BA810 Team Project

Team3: David Hutchens, Jingcheng Huang, Yue Gong, Youming Qiu, Minna Tang, Yishuang Song

2019/10/10

Table of Content

Introduction

Data Overview

- This is a fictional data set created by IBM data scientists.
- There are 1470 rows and 35 columns.
- The source of the data was derived from Kaggle.

Preparation

```
options(repr.plot.width=8, repr.plot.height=4)
```

Load Libraries

```
#Loading libraries
library(readr)
library(tidyverse)
library(ggplot2)
library(ggthemes)
library(cowplot)
library(ggcorrplot)
library(gridExtra)
library(scales)
library(rpart)
library(rpart.plot)
library(ggcorrplot)
library(caret)
library(RColorBrewer)
library(randomForest)
library(gbm)
library(glmnet)
library(plotmo)
theme_set(theme_bw())
```

Read in Data

```
employee <- read_csv('WA_Fn-UseC_-HR-Employee-Attrition.csv')
```

inspect the dataset

```
dim(employee)
```

```
## [1] 1470   35
```

```
head(employee)
```

```
## # A tibble: 6 x 35
##     Age Attrition BusinessTravel DailyRate Department DistanceFromHome
##   <dbl> <chr>     <chr>              <dbl> <chr>                  <dbl>
## 1    41 Yes       Travel_Rarely       1102 Sales                      1
## 2    49 No        Travel_Freque~       279 Research ~                 8
## 3    37 Yes       Travel_Rarely       1373 Research ~                2
## 4    33 No        Travel_Freque~      1392 Research ~                3
## 5    27 No        Travel_Rarely        591 Research ~                2
## 6    32 No        Travel_Freque~      1005 Research ~                2
## # ... with 29 more variables: Education <dbl>, EducationField <chr>,
## #   EmployeeCount <dbl>, EmployeeNumber <dbl>,
## #   EnvironmentSatisfaction <dbl>, Gender <chr>, HourlyRate <dbl>,
## #   JobInvolvement <dbl>, JobLevel <dbl>, JobRole <chr>,
## #   JobSatisfaction <dbl>, MaritalStatus <chr>, MonthlyIncome <dbl>,
## #   MonthlyRate <dbl>, NumCompaniesWorked <dbl>, Over18 <chr>,
## #   OverTime <chr>, PercentSalaryHike <dbl>, PerformanceRating <dbl>,
## #   RelationshipSatisfaction <dbl>, StandardHours <dbl>,
```

```
## #   StockOptionLevel <dbl>, TotalWorkingYears <dbl>,
## #   TrainingTimesLastYear <dbl>, WorkLifeBalance <dbl>,
## #   YearsAtCompany <dbl>, YearsInCurrentRole <dbl>,
## #   YearsSinceLastPromotion <dbl>, YearsWithCurrManager <dbl>
```

```
glimpse(employee)
```

```
## Observations: 1,470
## Variables: 35
## $ Age                      <dbl> 41, 49, 37, 33, 27, 32, 59, 30, 38, 3...
## $ Attrition                <chr> "Yes", "No", "Yes", "No", "No", "No",...
## $ BusinessTravel           <chr> "Travel_Rarely", "Travel_Frequently",...
## $ DailyRate                <dbl> 1102, 279, 1373, 1392, 591, 1005, 132...
## $ Department               <chr> "Sales", "Research & Development", "R...
## $ DistanceFromHome         <dbl> 1, 8, 2, 3, 2, 2, 3, 24, 23, 27, 16, ...
