near 75.5%, the model is a good one when predict that employees do not leave the company. However, it does not perform well when predict employee turnover, as it only gets around 15.6% accuracy rate. This indicates that we should try other models.

5.2 Decision Tree Model

Decision tree is used to find the correlations between attributes in an entire dataset and the corresponding graph can be produced as a decision support tool (Chesney, 2009). A Decision Tree Model would be built and accessed using AUC and confusion matrix as before.

Build the Decision Tree Model

```
dtree <- rpart(left ~., data = train)
preds <- predict(dtree, test, type = "class")
```

Build ROC Curve and check the model's AUC

```
rocv <- roc(as.numeric(test$left), as.numeric(preds))
rocv$auc
```

Result:

```
Area under the curve: 0.9524
```

The AUC is above 90%, which indicate the Decision Tree Model is much better than the Naive Bayes Model.

Use confusion matrix to check the model's accuracy

```
prop.table(table(test$left, preds, dnn = c("Actual", "Predicted")), 1)
```

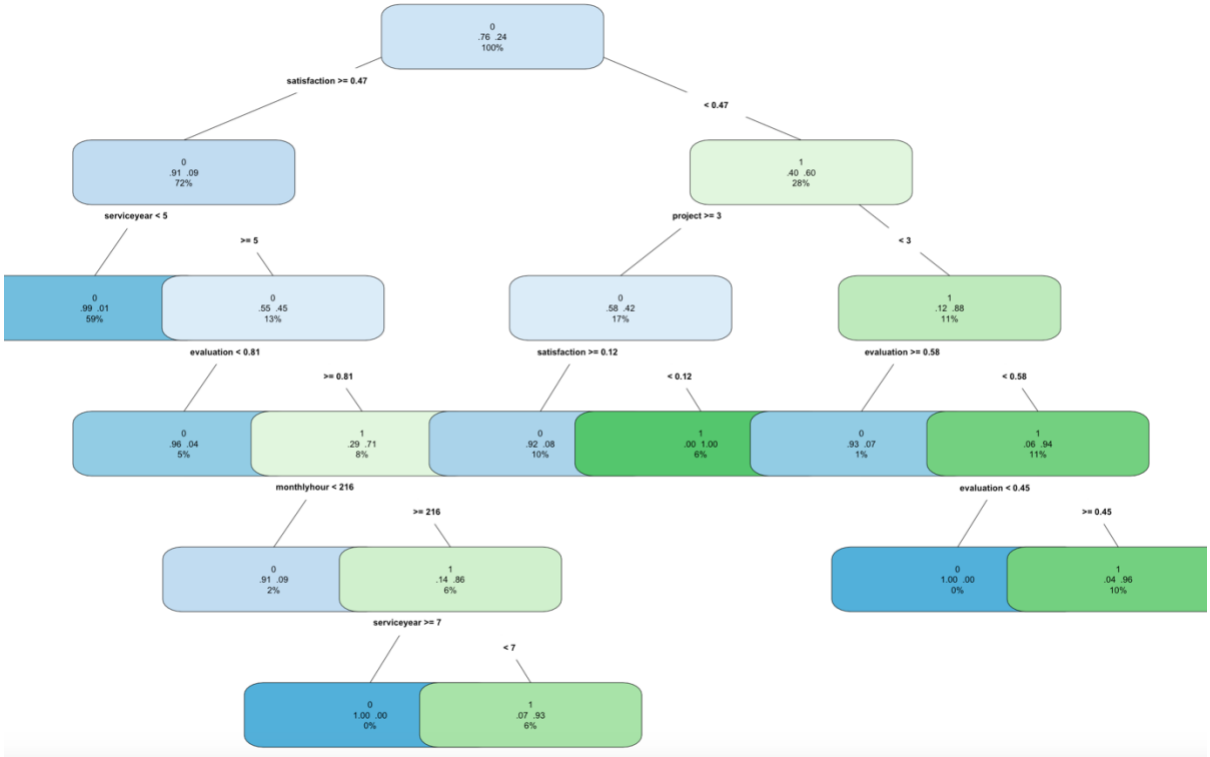
Result:

```
        Predicted
Actual          0          1
     0 0.98861646 0.01138354
     1 0.08379888 0.91620112
```

Again, the model has high accuracy of predicting both employees who leave and do not leave the company. Therefore, we can conclude that the Decision Tree Model is better than Naive Bayes Model in this case.

Draw the decision tree diagram

```
1.  rpart.plot(dtree,
2.                type = 4,
3.                extra = 104,
4.               tweak = 0.8,
5.                fallen.leaves = F,
6.                cex = 0.7)
```



As the decision tree showed above, satisfaction is the most important factor of employee attrition. Number of projects involved, serve year, monthly working hours, last evaluation are also important factors. This Decision Tree can help managers predict an employee will leave or not intuitively.

