Why emoployee turnover

Zhong guangbin

2021/7/13

1 Load package

```
library(tidyverse)
library(VIM)
library(randomForest)
library(GGally)
library(glmnet)
```

2 explore data

2.1 Load data

```
d <- read.csv("HR_comma_sep.csv")
```

2.2 view data

2.2.1 What's the turnover rate in this company

```
head(d)
```

```
##   satisfaction_level last_evaluation number_project average_montly_hours
## 1               0.38            0.53              2                  157
## 2               0.80            0.86              5                  262
## 3               0.11            0.88              7                  272
## 4               0.72            0.87              5                  223
## 5               0.37            0.52              2                  159
## 6               0.41            0.50              2                  153
##   time_spend_company Work_accident left promotion_last_5years sales salary
## 1                  3             0    1                     0 sales    low
## 2                  6             0    1                     0 sales medium
## 3                  4             0    1                     0 sales medium
## 4                  5             0    1                     0 sales    low
## 5                  3             0    1                     0 sales    low
## 6                  3             0    1                     0 sales    low
```

```
str(d)
```

```
## '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 ...
##  $ salary               : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 ...
```

```
summary(d)
```

```
##  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
##
##          sales         salary
##  sales      :4140   high  :1237
##  technical  :2720   low   :7316
##  support    :2229   medium:6446
##  IT         :1227
##  product_mng: 902
##  marketing  : 858
##  (Other)    :2923
```

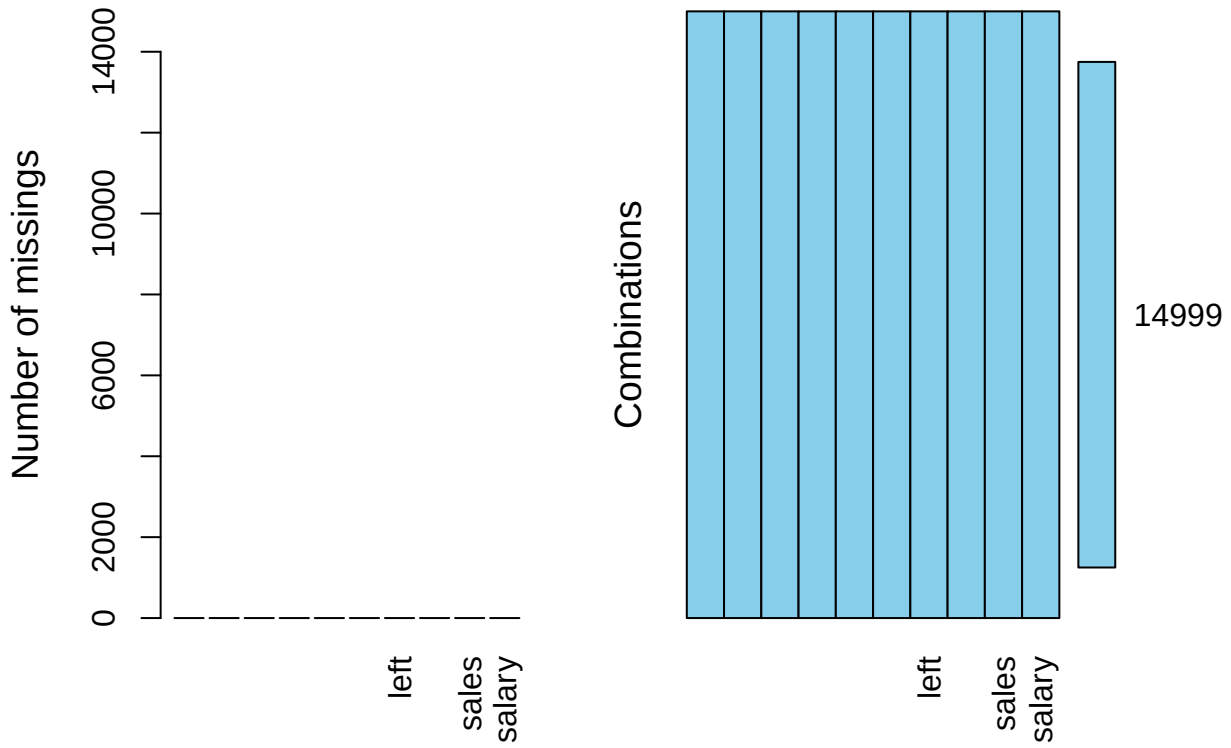
```
 d %>%
  group_by(factor(left)) %>%
  summarise(counts = n()) %>%
  mutate(ratio = counts/sum(counts))
```

```
## # A tibble: 2 x 3
##   `factor(left)` counts ratio
##   <fct>           <int> <dbl>
## 1 0               11428 0.762
## 2 1                3571 0.238
```