## $ Education                <dbl> 2, 1, 2, 4, 1, 2, 3, 1, 3, 3, 3, 2, 1...
## $ EducationField           <chr> "Life Sciences", "Life Sciences", "Ot...
## $ EmployeeCount            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ EmployeeNumber           <dbl> 1, 2, 4, 5, 7, 8, 10, 11, 12, 13, 14,...
## $ EnvironmentSatisfaction  <dbl> 2, 3, 4, 4, 1, 4, 3, 4, 4, 3, 1, 4, 1...
## $ Gender                   <chr> "Female", "Male", "Male", "Female", "...
## $ HourlyRate               <dbl> 94, 61, 92, 56, 40, 79, 81, 67, 44, 9...
## $ JobInvolvement           <dbl> 3, 2, 2, 3, 3, 3, 4, 3, 2, 3, 4, 2, 3...
## $ JobLevel                 <dbl> 2, 2, 1, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1...
## $ JobRole                  <chr> "Sales Executive", "Research Scientis...
## $ JobSatisfaction          <dbl> 4, 2, 3, 3, 2, 4, 1, 3, 3, 3, 2, 3, 3...
## $ MaritalStatus            <chr> "Single", "Married", "Single", "Marri...
## $ MonthlyIncome            <dbl> 5993, 5130, 2090, 2909, 3468, 3068, 2...
## $ MonthlyRate              <dbl> 19479, 24907, 2396, 23159, 16632, 118...
## $ NumCompaniesWorked       <dbl> 8, 1, 6, 1, 9, 0, 4, 1, 0, 6, 0, 0, 1...
## $ Over18                   <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y...
## $ OverTime                 <chr> "Yes", "No", "Yes", "Yes", "No", "No"...
## $ PercentSalaryHike        <dbl> 11, 23, 15, 11, 12, 13, 20, 22, 21, 1...
## $ PerformanceRating        <dbl> 3, 4, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3...
## $ RelationshipSatisfaction <dbl> 1, 4, 2, 3, 4, 3, 1, 2, 2, 2, 3, 4, 4...
## $ StandardHours            <dbl> 80, 80, 80, 80, 80, 80, 80, 80, 80, 8...
## $ StockOptionLevel         <dbl> 0, 1, 0, 0, 1, 0, 3, 1, 0, 2, 1, 0, 1...
## $ TotalWorkingYears        <dbl> 8, 10, 7, 8, 6, 8, 12, 1, 10, 17, 6, ...
## $ TrainingTimesLastYear    <dbl> 0, 3, 3, 3, 3, 2, 3, 2, 2, 3, 5, 3, 1...
## $ WorkLifeBalance          <dbl> 1, 3, 3, 3, 3, 2, 2, 3, 3, 2, 3, 3, 2...
## $ YearsAtCompany           <dbl> 6, 10, 0, 8, 2, 7, 1, 1, 9, 7, 5, 9, ...
## $ YearsInCurrentRole       <dbl> 4, 7, 0, 7, 2, 7, 0, 0, 7, 7, 4, 5, 2...
## $ YearsSinceLastPromotion  <dbl> 0, 1, 0, 3, 2, 3, 0, 0, 1, 7, 0, 0, 4...
## $ YearsWithCurrManager     <dbl> 5, 7, 0, 0, 2, 6, 0, 0, 8, 7, 3, 8, 3...
```

```
summary(employee)
```

```
##       Age         Attrition         BusinessTravel       DailyRate
##  Min.   :18.00   Length:1470        Length:1470        Min.   : 102.0
##  1st Qu.:30.00   Class :character   Class :character   1st Qu.: 465.0
##  Median :36.00   Mode  :character   Mode  :character   Median : 802.0
##  Mean   :36.92                                         Mean   : 802.5
##  3rd Qu.:43.00                                         3rd Qu.:1157.0
##  Max.   :60.00                                         Max.   :1499.0
##   Department        DistanceFromHome   Education     EducationField
```

