R Notebook

Human Resources Analytics: Kaggle

Why are our best and most experienced employees leaving prematurely? Have fun with this database and try to predict which valuable employees will leave next. Fields in the dataset include:

Employee satisfaction level

Last evaluation

Number of projects

Average monthly hours

Time spent at the company

Whether they have had a work accident

Whether they have had a promotion in the last 5 years

Sales

Salary

Whether the employee has left

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(ggplot2)
library(ggvis)
```

```
##
## Attaching package: 'ggvis'

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

```r
library(corrplot)
library(DT)
library(readr)
suppressMessages(alldata <-read_csv("HR_comma_sep.csv"))
```

```r
dim(alldata)
```

```
## [1] 14999    10
```

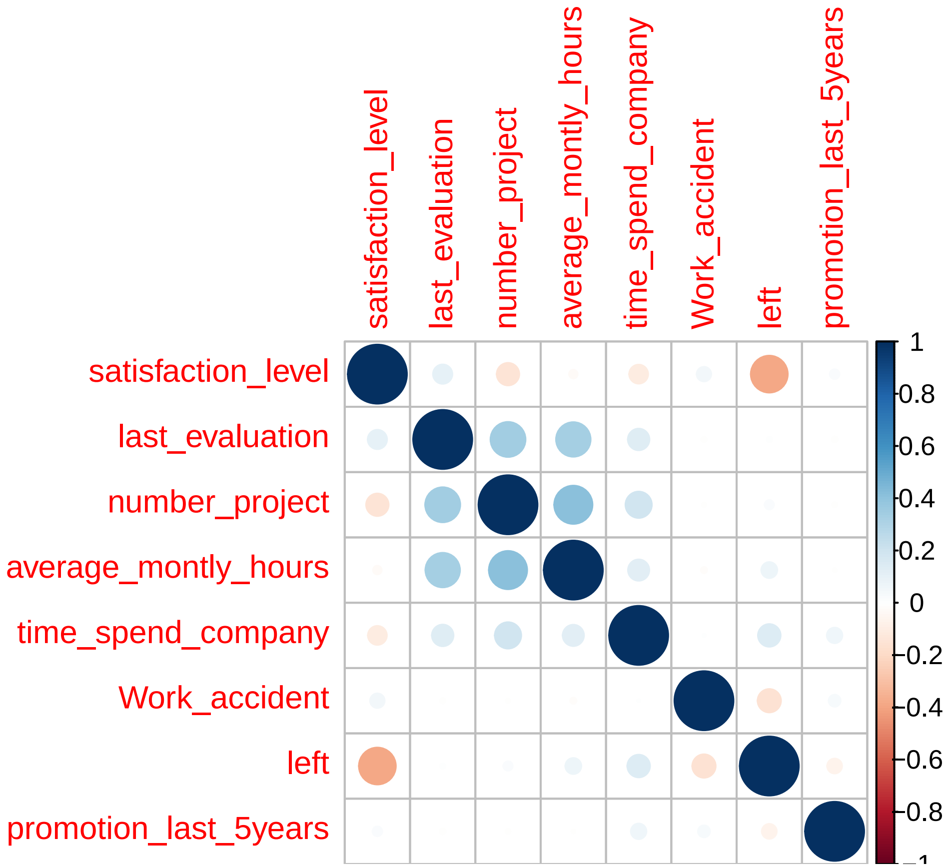
```r
summary(alldata)
```

```
##  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
##  Min.   : 2.000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 3.000     Median :0.0000   Median :0.0000
##  Mean   : 3.498     Mean   :0.1446   Mean   :0.2381
##  3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :10.000     Max.   :1.0000   Max.   :1.0000
##  promotion_last_5years    sales              salary
##  Min.   :0.00000       Length:14999       Length:14999
##  1st Qu.:0.00000       Class :character   Class :character
##  Median :0.00000       Mode  :character   Mode  :character
##  Mean   :0.02127
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

We see from the above summary that 24% of the employees in the data left the company. The average level of satisfaction is 62%, average number of projects worked on was 3.8, average monthly hours of work is 201, average time spent in the company is 3.5 years and the average number of promotions in the last 5 years has been 0.02.

Let us now look at the correlations between our variables:

```r
HR_correlation <- alldata %>% select(satisfaction_level:promotion_last_5years)
M <- cor(HR_correlation)
corrplot(M, method="circle")
```



We see that reported job satisfaction level has a significant inverse correlation with leaving. That is, those of low reported job satisfaction were likely to leave. We also see that there was small correlation between having spent much time in the company and not leaving. Those who have had a work accident were more likely to leave as well. Interestingly, having been promoted in the last 5 years did not correlation with likelihood of leaving.

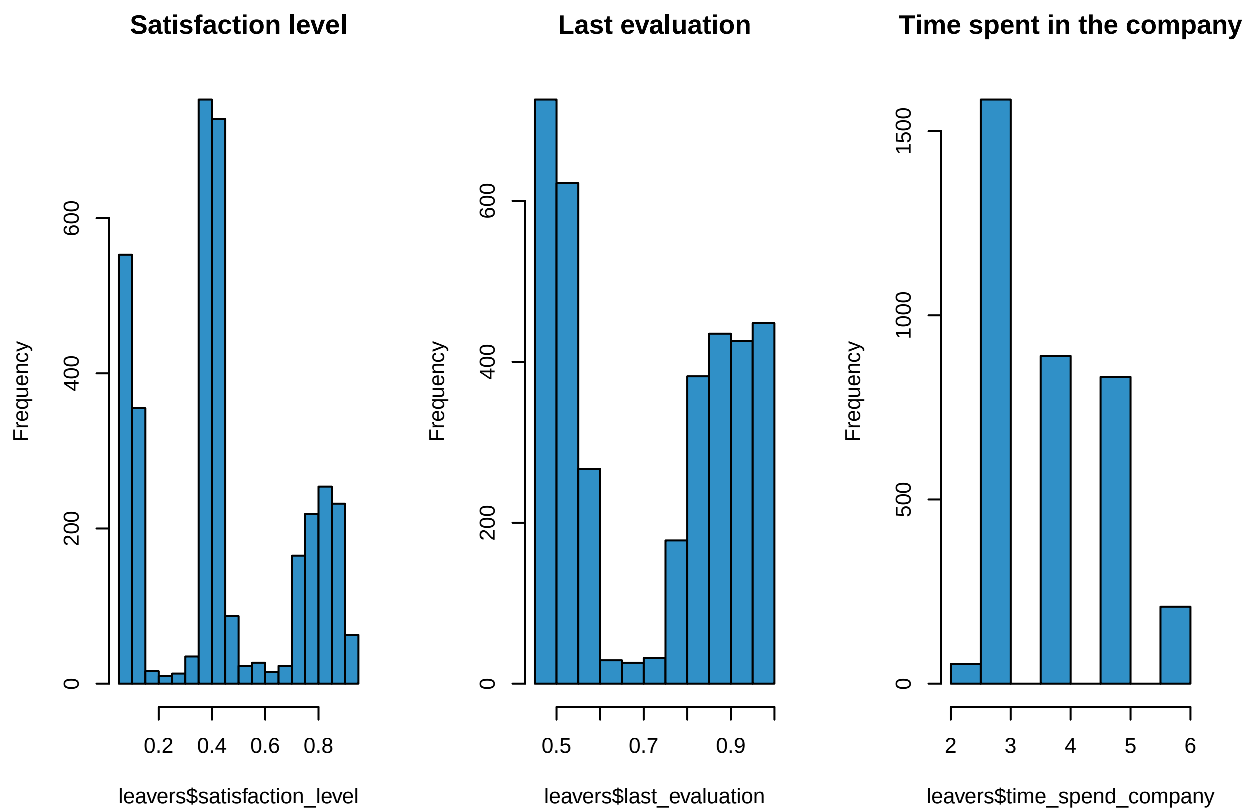
Let's consider now only those who leave:

```
leavers <- alldata %>% filter(left==1)
nrow(leavers)
```

```
## [1] 3571
```

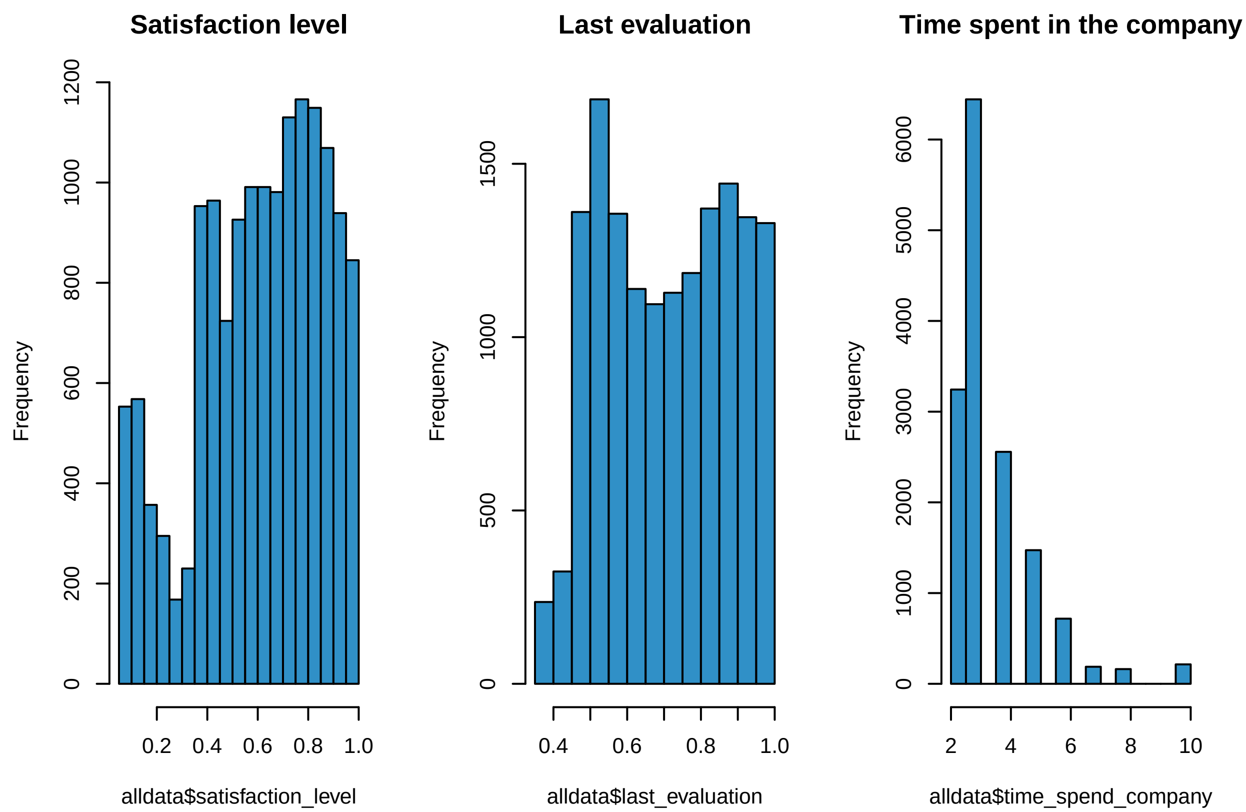
and let's see what features they had

```
par(mfrow=c(1,3))

hist(leavers$satisfaction_level,col="#3090C7", main = "Satisfaction level")

hist(leavers$last_evaluation,col="#3090C7", main = "Last evaluation")

hist(leavers$time_spend_company,col="#3090C7", main = "Time spent in the company")
```



Looking at the "Last evaluation" graph, we see an interesting distribution. It seems that those who who received a poor evaluation and those who scored highly on the evaluation were likely to leave. Those who were in-between, were likely to stay. However, we need to see whether the dstribution of evaluations was uniform. If there are very few lukewarm evaluations, then we can't conclude much.

```
par(mfrow=c(1,3))
hist(alldata$satisfaction_level,col="#3090C7", main = "Satisfaction level")
hist(alldata$last_evaluation,col="#3090C7", main = "Last evaluation")
hist(alldata$time_spend_company,col="#3090C7", main = "Time spent in the company")
```



Now we are much more confident in saying that "mediocre" employees will stay, while bad ones and good ones will leave. Of course we also see that those who leave are statistically less satisfied.

The number of leavers is

```
nrow(leavers)
```

```
## [1] 3571
```

Let's look at employees that should have been retained. These are the ones that either received a high evaluation or worked on many projects at once.

