Employee Retention

Mitul Shah

8/12/2017

Load the data

```
# read data set
data = read.csv("employee_retention_data.csv")

# check the structure
str(data)
```

```
## 'data.frame':    24702 obs. of  7 variables:
##  $ employee_id: num  13021 825355 927315 662910 256971 ...
##  $ company_id : int  7 7 4 7 2 4 4 2 9 1 ...
##  $ dept       : Factor w/ 6 levels "customer_service",..: 1 5 5 1 2 2 1 1 4 6 ...
##  $ seniority  : int  28 20 14 20 23 14 21 4 7 7 ...
##  $ salary     : num  89000 183000 101000 115000 276000 165000 107000 30000 160000 104000 ...
##  $ join_date  : Factor w/ 995 levels "2011-01-24","2011-01-25",..: 643 459 758 264 148 205 558 633 38
##  $ quit_date  : Factor w/ 664 levels "2011-10-13","2011-10-14",..: 643 364 NA 229 428 267 NA NA 640 N
```

```
data$company_id = as.factor(data$company_id) # this is a categorical var
data$join_date = as.Date(data$join_date) #make it a date
```

```
## Warning in strptime(xx, f <- "%Y-%m-%d", tz = "GMT"): unknown timezone
## 'zone/tz/2018c.1.0/zoneinfo/America/New_York'
```

```
data$quit_date = as.Date(data$quit_date) #make it a date

summary(data)
```

```
##   employee_id       company_id                 dept        seniority
##  Min.   :    36   1      :8486   customer_service:9180   Min.   : 1.00
##  1st Qu.:250134   2      :4222   data_science    :3190   1st Qu.: 7.00
##  Median :500793   3      :2749   design          :1380   Median :14.00
##  Mean   :501604   4      :2062   engineer        :4613   Mean   :14.13
##  3rd Qu.:753137   5      :1755   marketing       :3167   3rd Qu.:21.00
##  Max.   :999969   6      :1291   sales           :3172   Max.   :99.00
##                   (Other):4137
##      salary         join_date            quit_date
##  Min.   : 17000   Min.   :2011-01-24   Min.   :2011-10-13
##  1st Qu.: 79000   1st Qu.:2012-04-09   1st Qu.:2013-06-28
##  Median :123000   Median :2013-06-24   Median :2014-06-20
##  Mean   :138183   Mean   :2013-06-29   Mean   :2014-05-02
##  3rd Qu.:187000   3rd Qu.:2014-09-17   3rd Qu.:2015-03-27
##  Max.   :408000   Max.   :2015-12-10   Max.   :2015-12-09
##                                        NA's   :11192
```

Create a table with 3 columns: day, company_id, employee headcount

Let's answer this question: You should create a table with 3 columns: day, employee headcount, company id.

```r
# libraries needed
require(dplyr)
```

```
## Loading required package: 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
require(rpart)
```

```
## Loading required package: rpart
```

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
require(scales)
```

```
## Loading required package: scales
```

```r
unique_dates = seq(as.Date("2011/01/24"), as.Date("2015/12/13"), by = "day")

unique_companies = unique(data$company_id)

data_headcount = merge(unique_dates, unique_companies, by = NULL)

colnames(data_headcount) = c("date", "company_id")

data_join = data %>%
            group_by(join_date, company_id) %>%
            summarise(join_count = length(join_date))

## This is also correct:
# data_join = data %>% group_by(join_date, company_id) %>% summarise(join_count = n_distinct(employee_i

## data_quit = data %>% group_by(quit_date, company_id) %>% summarise(quit_count = n_distinct(employee_
data_quit = data %>%
            group_by(quit_date, company_id) %>%
            summarise(quit_count = length(quit_date))


data_headcount = merge (data_headcount, data_join,
                        by.x = c("date", "company_id"),
                        by.y = c("join_date", "company_id"),
                        all.x = TRUE)


data_headcount = merge (data_headcount, data_quit,
                        by.x = c("date", "company_id"),
                        by.y = c("quit_date", "company_id"),
                        all.x = TRUE)


data_headcount$join_count[is.na(data_headcount$join_count)] = 0
data_headcount$quit_count[is.na(data_headcount$quit_count)] = 0


data_headcount = data_headcount %>%
            group_by(company_id) %>%
            mutate(join_cumsum = cumsum(join_count),
                     quit_cumsum = cumsum(quit_count))

data_headcount$count = data_headcount$join_cumsum - data_headcount$quit_cumsum

data_headcount_table = data.frame(data_headcount[, c("date", "company_id","count")])
```

