Identification of Employee Resignation Predictors

Mikhail Lara

Load Data & Packages

```
library(data.table)
library(ggplot2)
library(randomForest)
library(gmodels)
library(gridExtra)
library(grid)
library(caret)
library(vcd)
library(corrplot)

setwd('/Users/Mikey/Documents/ML-Case-Studies/Human Resources Analytics')
file_name<-'HR_comma_sep.csv'

hr<-read.csv(file_name)
hr_dat<-as.data.table(hr)
```

Data Summary & EDA

- The data set includes 5 categorical variables that describe each employee's resignation status, promotion status, salary level and department within the company.
  - Whether the Employee Experienced a Workplace Accident (0: False , 1: True)
  - Whether the Employee Received a Promotion in the Last 5 Years (0: False , 1: True)
  - Department{sales}
  - Salary Bracket{salary} (low, medium, high)
  - Resignation Status{left} (0: Remained with Company , 1: Resigned)

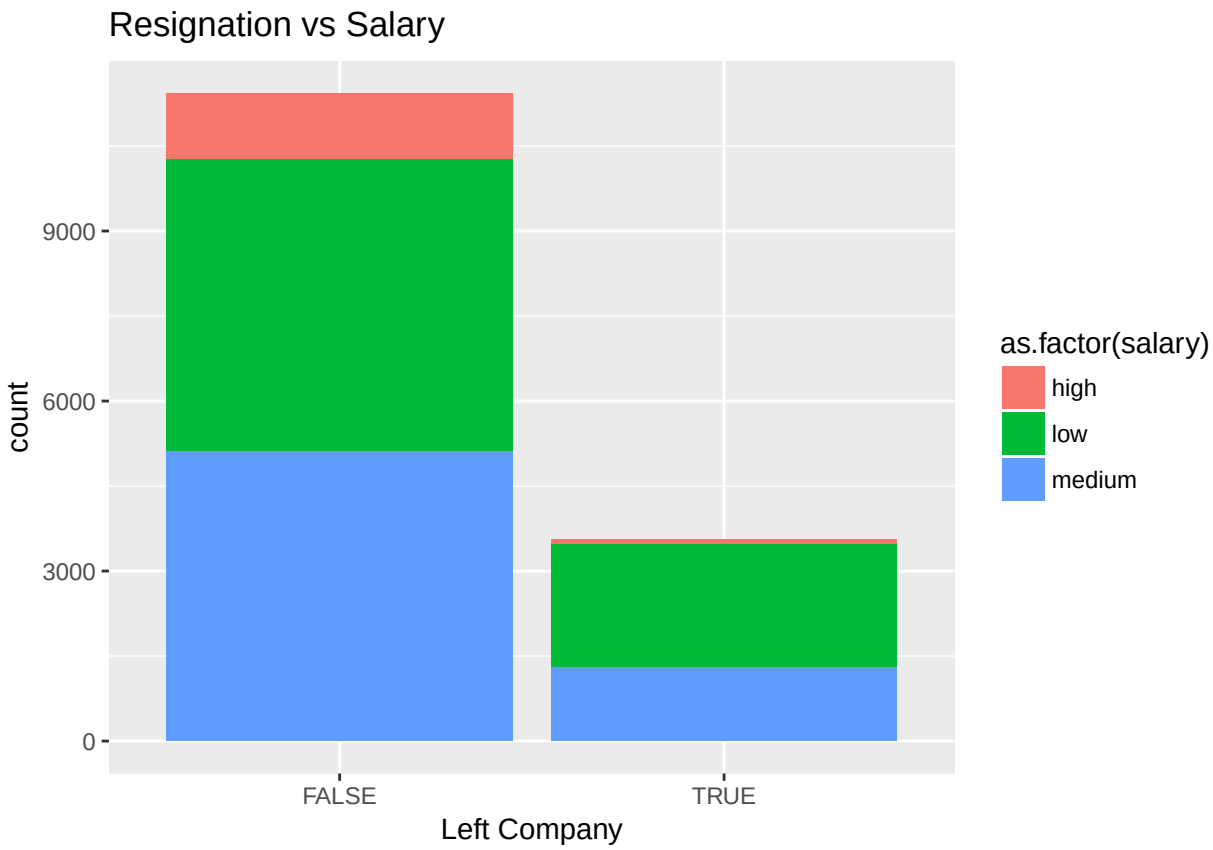
Correlation Plot

- Pairs plot indicates that 'Left' has >10% Correlation with:
  - Satisfaction Level (-38.8%)
  - Work Accident (-15.5%)
  - Time Spend at Company (14.5%)
- The Average Monthly Hours data has the highest correlation with 'Left' that is less than 10% (7.13%)
- Of these 3 variables, the company can exert the greatest control over the satisfaction level by modifying the workplace environment and setting company policies. Focus on characterizing this before moving on to work accident experience and the time spent at the company.