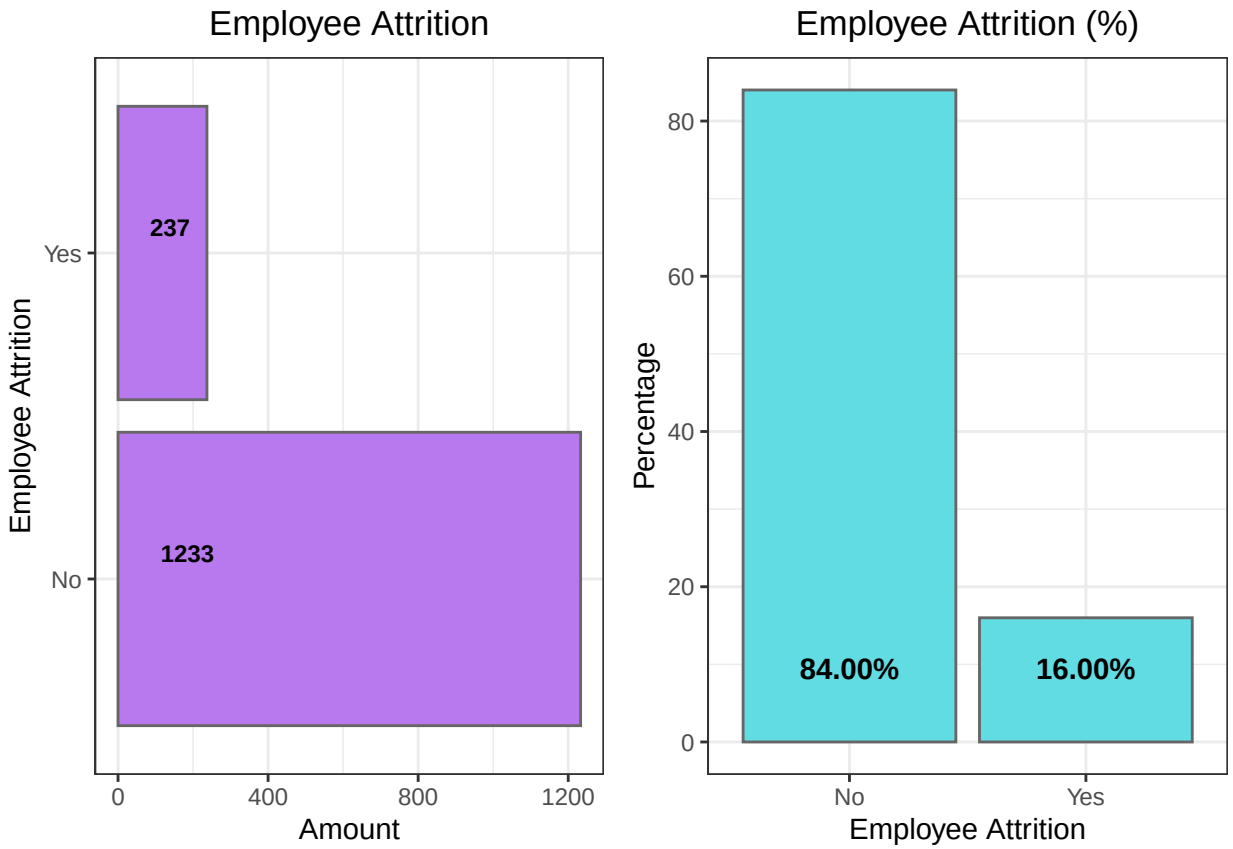
```
##  Length:1470        Min.   : 1.000   Min.   :1.000   Length:1470
##  Class :character   1st Qu.: 2.000   1st Qu.:2.000   Class :character
##  Mode  :character   Median : 7.000   Median :3.000   Mode  :character
##                     Mean   : 9.193   Mean   :2.913
##                     3rd Qu.:14.000   3rd Qu.:4.000
##                     Max.   :29.000   Max.   :5.000
##  EmployeeCount EmployeeNumber   EnvironmentSatisfaction    Gender
##  Min.   :1     Min.   :   1.0   Min.   :1.000           Length:1470
##  1st Qu.:1     1st Qu.: 491.2   1st Qu.:2.000           Class :character
##  Median :1     Median :1020.5   Median :3.000           Mode  :character
##  Mean   :1     Mean   :1024.9   Mean   :2.722
##  3rd Qu.:1     3rd Qu.:1555.8   3rd Qu.:4.000
##  Max.   :1     Max.   :2068.0   Max.   :4.000
##    HourlyRate     JobInvolvement    JobLevel       JobRole
##  Min.   : 30.00   Min.   :1.00   Min.   :1.000   Length:1470
##  1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000   Class :character
##  Median : 66.00   Median :3.00   Median :2.000   Mode  :character
##  Mean   : 65.89   Mean   :2.73   Mean   :2.064
##  3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000
##  Max.   :100.00   Max.   :4.00   Max.   :5.000
##  JobSatisfaction MaritalStatus      MonthlyIncome    MonthlyRate
##  Min.   :1.000   Length:1470        Min.   : 1009   Min.   : 2094
##  1st Qu.:2.000   Class :character   1st Qu.: 2911   1st Qu.: 8047
##  Median :3.000   Mode  :character   Median : 4919   Median :14236
##  Mean   :2.729                      Mean   : 6503   Mean   :14313
##  3rd Qu.:4.000                      3rd Qu.: 8379   3rd Qu.:20462
##  Max.   :4.000                      Max.   :19999   Max.   :26999
##  NumCompaniesWorked    Over18            OverTime
##  Min.   :0.000      Length:1470        Length:1470
##  1st Qu.:1.000      Class :character   Class :character
##  Median :2.000      Mode  :character   Mode  :character
##  Mean   :2.693
##  3rd Qu.:4.000
##  Max.   :9.000
##  PercentSalaryHike PerformanceRating RelationshipSatisfaction
##  Min.   :11.00     Min.   :3.000     Min.   :1.000
##  1st Qu.:12.00     1st Qu.:3.000     1st Qu.:2.000
##  Median :14.00     Median :3.000     Median :3.000
##  Mean   :15.21     Mean   :3.154     Mean   :2.712
##  3rd Qu.:18.00     3rd Qu.:3.000     3rd Qu.:4.000
##  Max.   :25.00     Max.   :4.000     Max.   :4.000
##  StandardHours StockOptionLevel TotalWorkingYears TrainingTimesLastYear
##  Min.   :80    Min.   :0.0000   Min.   : 0.00     Min.   :0.000
##  1st Qu.:80    1st Qu.:0.0000   1st Qu.: 6.00     1st Qu.:2.000
##  Median :80    Median :1.0000   Median :10.00     Median :3.000
##  Mean   :80    Mean   :0.7939   Mean   :11.28     Mean   :2.799
##  3rd Qu.:80    3rd Qu.:1.0000   3rd Qu.:15.00     3rd Qu.:3.000
##  Max.   :80    Max.   :3.0000   Max.   :40.00     Max.   :6.000
##  WorkLifeBalance YearsAtCompany   YearsInCurrentRole
##  Min.   :1.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:2.000   1st Qu.: 3.000   1st Qu.: 2.000
##  Median :3.000   Median : 5.000   Median : 3.000
##  Mean   :2.761   Mean   : 7.008   Mean   : 4.229
##  3rd Qu.:3.000   3rd Qu.: 9.000   3rd Qu.: 7.000
```

```
##  Max.   :4.000   Max.   :40.000   Max.   :18.000
##  YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000          Min.   : 0.000
##  1st Qu.: 0.000          1st Qu.: 2.000
##  Median : 1.000          Median : 3.000
##  Mean   : 2.188          Mean   : 4.123
##  3rd Qu.: 3.000          3rd Qu.: 7.000
##  Max.   :15.000          Max.   :17.000
```