```
good_leavers <- leavers %>% filter(last_evaluation >= 0.75 | number_project >= 5)
nrow(good_leavers)
```

```
## [1] 1946
```

This turns out to be the majority of employees that left the company. So there is indeed a potential for improvement when it comes to retaining the desirable employees.

Next, we will build a predictive model for which employees will leave next.

I'll add a 0-1 column that specifies whether a worker is a "good leaver" or not and remove the column for left.

```
alldata$goodleft <- 1*((alldata$last_evaluation>=0.75 | alldata$number_project>=5) & alldata$left==1)
alldata$left <- NULL
head(alldata)
```

```
## # A tibble: 6 × 10
##   satisfaction_level last_evaluation number_project average_montly_hours
##                <dbl>           <dbl>          <int>                <int>
## 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
## # ... with 6 more variables: time_spend_company <int>,
## #   Work_accident <int>, promotion_last_5years <int>, sales <chr>,
## #   salary <chr>, goodleft <dbl>
```

```
library("caret")
```

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

```
split=0.80
trainIndex <- createDataPartition(alldata$goodleft, p=split, list=FALSE)
train <- alldata[ trainIndex,]
test <- alldata[-trainIndex,]
```

```
model <- glm (goodleft ~ ., data = train, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = goodleft ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7738  -0.1906  -0.0560  -0.0087   3.8511
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -26.037875   0.734143 -35.467  < 2e-16 ***
## satisfaction_level     -0.090993   0.176054  -0.517    0.605
## last_evaluation         9.789953   0.415831  23.543  < 2e-16 ***
## number_project          1.297506   0.054647  23.744  < 2e-16 ***
## average_montly_hours    0.027627   0.001302  21.212  < 2e-16 ***
## time_spend_company      0.530503   0.027665  19.176  < 2e-16 ***
## Work_accident          -1.379799   0.167534  -8.236  < 2e-16 ***
## promotion_last_5years  -2.651223   0.585944  -4.525 6.05e-06 ***
## saleshr                 0.213785   0.282381   0.757    0.449
## salesIT                -0.135591   0.244432  -0.555    0.579
## salesmanagement        -0.263808   0.312769  -0.843    0.399
## salesmarketing         -0.251492   0.270934  -0.928    0.353
## salesproduct_mng       -0.347942   0.261437  -1.331    0.183
## salesRandD             -0.120625   0.274493  -0.439    0.660
## salessales              0.016294   0.205452   0.079    0.937
## salessupport            0.172628   0.218115   0.791    0.429
## salestechnical          0.143267   0.210986   0.679    0.497
## salarylow               2.306289   0.267684   8.616  < 2e-16 ***
## salarymedium            1.770848   0.268213   6.602 4.05e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9223.7  on 11999  degrees of freedom
```

```
## Residual deviance: 3434.6  on 11981  degrees of freedom
## AIC: 3472.6
##
## Number of Fisher Scoring iterations: 8
```

```
predict <- predict(model, type = 'response')
confusion_train=table(train$goodleft, predict > 0.5)
# accuracy on test data
(confusion_train[1,1]+confusion_train[2,2])/nrow(train)
```

```
## [1] 0.951
```

```
prediction<-(predict.glm(model, test[,-10],type='response')>0.5)*1
# accuracy on test data
sum((prediction==test[,10])*1)/nrow(test)
```

```
## [1] 0.9486495
```

It looks like logistic regression does a very good job on the data. We can now predict which good workers are likely to leave. It is interesting to look at the coefficients in the regression to get a sense of what factors can predict that a worker is both good and likely to leave:

```
coefficients(model)
```

```
##           (Intercept)    satisfaction_level       last_evaluation
##          -26.03787504           -0.09099267            9.78995327
##        number_project  average_montly_hours    time_spend_company
##            1.29750646            0.02762679            0.53050279
##         Work_accident promotion_last_5years               saleshr
##           -1.37979941           -2.65122325            0.21378543
##               salesIT       salesmanagement        salesmarketing
##           -0.13559118           -0.26380795           -0.25149231
##      salesproduct_mng            salesRandD            salessales
##           -0.34794181           -0.12062506            0.01629388
##          salessupport        salestechnical             salarylow
##            0.17262831            0.14326691            2.30628949
##          salarymedium
##            1.77084845
```