The turnover rate was as high as 23.8%

2.2.2 Check for missing values

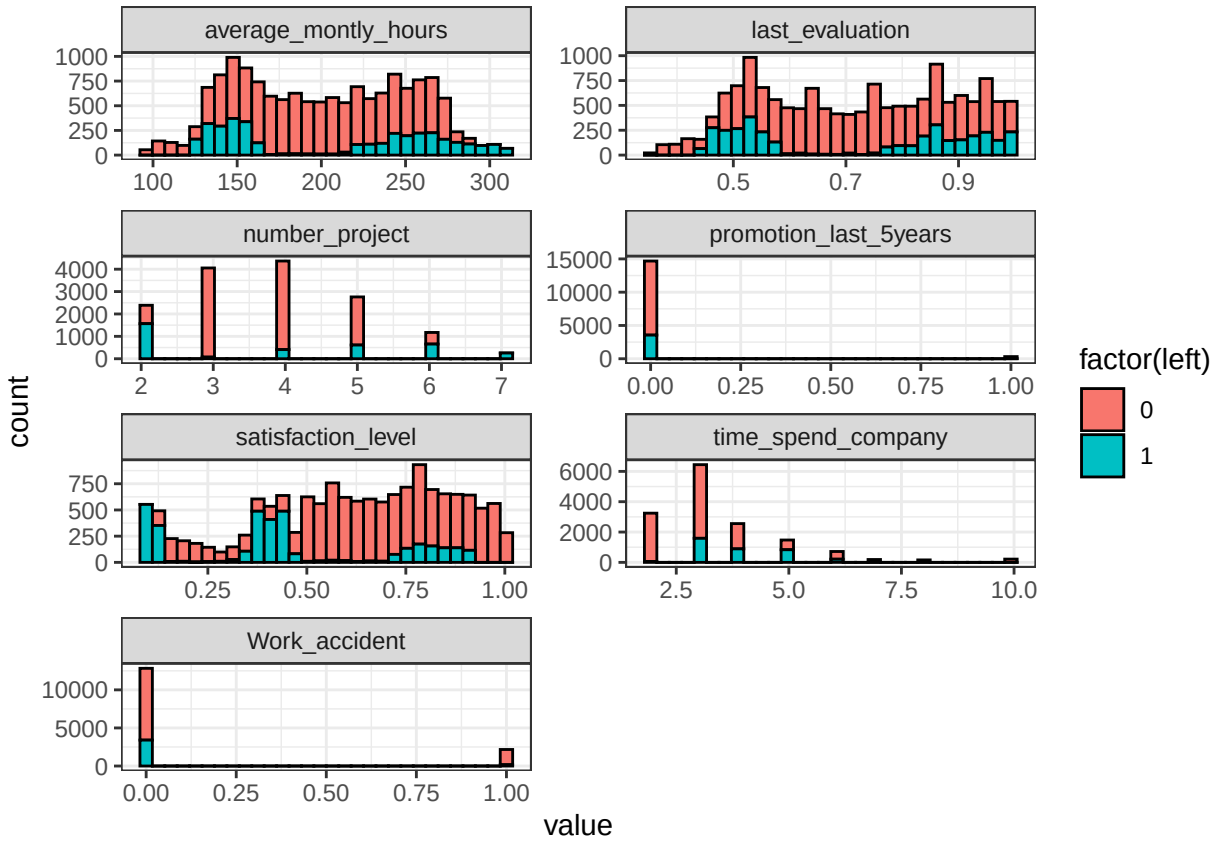
```
aggr(d,prop = F, number = T)
```



No missing data

2.2.3 Distribution of characteristics of departing employees and existing employees

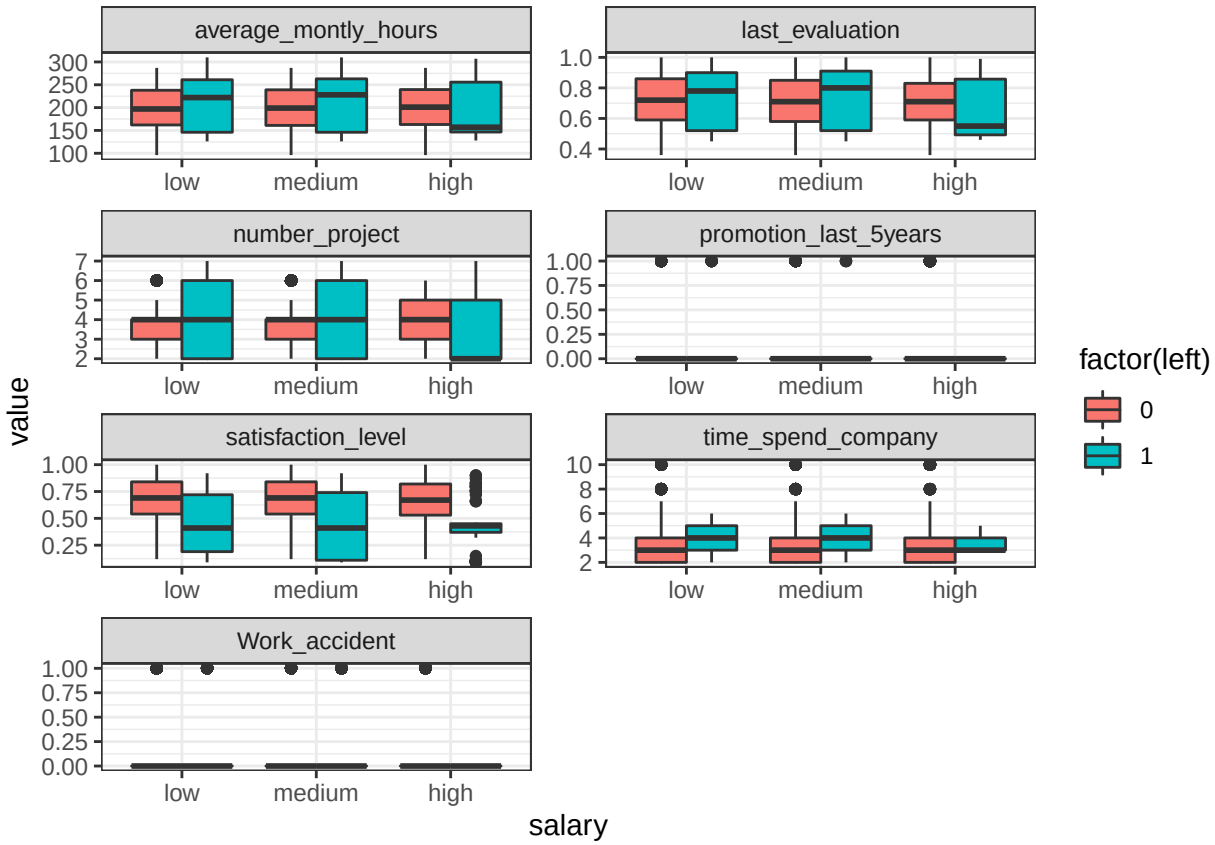
```
d %>%
   pivot_longer(c(1:6,8),names_to = "type",values_to = "value") %>%
   ggplot(aes(value)) +
   theme_bw() +
   geom_histogram(aes(fill = factor(left)),color = "black") +
   facet_wrap(~type,ncol = 2,scales = "free")
```



The employees who left were involved in more projects, and all of the employees who left were promoted within five years

```
d$salary <- factor(d$salary,
                   levels = c("low","medium","high"))

d %>%
   pivot_longer(c(1:6,8),names_to = "type",values_to = "value") %>%
   ggplot(aes(salary,value)) +
   theme_bw() +
   geom_boxplot(aes(fill = factor(left)),position = "dodge") +
   facet_wrap(~type,ncol = 2,scales = "free")
```



Among the departed employees, those with low and medium salaries devote more time and projects, and their satisfaction has decreased significantly compared with last time. This may be one of the reasons for the resignation of the employees (the salary is not fully paid). While the time spent on a high salary didn't differ much, the number of projects they participated in declined (marginalization), notice that the high salary group also reported low levels of satisfaction last month, and this may have been the case for a long time.