Assess Gross Population Effect

- A Chi-squared test for independence indicates that the the relative proportions of employees that resigned does not match those of the salary distribution of all employees reported ($p<0.05$). This implies that each employee's salary mey be associated with his, or her, decision to resign from the company. Although this seems obvious, it is important because it rules out the possibility that employee resignation cannot be decreased by increasing salaries and bonuses.
  - Reported Total Salary Distribution:
    - High(8.2%) , Medium(43.0%) , Low(48.8%)
  - Reported Resignation-Salary Distribution:
    - High(2.3%) , Medium(36.9%) , Low(60.8%)

```
ggplot(data=hr_dat,aes(x=left,fill=as.factor(salary)))+geom_bar()+
    xlab('Left Company')+labs(title='Resignation vs Salary')
```




```
CrossTable(hr_dat$left,hr_dat$salary)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
```

```
##
## Total Observations in Table:  14999
##
##
##              | hr_dat$salary
##  hr_dat$left |      high |       low |    medium | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##        FALSE |      1155 |      5144 |      5129 |     11428 |
##              |    47.915 |    33.200 |     9.648 |           |
##              |     0.101 |     0.450 |     0.449 |     0.762 |
##              |     0.934 |     0.703 |     0.796 |           |
##              |     0.077 |     0.343 |     0.342 |           |
## -------------|-----------|-----------|-----------|-----------|
##         TRUE |        82 |      2172 |      1317 |      3571 |
##              |   153.339 |   106.247 |    30.876 |           |
##              |     0.023 |     0.608 |     0.369 |     0.238 |
##              |     0.066 |     0.297 |     0.204 |           |
##              |     0.005 |     0.145 |     0.088 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |      1237 |      7316 |      6446 |     14999 |
##              |     0.082 |     0.488 |     0.430 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

```
resign_tot<-data.table(rbind(c(2172,1317,82),c(7316,6446,1237)))
setnames(resign_tot,c('Low Salary','Medium Salary','High Salary'))
rownames(resign_tot)<-c('Resigned','Total')
print.data.frame(resign_tot)
```

```
##          Low Salary Medium Salary High Salary
## Resigned       2172          1317          82
## Total          7316          6446        1237
```

```
chisq.test(resign_tot)
```

```
##
##  Pearson's Chi-squared test
##
## data:  resign_tot
## X-squared = 251.37, df = 2, p-value < 2.2e-16
```

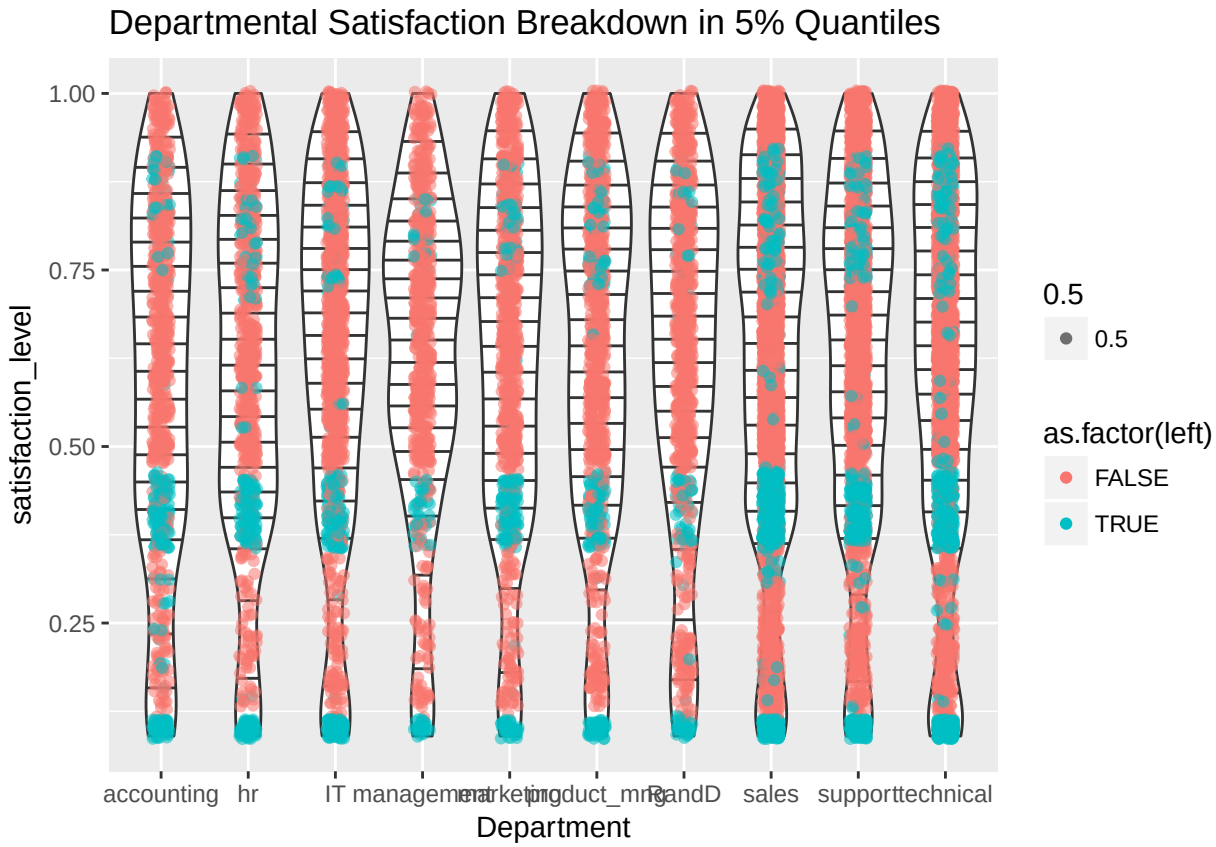
Employee Satisfaction Level

- Two Populations clusters appear to be associated with employees resigning:
  - Cluster 1: $0.101 <$ Satisfaction Level < 0.108 & Satisfaction Level is in the 5% departmental quantile
  - Cluster 2: $0.403 <$ Satisfaction Level < 0.409 & Satisfaction Level is in 15%-30% departmental quantiles
  - A much smaller cluster employees also resigned despite having a Satisfaction Level greater than 0.75