5.3 Random Forest Model

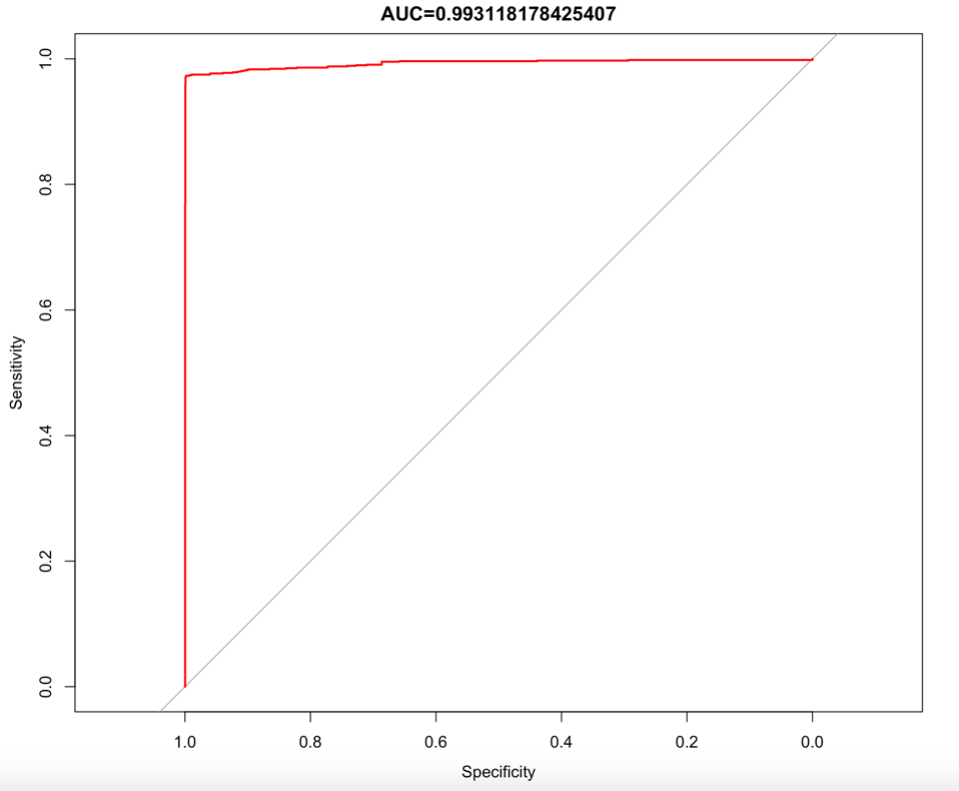
Random Forest builds many decision trees and then combine them to perform a better prediction (Bhartiya et al., 2019). A Random Forest Model would be established and evaluated using the same criteria as before.

Build Random Forest Model

```
1. index <- sample(2, nrow(hr), replace = T, prob = c(0.7,0.3))
2. train <- hr[index == 1,]; test <- hr[index == 2,]
3. rfmodel <- randomForest(left~., data = train)
4. predict.hr <- predict(rfmodel, test)
```

Build ROC Curve and check the model's AUC

```
1. prob.hr <- predict(rfmodel, test, type = "prob")
2. roc.hr <- roc(test$left, prob.hr[,2], levels = levels(test$left))
3. plot(roc.hr,type="S", col ="red", main = paste("AUC=", roc.hr$auc, sep = ""))
```



The AUC is extremely high, so among the tree models above, this company should choose the Random Forest Model to predict employee attrition.

Use confusion matrix to check the model's accuracy

```
confusionMatrix(test$left, predict.hr)
```

Result:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3362    3
         1   32 1039

               Accuracy : 0.9921
                 95% CI : (0.989, 0.9945)
    No Information Rate : 0.7651
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9783

 Mcnemar's Test P-Value : 2.214e-06

            Sensitivity : 0.9906
            Specificity : 0.9971
         Pos Pred Value : 0.9991
         Neg Pred Value : 0.9701
             Prevalence : 0.7651
         Detection Rate : 0.7579
   Detection Prevalence : 0.7586
      Balanced Accuracy : 0.9938

       'Positive' Class : 0
```

Again, Random Forest Model is the best one, having extremely high accuracy in predicting employees who would leave and not leave.

5.4 Interactive Table of Predicting Employee Attrition

```
1. predict.hr <- as.numeric(as.character(predict.hr))
2. pred_end <- predict(rfmodel, test[-7], type = 'prob')
3. data_end <- cbind(round(pred_end, 3), predict.hr)
4. datatable(data_end)
```

Show 25 entries Search:

	0	1	predict.hr
2	0.032	0.968	1
6	0	1	1
7	0	1	1
10	0	1	1
11	0	1	1
12	0	1	1
14	0	1	1
16	0	1	1
18	0.002	0.998	1
19	0.066	0.934	1
23	0	1	1
25	0	1	1
31	0.01	0.99	1
32	0	1	1
34	0	1	1
36	0.002	0.998	1
38	0.002	0.998	1
43	0	1	1

Showing 1 to 25 of 4,436 entries

Previous 1 2 3 4 5 … 178 Next

Based on the Random Forest Model which has 99.3% accuracy rate, we build an interactive table. In this table, 1 for predicting employee turnover and 0 for predicting employees will not leave, and 'predict.hr' represent the prediction result. For example, employee whose id is 2 has 96.8% for leaving and this employee actually leave. This table would be useful for managers to check which employee would leave intuitively and the corresponding possibility. After knowing which employees would leave, and the major reasons analysed in the previous section, managers may take actions to prevent employees from leaving efficiently.

REFERENCES

Bhartiya N., Jannu S., Shukla P. and Chapaneri R. (2019), 'Employee Attrition Prediction Using Classification Models'*, 2019 IEEE 5th International Conference for Convergence in Technology (I2CT),* Bombay, India, pp.1-6.

Chesney, T. (2009) *Searching For Patterns: How We Can Know without Asking*. Nottingham: Nottingham University Press.

Fawcett, T. (2006), 'An introduction to ROC analysis', *Pattern Recognition Letters,* 27(8), pp.861-874.

Lu, B. (2019), *Demystifying The Correlation Matrix,* Data Driven Investor. Available at: https://www.datadriveninvestor.com/2019/04/24/demystifying-the-correlation-matrix/ (Accessed: 01 May 2020)

Rohit Hebbar A., Patil S. H., Rajeshwari S. B. and Saqquaf, S. S. M. (2018), 'Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees', *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT),* Bangalore, India, pp. 934-938