2.2.4 Which department has the higher turnover rate

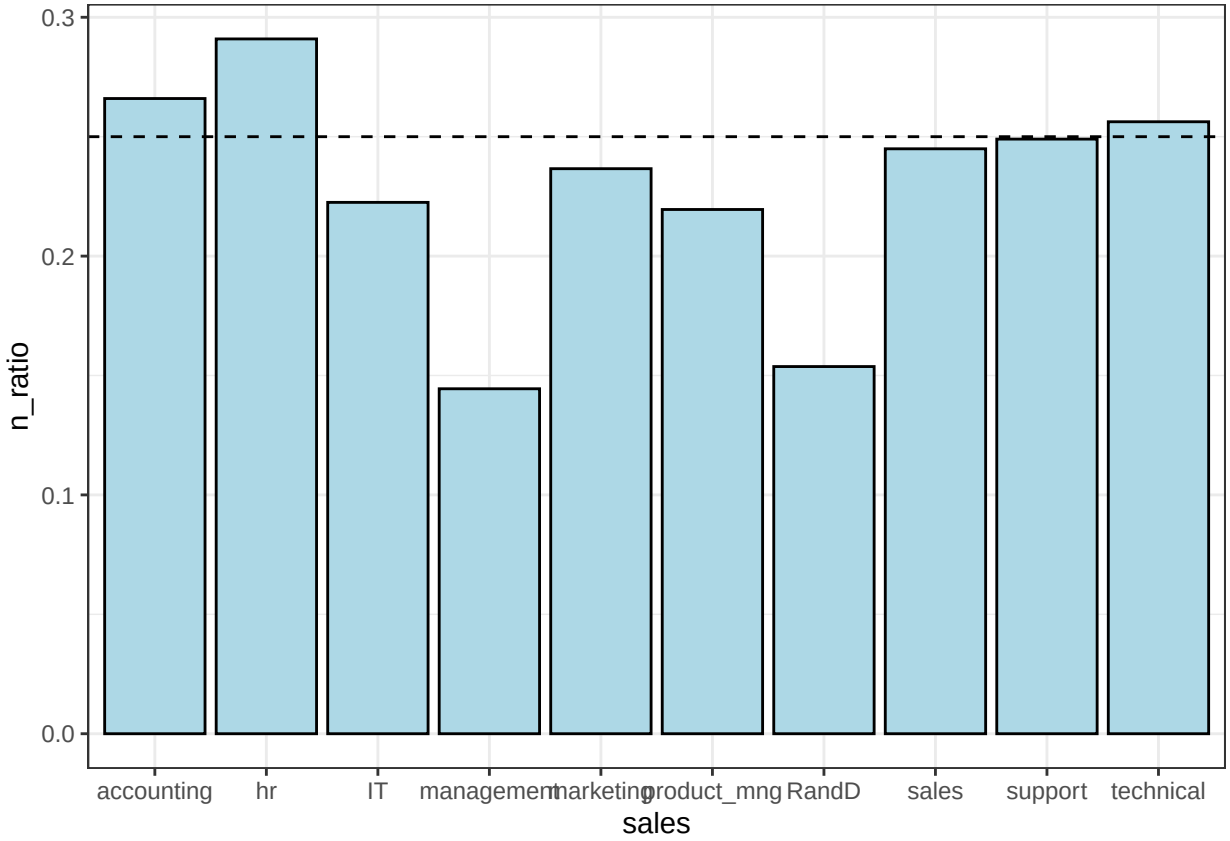
```
d1 <- d %>%
   group_by(sales,left) %>%
   mutate(n = n()) %>%
   group_by(sales) %>%
   mutate(total_n = n(),n_ratio = n/total_n) %>%
   distinct(n_ratio,.keep_all = T) %>%
   filter(left == 1) %>%
   arrange(desc(n_ratio))
   head(d1[,c(9,13)])
```

```
## # A tibble: 6 x 2
## # Groups:   sales [6]
##   sales      n_ratio
##   <fct>        <dbl>
## 1 hr           0.291
## 2 accounting   0.266
## 3 technical    0.256
```

```
## 4 support      0.249
## 5 sales        0.245
## 6 marketing    0.237
```