```
sales.count<-hr_dat[,.N,by=sales]
setorder(sales.count,N)
```

```
s.sales<-ggplot(hr_dat, aes(x=as.factor(sales),y=satisfaction_level))+
    geom_violin(draw_quantiles=seq(from=0.05,to=1,by=.05))+
    geom_jitter(width=0.1,aes(colour=as.factor(left),alpha=0.5))+
    xlab('Department')+
    ggtitle('Departmental Satisfaction Breakdown in 5% Quantiles')

s.sales
```



Employee Satisfaction Level: Cluster 1

- The mean departmental satisfaction levels in Cluster 1 are skewed because of the accounting department with a value of 0.116. Hypothesis testing indicates that it's satisfaction level is significantly greater than other departments, implying that the unhappiest employees who resigned in accounting enjoyed their jobs more than unhappiest employees who resigned in other departments.

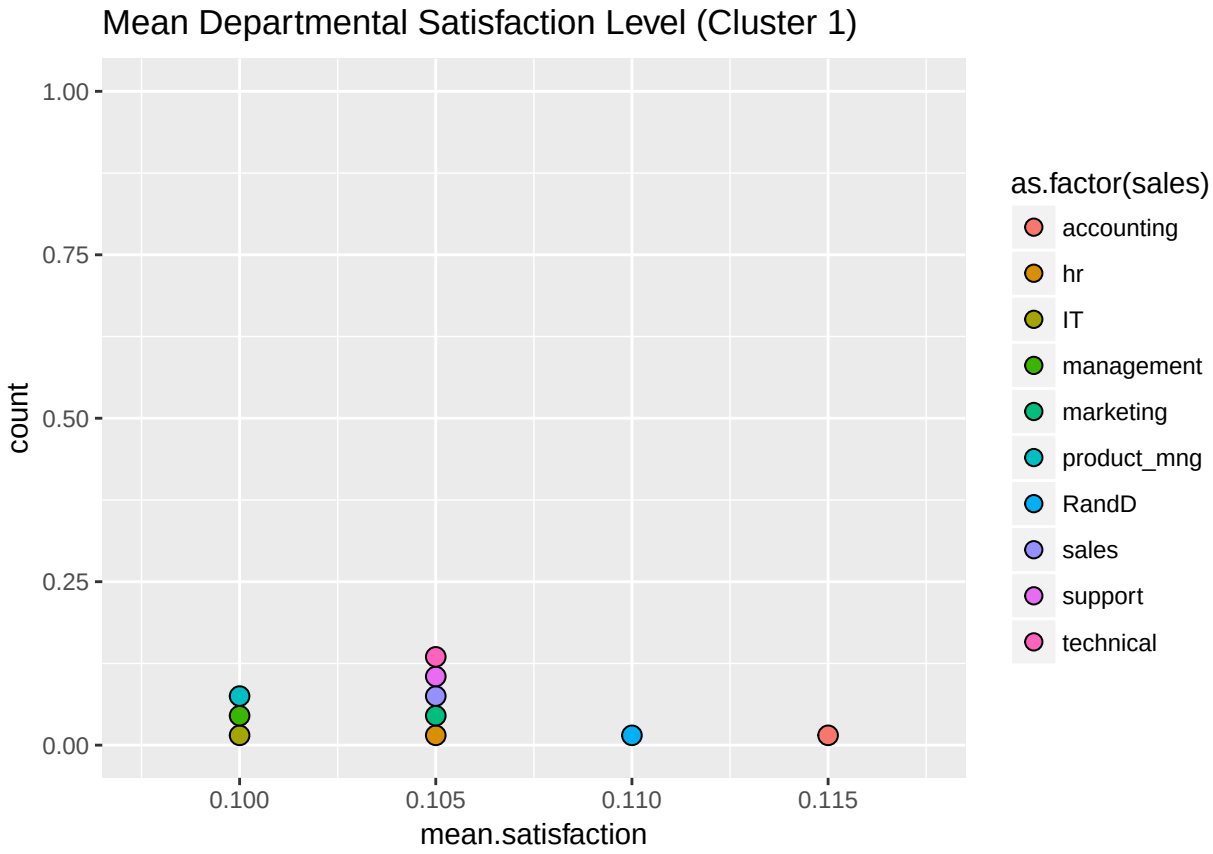
```
##            sales satisfaction_level
##  1:  accounting          0.1169841
##  2:       RandD          0.1092308
##  3:       sales          0.1049583
##  4:   marketing          0.1047619
##  5:     support          0.1045985
##  6:          hr          0.1026000
##  7:   technical          0.1025980
##  8:          IT          0.1023864
##  9:  management          0.1021429
## 10: product_mng          0.1000000
```