```r
## Another way

loop_cumsum = c() #intialize empty vector
loop_date = c()
loop_company = c()


for (i in seq(as.Date("2011/01/24"), as.Date("2015/12/13"), by = "day")) {
   for (j in unique(data$company_id)){ # loop through all companies
        tmp_join = nrow(subset(data, join_date <= i & company_id == j))
        tmp_quit = nrow(subset(data, quit_date <= i & company_id == j))
        loop_cumsum = c(loop_cumsum, tmp_join - tmp_quit )
        loop_date = c(loop_date, i)
        loop_company = c(loop_company, j)
   }
data_headcount_table_loop = data.frame(date = as.Date(loop_date, origin = '1970-01-01'), company_id = l
}



identical(data_headcount_table[order(data_headcount_table[,1],                as.numeric(as.character(da
```

```
## [1] TRUE
```

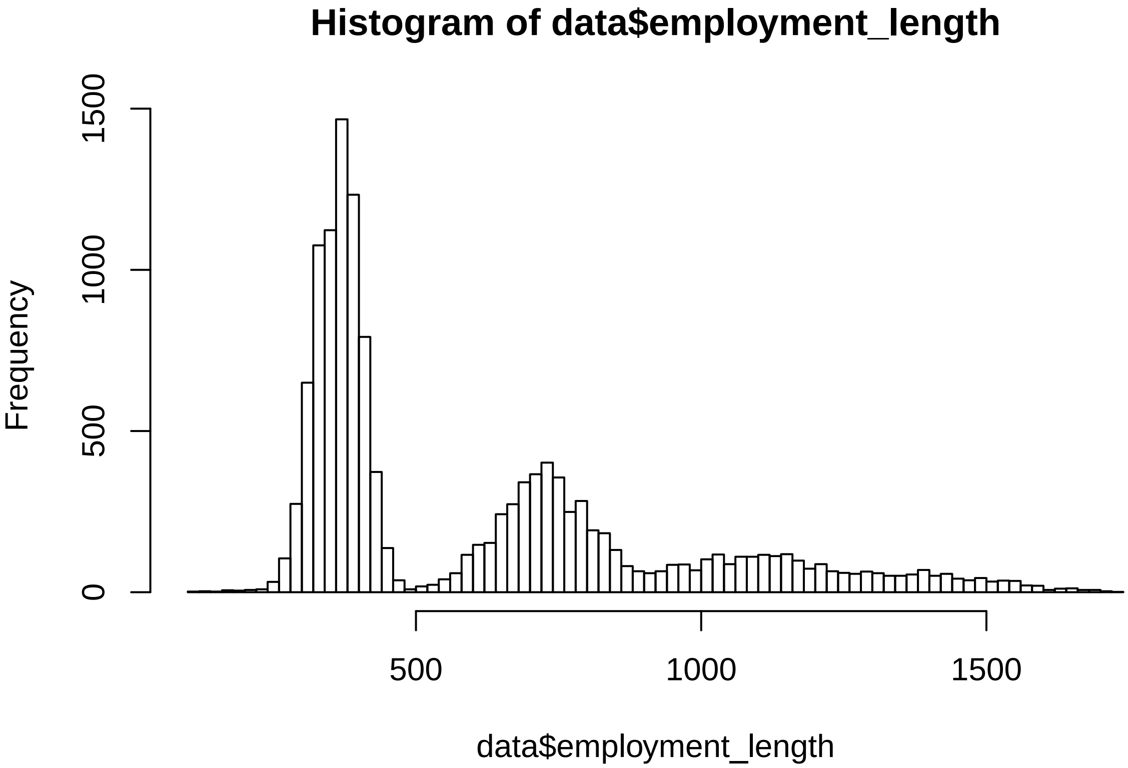
Main factors driving Employee Retention

Now, let's try to understand employee retention. Here, the main challenge is about feature engineering. That is, extract variables from the quitting date column.

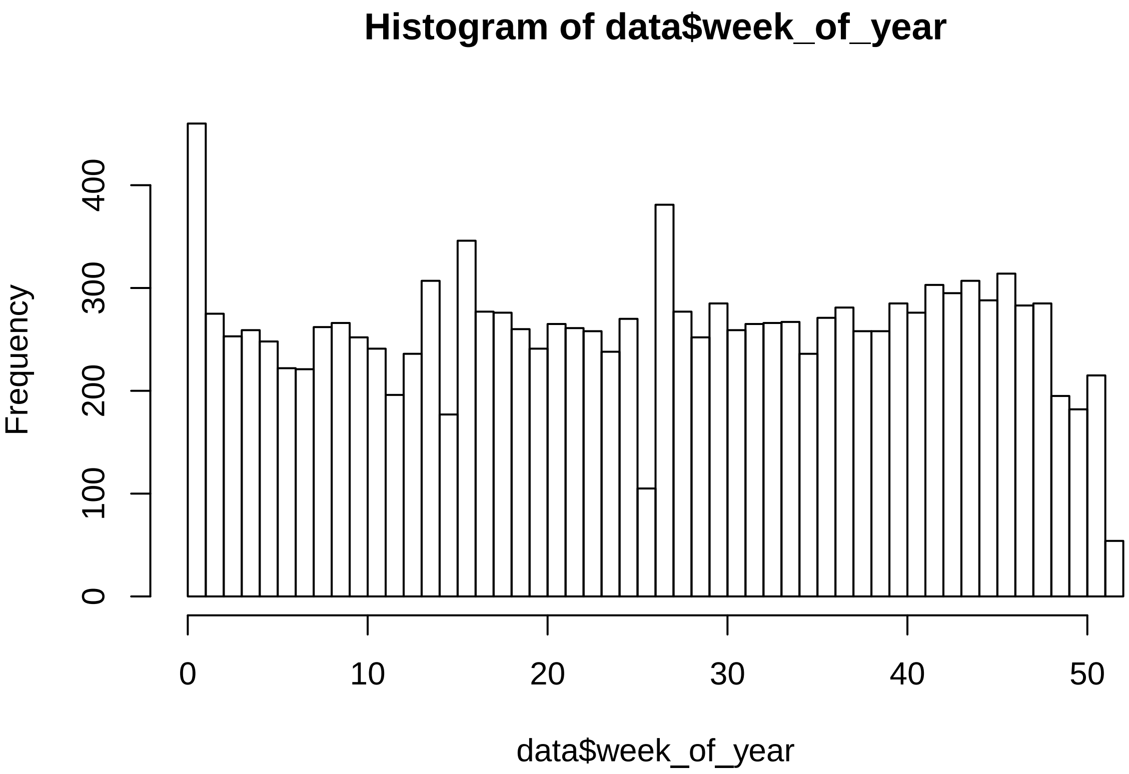
```
data$employment_length = as.numeric(data$quit_date - data$join_date)