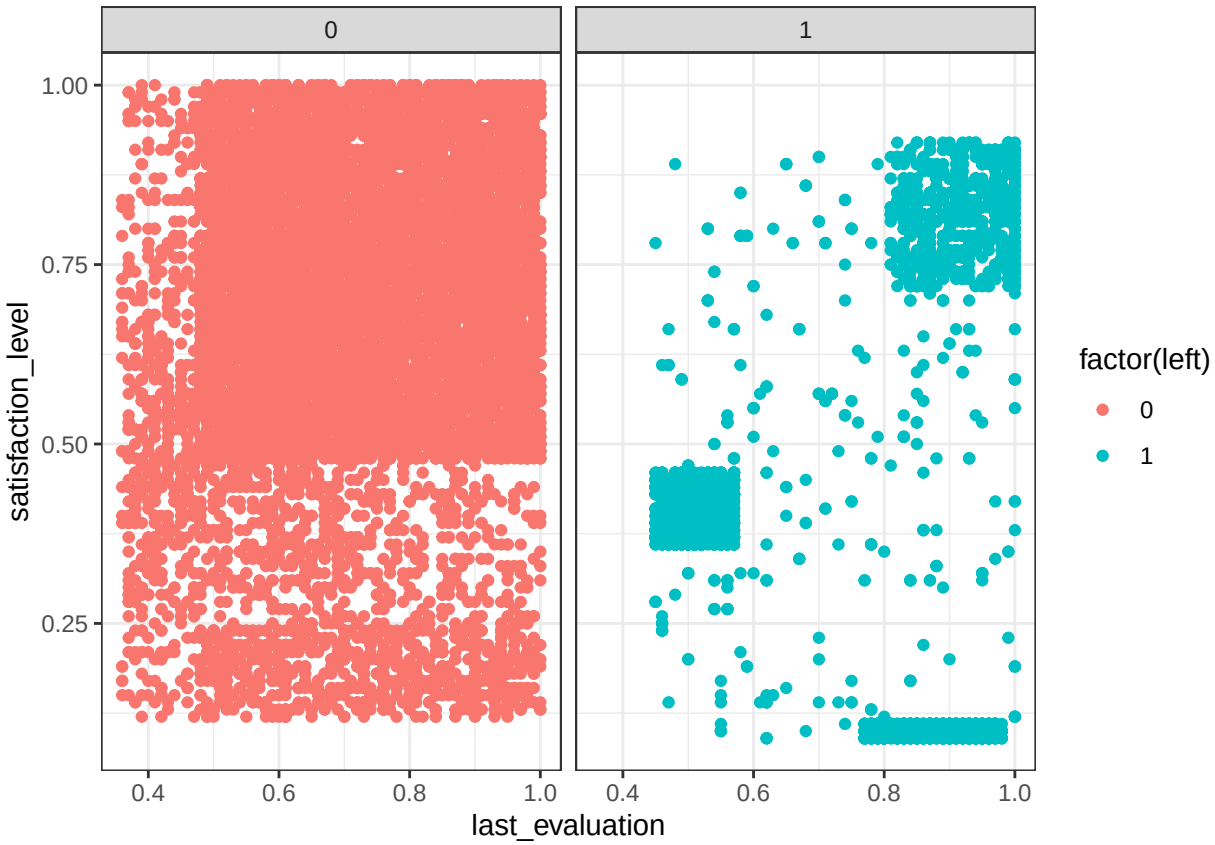
The top three parts of the turnover rate are HR, Accounting and Technical in turn,respectively, at 29.1%, 26.6% and 25.6%

```
  ggplot(d1, aes(sales,n_ratio)) +
  theme_bw() +
  geom_col(color = "black",fill = "lightblue") +
  geom_hline(yintercept = 0.25,linetype = "dashed")
```



The turnover rate of managment department and RandD department is significantly lower than that of other departments

```
ggplot(data = d,aes(last_evaluation,satisfaction_level)) +
  theme_bw() +
  geom_point(aes(color = factor(left))) +
  facet_wrap(~factor(left))
```



Judging from the two ratings, there are three main types of employees who leave the company: The first, whose scores dropped significantly (bottom right), may have been unhappy at work during that time; The second, rated highly both times (top right), may have been lured away by competing jobs; The third kind, two grades are not high, may be long-term work is not happy