```
##
##  Shapiro-Wilk normality test
##
## data:  temp$mean.satisfaction
## W = 0.78917, p-value = 0.01071
```

```
## [1] "Cluster 1 Mean Satisfaction Confidence Interval"
```

```
## [1] 0.1085083 0.1015439
```

```
ggplot(data=temp,aes(x=mean.satisfaction,fill=as.factor(sales)))+
    geom_dotplot(binaxis = "x", stackgroups = TRUE, binwidth = .005,dotsize = 0.1 ,method = "histodot")+
    ggtitle('Mean Departmental Satisfaction Level (Cluster 1)')
```



Employee Satisfaction Level: Cluster 2

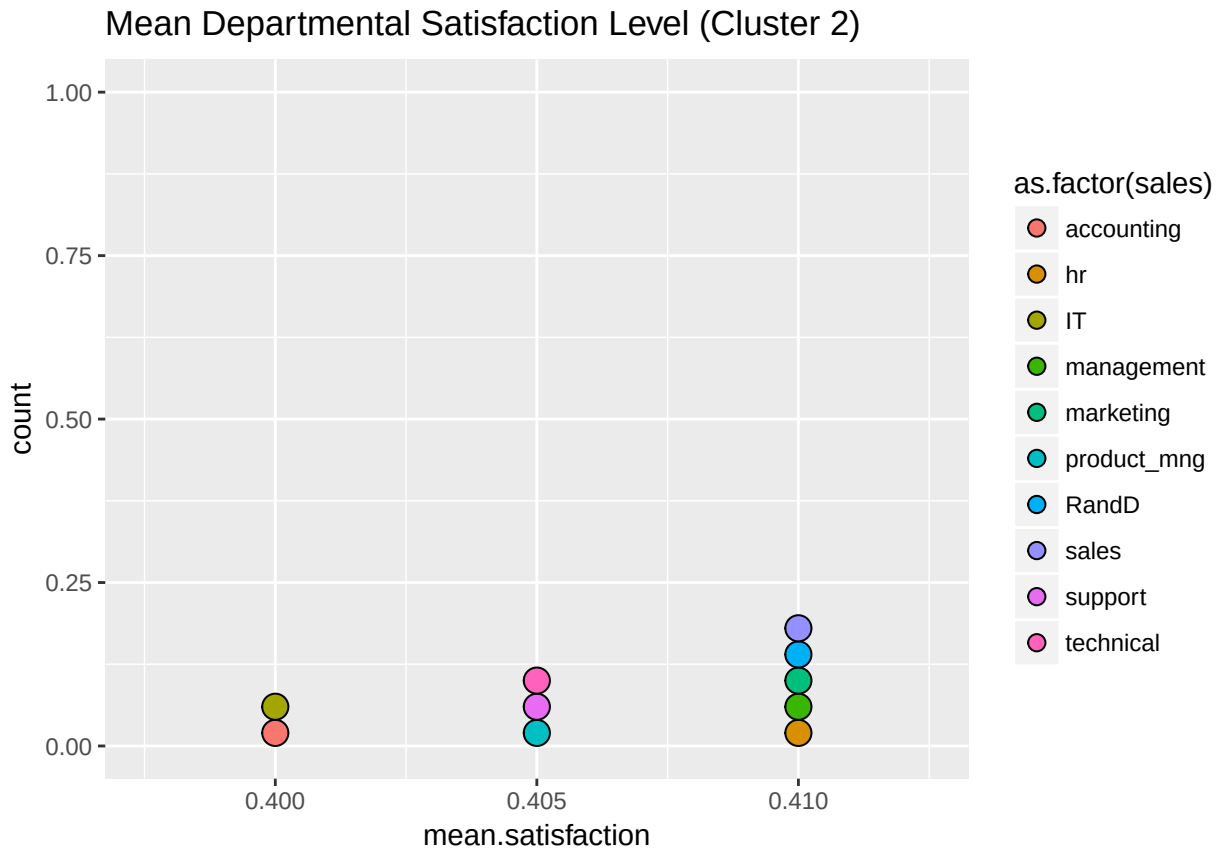
- The salary-resignation distribution exhibits the similar clustering as the breakdown by departments at low and medium salaries. However, the clusters are not as pronounced for high salaried employees. It is not surprising that the locations of the clusters are approximately equal across all salary brackets. This suggests that dissatisfaction that results in resignation is an intrinsic feature of the company that is not influenced by career paths or incentives.