data$week_of_year = as.numeric(format(data$quit_date, "%U"))

hist(data$employment_length, breaks = 100)
```



```
hist(data$week_of_year, breaks = length(unique(data$week_of_year)))
```



Very interesting, there are peaks around each employee year anniversary! And it also peaks around the new year. Makes sense, companies have much more budget to hire at the beginning of the year.

Now, let's see if we can find the characteristics of people who quit early. Looking at the class employement_length, it looks like we could define early quitters as those people who quit within 1 year or so. So, let's create two classes of users: quit within 13 months or not (if they haven't been in the company for atleast 13 months, we remove them)

```
## Only keep people who had enough time to age
data = subset(data, data$join_date < as.Date("2015/12/13") - (365 + 31))

## Early Quitters column
data$early_quitter = as.factor(ifelse(is.na(data$quit_date) | as.numeric(data$quit_date - data$join_dat
```

```
tree = rpart(early_quitter ~ .,data[, c("company_id", "dept", "seniority", "early_quitter", "salary")],

tree
```

```
## n= 19270
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 19270 9635.0000 0 (0.5000000 0.5000000)
##   2) salary>=224500 2764  855.3351 0 (0.6528040 0.3471960) *
##   3) salary< 224500 16506 8026.7840 1 (0.4776014 0.5223986)
##     6) salary< 62500 2887 1249.7210 0 (0.5498859 0.4501141) *
##     7) salary>=62500 13619 6500.0510 1 (0.4632968 0.5367032) *
```

Not very surprising! Salary matters the most. After all, it probably has the information about other variables too within itself like seniority, dept, etc.

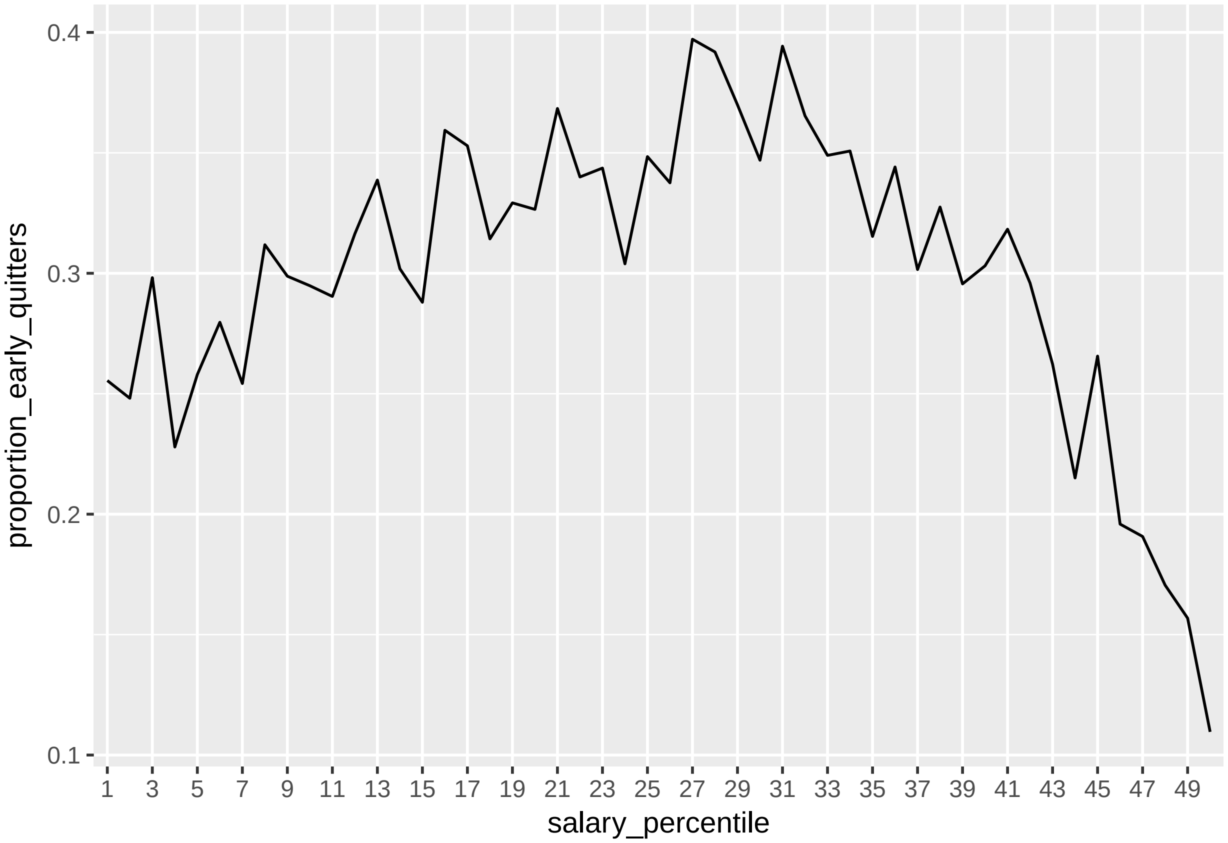
It is interesting though that people with salary betweem 62500 and 224500 have higher probability of being an early quitter. That means the people with very high salary and people who make very little are not likely to quit early.

By plotting the proportion of early quitter by salary percentile, this becomes quite clear.

```
data$salary_percentile = cut(data$salary, breaks = quantile(data$salary, probs = seq(0, 1, 0.02)), inclu

data_proportion_by_percentile = data %>%
                                  group_by(salary_percentile) %>%
                                  summarize(proportion_early_quitters = length(early_quitter[early_quitt
                                            )

qplot(salary_percentile, proportion_early_quitters, data=data_proportion_by_percentile, geom="line", gr
```



Conclusion

1. Given how important is salary, I would definitely love to have as a variable the salary the employee was offered in the nect job. Otherwise, things like promotions or raises received during the employee tenure would be interesting.

2. The major findings are that the employees quit at year anniversaries or at the beginning of the year. Both cases make sense. Even if you don't like your current job, you often stay for 1 year before quitting plus you often get stocks after 1 year so it makes sense to wait. Also, the beginning of the year is

well-known to be the best time to change job: companies are hiring more and you often want to stay until the end of Dec to get the calender year bonus.

3. Employees with low and high salaries are less likely to quit. Probably because employees with high salaries are happy there and employees with low salaries are not that marketable, so they have a hard time finding a new job.