3 Data analysis

3.1 which cause left

3.1.1 Pre-predictive processing

```
str(d)
```

```
## '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 "low","medium",..: 1 2 2 1 1 1 1 1 1 1 ...
```

```r
cor(d[,-c(9,10)])
```

```
##                        satisfaction_level last_evaluation number_project
## satisfaction_level            1.00000000     0.105021214   -0.142969586
## last_evaluation               0.10502121     1.000000000    0.349332589
## number_project               -0.14296959     0.349332589    1.000000000
## average_montly_hours         -0.02004811     0.339741800    0.417210634
## time_spend_company           -0.10086607     0.131590722    0.196785891
## Work_accident                 0.05869724    -0.007104289   -0.004740548
## left                         -0.38837498     0.006567120    0.023787185
## promotion_last_5years         0.02560519    -0.008683768   -0.006063958
##                        average_montly_hours time_spend_company Work_accident
## satisfaction_level            -0.020048113       -0.100866073   0.058697241
## last_evaluation                0.339741800        0.131590722  -0.007104289
## number_project                 0.417210634        0.196785891  -0.004740548
## average_montly_hours           1.000000000        0.127754910  -0.010142888
## time_spend_company             0.127754910        1.000000000   0.002120418
## Work_accident                 -0.010142888        0.002120418   1.000000000
## left                           0.071287179        0.144822175  -0.154621634
## promotion_last_5years         -0.003544414        0.067432925   0.039245435
##                                      left promotion_last_5years
## satisfaction_level           -0.38837498          0.025605186
## last_evaluation               0.00656712         -0.008683768
## number_project                0.02378719         -0.006063958
## average_montly_hours          0.07128718         -0.003544414
## time_spend_company            0.14482217          0.067432925
## Work_accident                -0.15462163          0.039245435
## left                          1.00000000         -0.061788107
## promotion_last_5years        -0.06178811          1.000000000
```

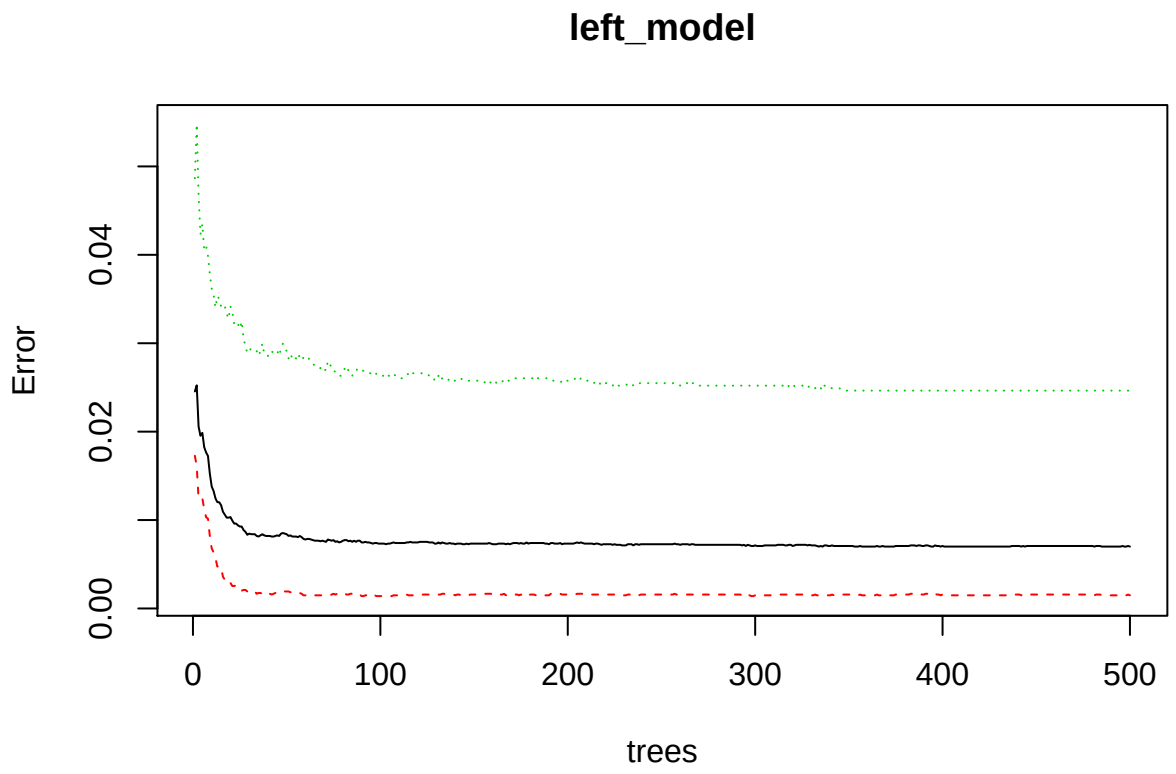
```r
factor_vars <- c("Work_accident","left","promotion_last_5years")
d[factor_vars] <- lapply(d[factor_vars], function(x) as.factor(x))
```