```
##
##  Shapiro-Wilk normality test
##
## data:  temp$mean.satisfaction
## W = 0.91132, p-value = 0.2902
```

```
## [1] "Cluster 2 Mean Satisfaction Confidence Interval"
```

```
## [1] 0.4091665 0.4034531
```

```
ggplot(data=temp,aes(x=mean.satisfaction,fill=as.factor(sales)))+
    geom_dotplot(binaxis = "x", stackgroups = TRUE, binwidth = .005,dotsize = 0.1 ,method = "histodot")+
    ggtitle('Mean Departmental Satisfaction Level (Cluster 2)')
```



- From the previous cross-tabulation, it is known that 6.6% of high salaried employess resigned whereas 29%.7 and 20.4% of low and medium salary employees resigned respectively. The disparity supports the hypothesis that salary is correlated to an employee's decision to resign.

```
s.salary<-ggplot(hr_dat, aes(x=as.factor(salary),y=satisfaction_level))+
    geom_violin(draw_quantiles=c(0.25,0.5,0.75))+
    geom_jitter(width=0.1,aes(colour=as.factor(left),alpha=0.5))+
    xlab('Salary Level')+
    ggtitle('Salary Satisfaction Breakdown in 25% Quantiles')

s.salary
```

- ANOVA on 'satisfaction level', excluding left, indicates a statistically significant dependence all fields except 'average monthly hours' ($p<.05$).
  - Consider omitting average monthly hours from ML model for left since it is poorly correlated with 'left' and isn't a significant predictor of satisfaction.

```
summary(aov(data=hr_dat,satisfaction_level~.-left))
```

```
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## last_evaluation           1   10.2   10.23 176.497  < 2e-16 ***
## number_project            1   34.1   34.08 588.291  < 2e-16 ***
## average_montly_hours      1    0.0    0.03   0.587 0.443636
## time_spend_company        1    6.9    6.85 118.290  < 2e-16 ***
## Work_accident             1    3.3    3.27  56.387 6.29e-14 ***
## promotion_last_5years     1    0.8    0.82  14.171 0.000168 ***
## sales                     9    1.3    0.14   2.484 0.007861 **
## salary                    2    2.6    1.30  22.363 2.01e-10 ***
## Residuals             14981  868.0    0.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