3.1.2 Prediction and Plot

```r
set.seed(2021)
left_model <- randomForest(left~.,data = d)
print(left_model)
```

```
##
## Call:
##  randomForest(formula = left ~ ., data = d)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 0.7%
## Confusion matrix:
##       0    1 class.error
## 0 11411   17 0.001487574
## 1    88 3483 0.024642957
```

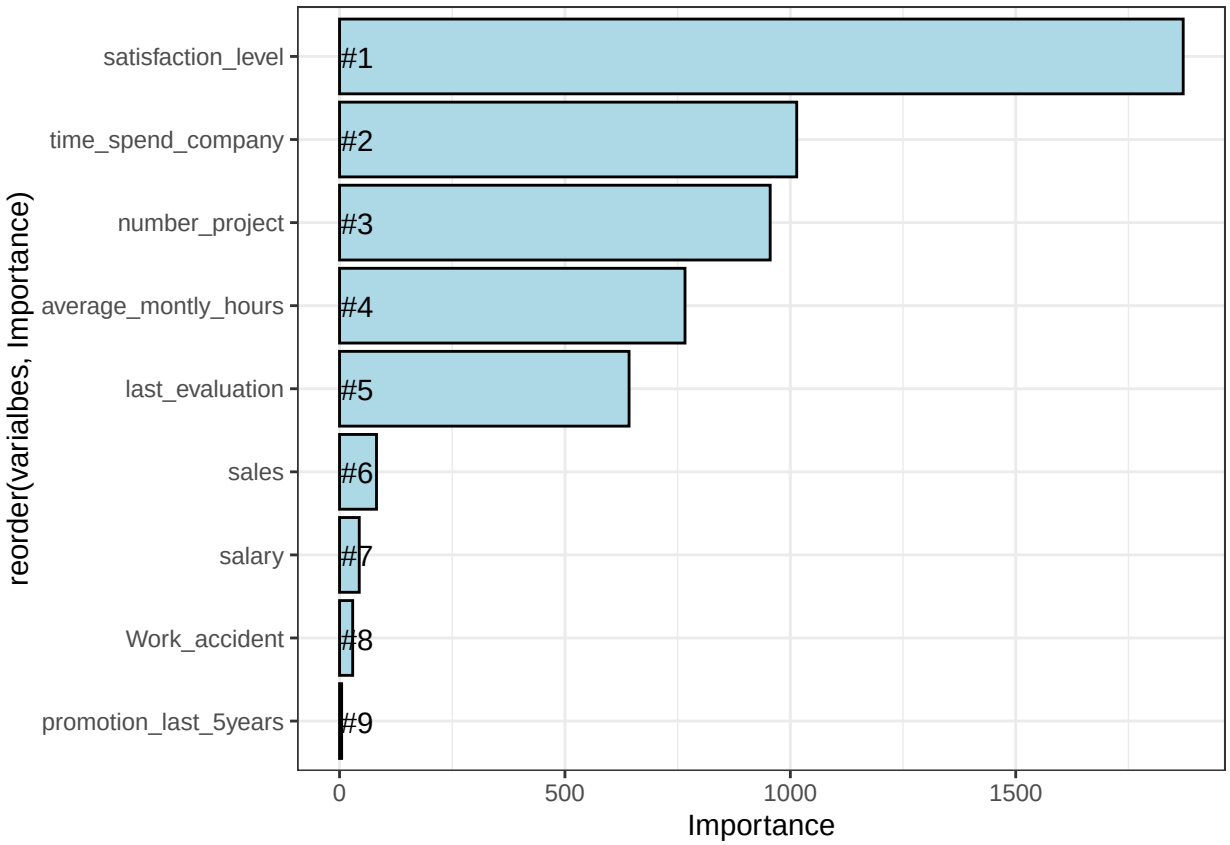
```
plot(left_model)
```



Importance rank

```
left_important <- importance(left_model)
varimportant <- data.frame(varialbes = row.names(left_important),
                           Importance = round(left_important[,'MeanDecreaseGini'],2))
```

```
varimportant %>%
   mutate(Rank = paste0("#",dense_rank(desc(left_important)))) %>%
   ggplot(aes(x = reorder(varialbes,Importance), y = Importance)) +
   theme_bw() +
   geom_col(color = "#000000",fill = "lightblue") +
     geom_text(aes(x = varialbes, y = 3, label = Rank),
    hjust=0, vjust=0.55, size = 4, colour = 'black') +
   coord_flip()
```




```
ind <- sample(2,nrow(d),replace = T,prob = c(0.7,0.3))
train <- d[ind == 1,]
test <- d[ind == 2,]
randomForest_model <- randomForest(left~.,data = train)
predicted_train <- predict(randomForest_model,newdata = train,type = "response")
Metrics::ce(train$left,predicted_train)
```

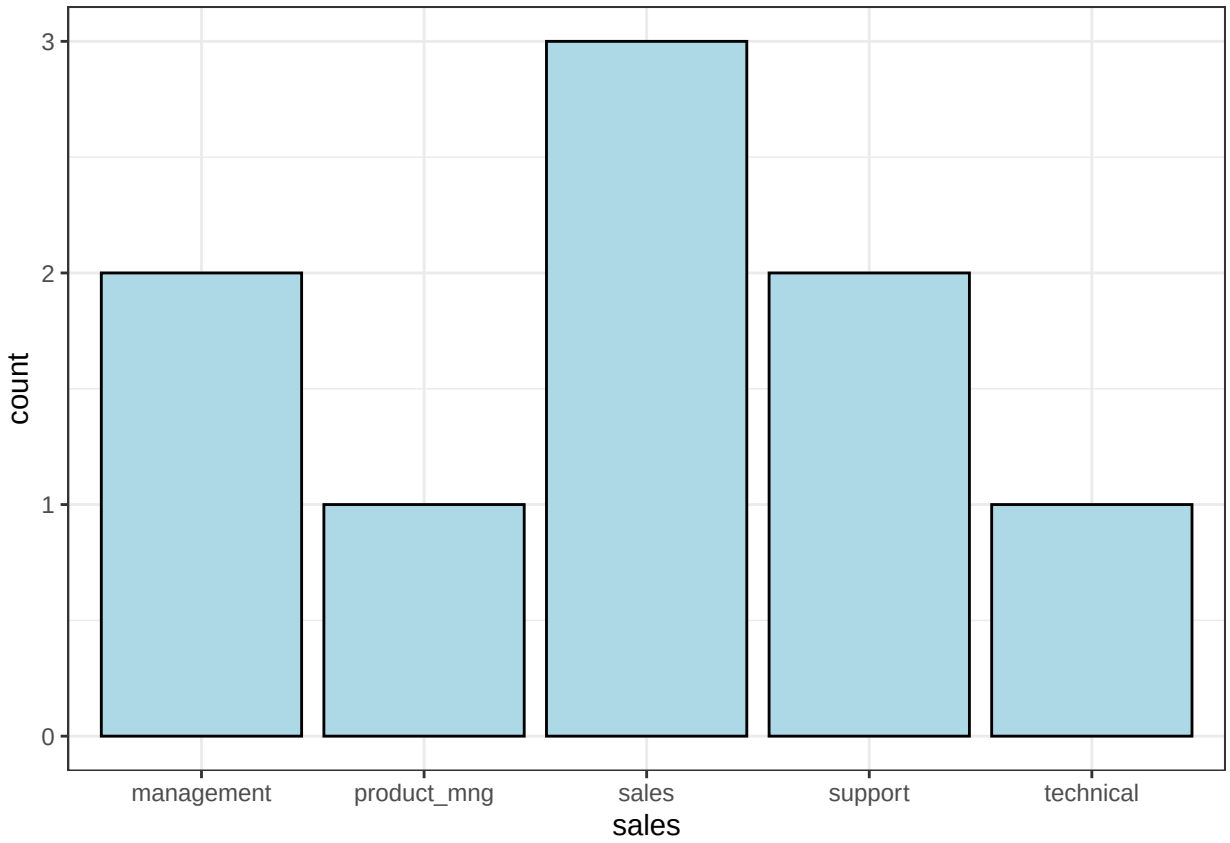
```
## [1] 9.48047e-05
```

```
predicted_test <- predict(randomForest_model,newdata = test,type = "response")
Metrics::ce(test$left,predicted_test)
```

```
## [1] 0.01258144
```

3.2 Which employees are potential quitters

```
d %>%
   mutate(predict_left = predict(randomForest_model,newdata = d,type = "response")) %>%
   filter(left == "0" & predict_left == "1") %>%
   ggplot(aes(sales,)) +
   theme_bw() +
   geom_bar(fill = "lightblue", color = "black")
```


3
2
1
0
count
management
product_mng
sales
support
technical
sales