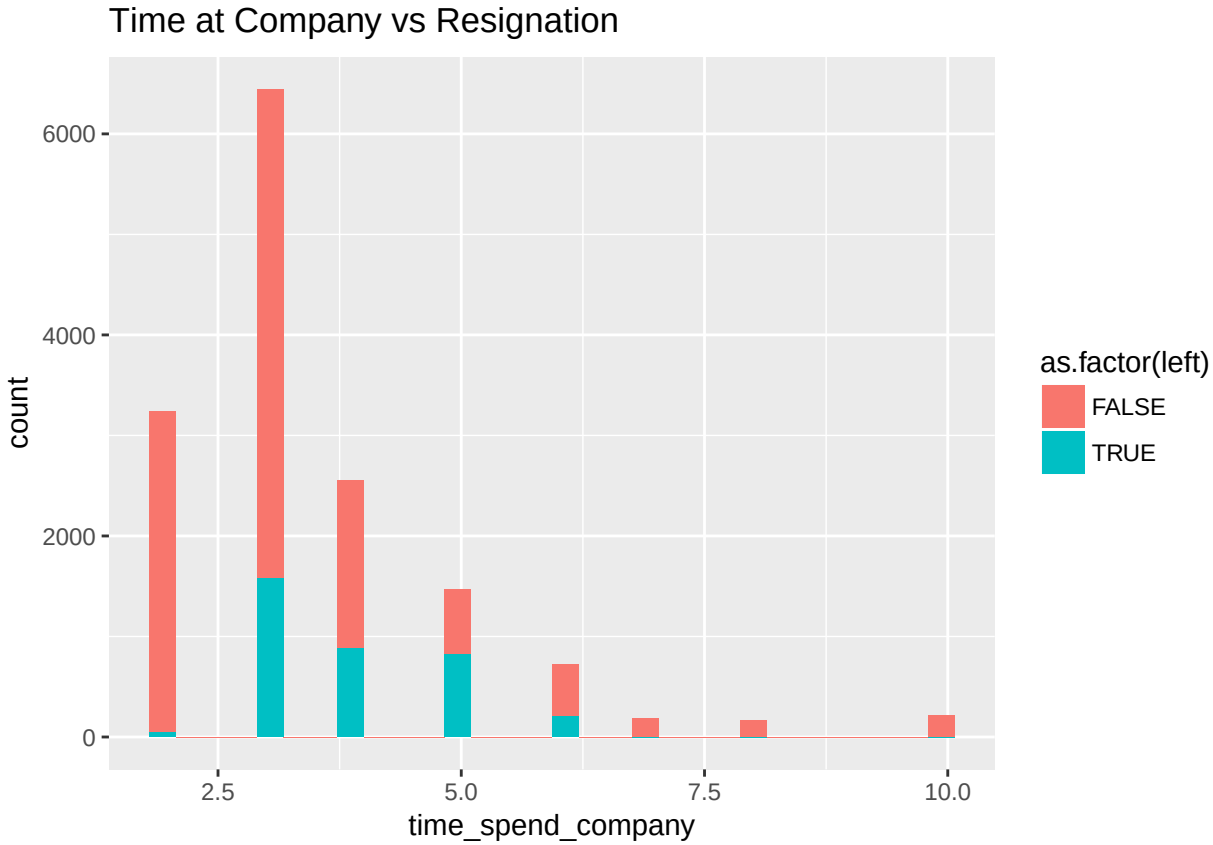
Time Spent at the Company

- Of all the employees that resigned, 98.5% did so in the first 6 years of employment with 92.6% of those employees leaving between the 3rd and 5th years.
- Very few employees resigned after only two years. This trend in employee retention can likely be explained by people fresh out of school who are just starting their careers and have an incentive to

remain employed to gain professional experience.

- After 6 years of employment at the company, the resignation rate drops to zero. This trend suggests the existence of a critical threshold beyond which HR no longer has to actively monitor employees to prevent resignation. Based on this, HR should consider providing promotions, raises, and company-sponsored training more frequently during the first 6 years of employment to get more employees to the 7 year mark where they are significantly less likely to resign.

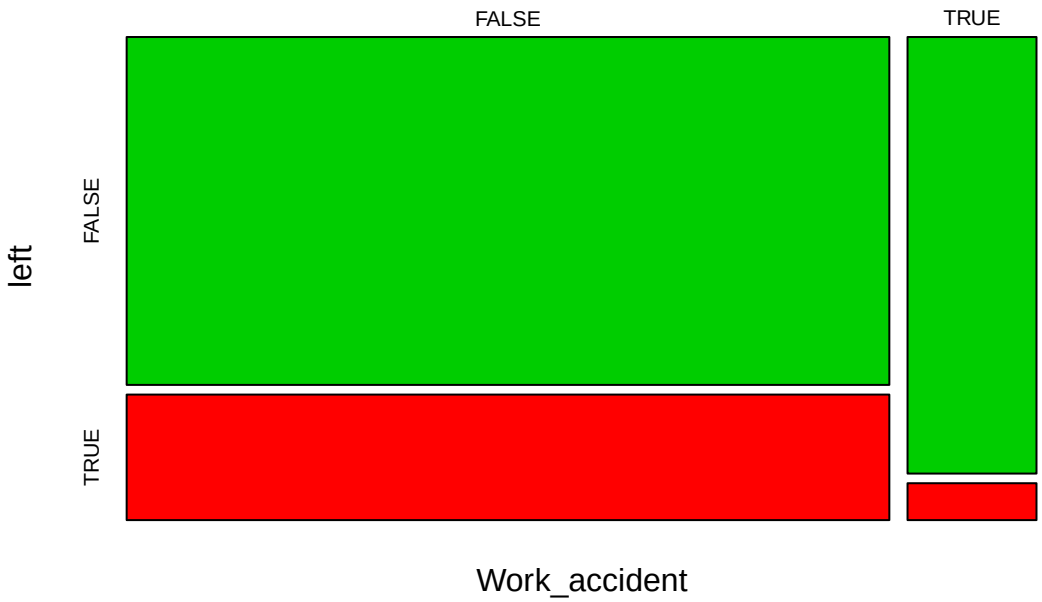
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Work Accident Experience

- The relationship between an employee experiencing a work accident and choosing to resign is unintuitive. Cross-tabulation shown in the mosaic plot, employee resignation is negatively correlated with experiencing an accident at work. This implies that employees who got injured on the job were more likely to choose to continue working at the company. Taken by itself, this result does not make sense. It may be explained by an interaction with either one of the reported data or one that is not present in the dataset.

```
mosaicplot(data=hr_dat,Work_accident~left,color=c(3,2),main= 'Dependence of Resignation on Work Acciden
```



Random Forest Classification & Variable Importance

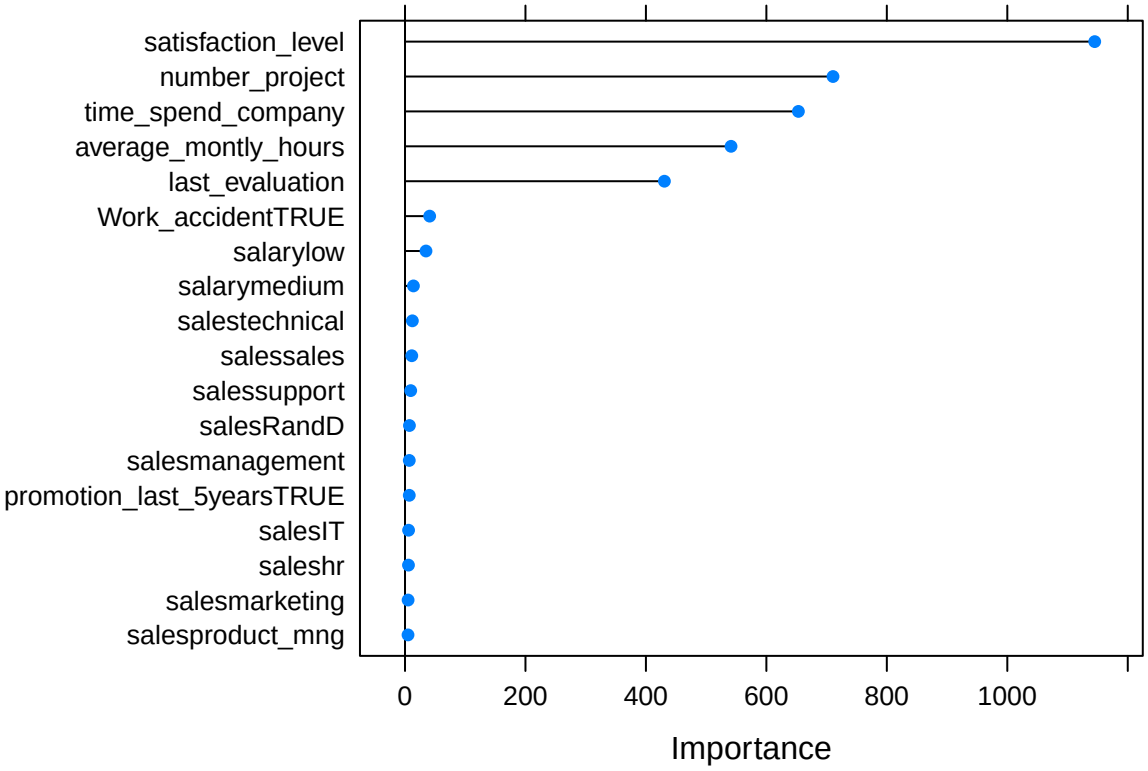
- The human resources data was split into training and validation data sets. The training data used to develop the random forest was comprised 75% of the original data set.

```
set.seed(7)
inTrain = createDataPartition(hr_dat$left,p=3/4,list=FALSE)
train_dat = hr_dat[inTrain,]
validate = hr_dat[-inTrain,]

control <- trainControl(method="repeatedcv", number=10, repeats=3)
seed <- 7
metric <- "Accuracy"
```

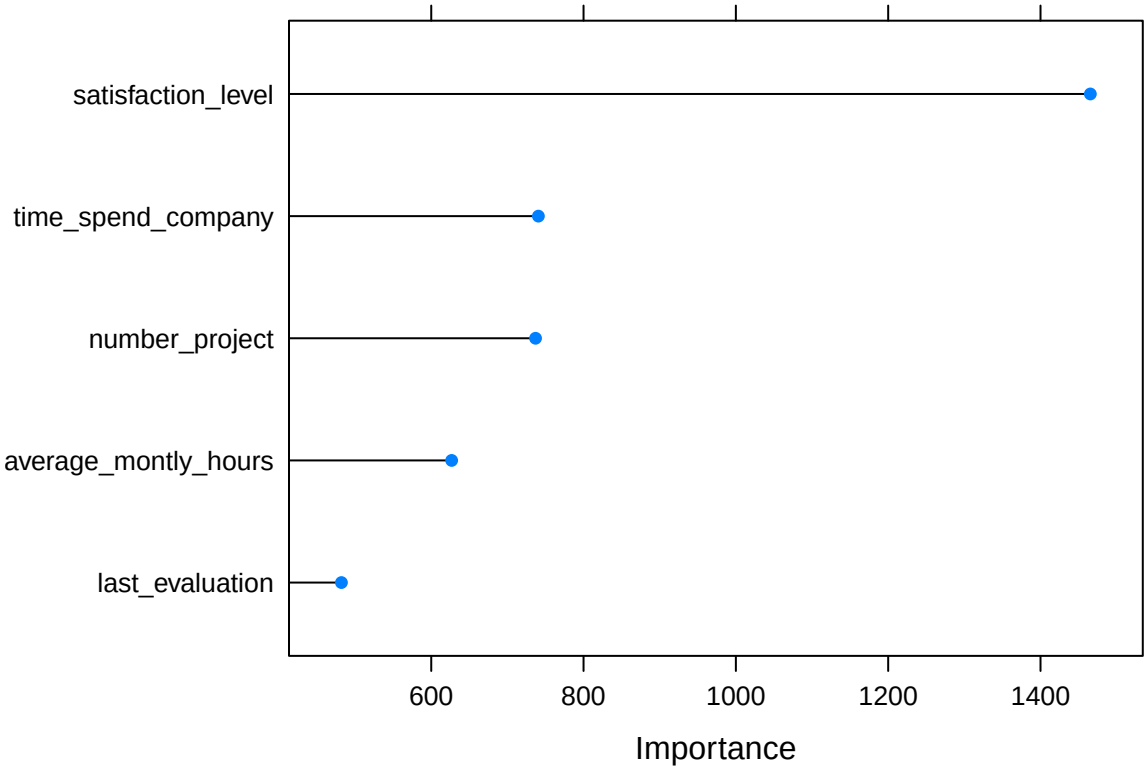
Random Forest

- Initial Model Using All Fields Except Average Monthly Hours as Predictors
  - In-Sample (Accuracy, Sensitivity, Specificity) = (0.9891, 0.9995, 0.9556)
  - Out-of-Sample (Accuracy, Sensitivity, Specificity) = (0.9805, 0.9965, 0.9294)
  - TOP Predictors & Relative Importance
    - Satisfaction Level (1127.152)
    - Number of Projects (705.512)
    - Time Spent with Company (668.831)
    - Average Monthly Hours (543.064)
    - Last Evaluation (441.229)



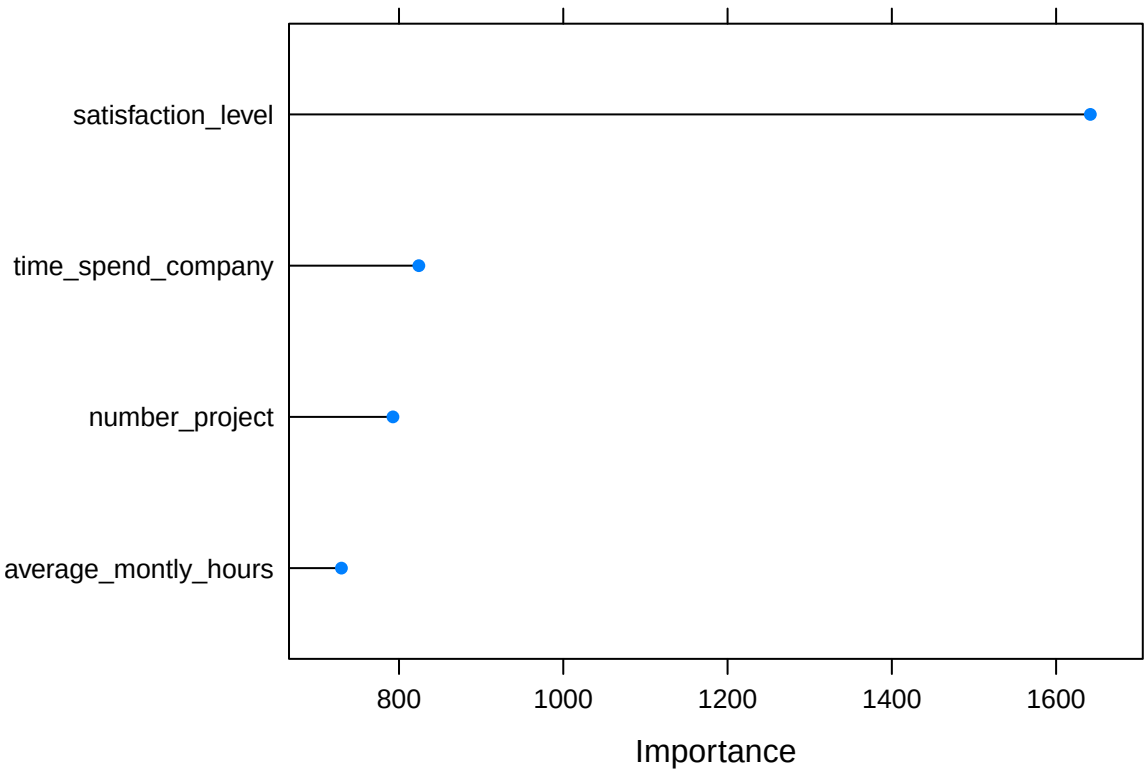
Reduced Predictor Model - 5 Predictors

- Reduced Order Model Using Most Significant Predictors Based on 1st Random Forest
  - In-Sample (Accuracy, Sensitivity, Specificity) = (0.9993, 0.9994, 0.9989)
  - Out-of-Sample (Accuracy, Sensitivity, Specificity) = (0.9893, 0.9961, 0.9675)
  - TOP Predictors & Unscaled Importance
    - Satisfaction Level (1490.3)
    - Time Spent with Company (750.7)
    - Number of Projects (715.9)
    - Average Monthly Hours (626.8)
    - Last Evaluation (467.2)



Reduced Predictor Model - 4 Predictors

- Reduced Order Model Using Most Significant Predictors Based on 2nd Random Forest
  - In-Sample (Accuracy, Sensitivity, Specificity) = (0.9956, 0.9974, 0.9899)
  - Out-of-Sample (Accuracy, Sensitivity, Specificity) = (0.9875, 0.9937, 0.9675)
  - TOP Predictors & Relative Importance
    - Satisfaction Level (1676.3)
    - Time Spent with Company (81)
    - Number of Projects (773.7)
    - Average Monthly Hours (726.2)



Model Remarks

Reducing the number of predictors from the entire set of fields to top 5 results in a highly accurate model with employee satisfaction being the most important variable. A 5 predictor model should be selected because the unscaled importance of the bottom 4 predictors are the same order of magnitude, indicating relatively equal contribution to the model's predictive power. In addition, the out-of-sample accuracy, sensitivity, and specificity of the 4 preditor model are less that those of the 5 predictor model, which suggests that the an employee's last evaluation score is a relevant predictor that cannot be omitted.