Exploratory Analysis

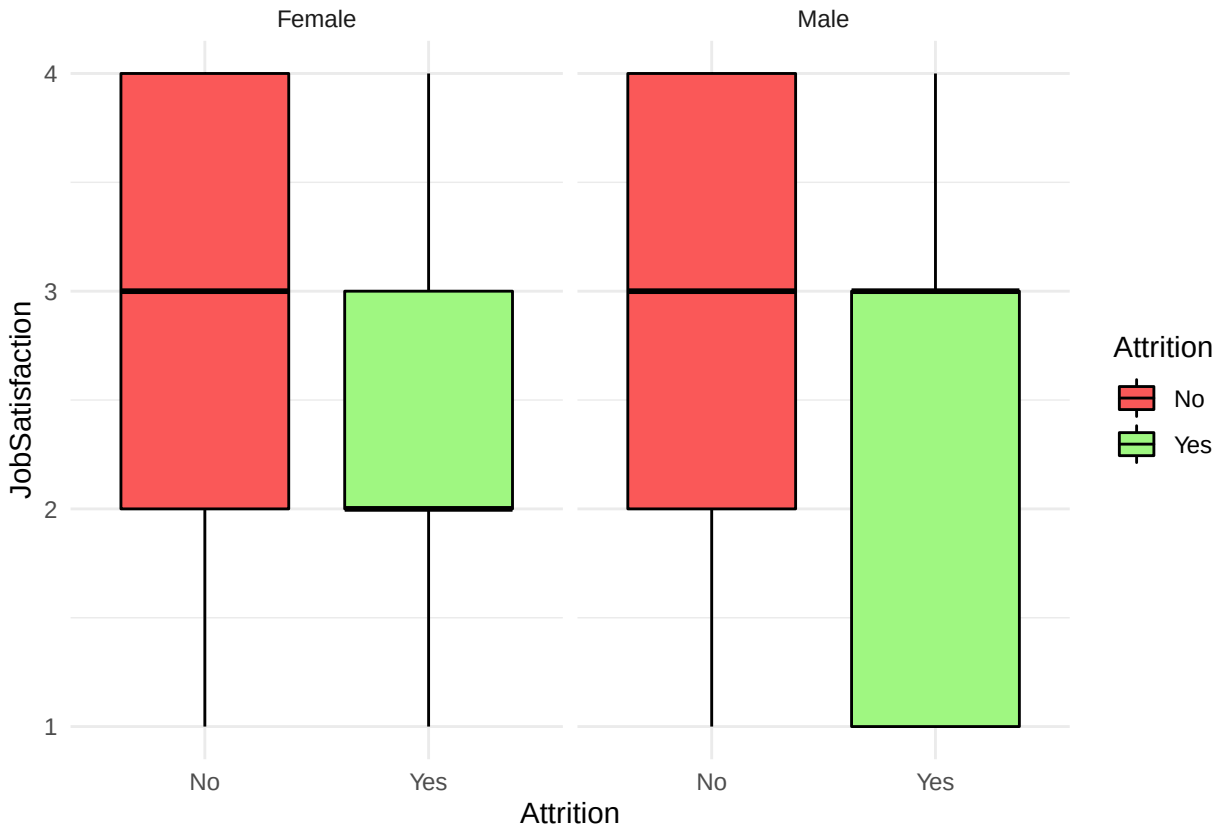
1. Distribution of labels in numbers and percentages

```
attritions_number <- employee %>%
  group_by(Attrition) %>%
  summarize(Count = n()) %>%
  ggplot(aes(x = Attrition, y = Count)) +
  geom_bar(stat = "identity", fill = "#b779ed", color = "grey40") +
  theme_bw() +
  coord_flip() +
  geom_text(aes(x = Attrition, y = 0.01, label = Count),
            hjust = -0.8, vjust = -1, size = 3,
            color = "black", fontface = "bold") +
  labs(title = "Employee Attrition", x = "Employee Attrition", y="Amount") +
  theme(plot.title=element_text(hjust=0.5))

attrition_percentage <- employee %>% group_by(Attrition) %>% summarise(Count=n()) %>%
  mutate(pct = round(prop.table(Count), 2) * 100) %>%
  ggplot(aes(x=Attrition, y=pct)) +
  geom_bar(stat = "identity", fill = "#62dce3", color="grey40") +
  geom_text(aes(x=Attrition, y=0.01, label= sprintf("%.2f%%", pct)),
            hjust=0.5, vjust=-3, size=4,
            color="black", fontface = "bold") +
  theme_bw() +
  labs(x = "Employee Attrition", y="Percentage") +
  labs(title = "Employee Attrition (%)") + theme(plot.title=element_text(hjust=0.5))

plot_grid(attritions_number, attrition_percentage, align="h", ncol=2)
```



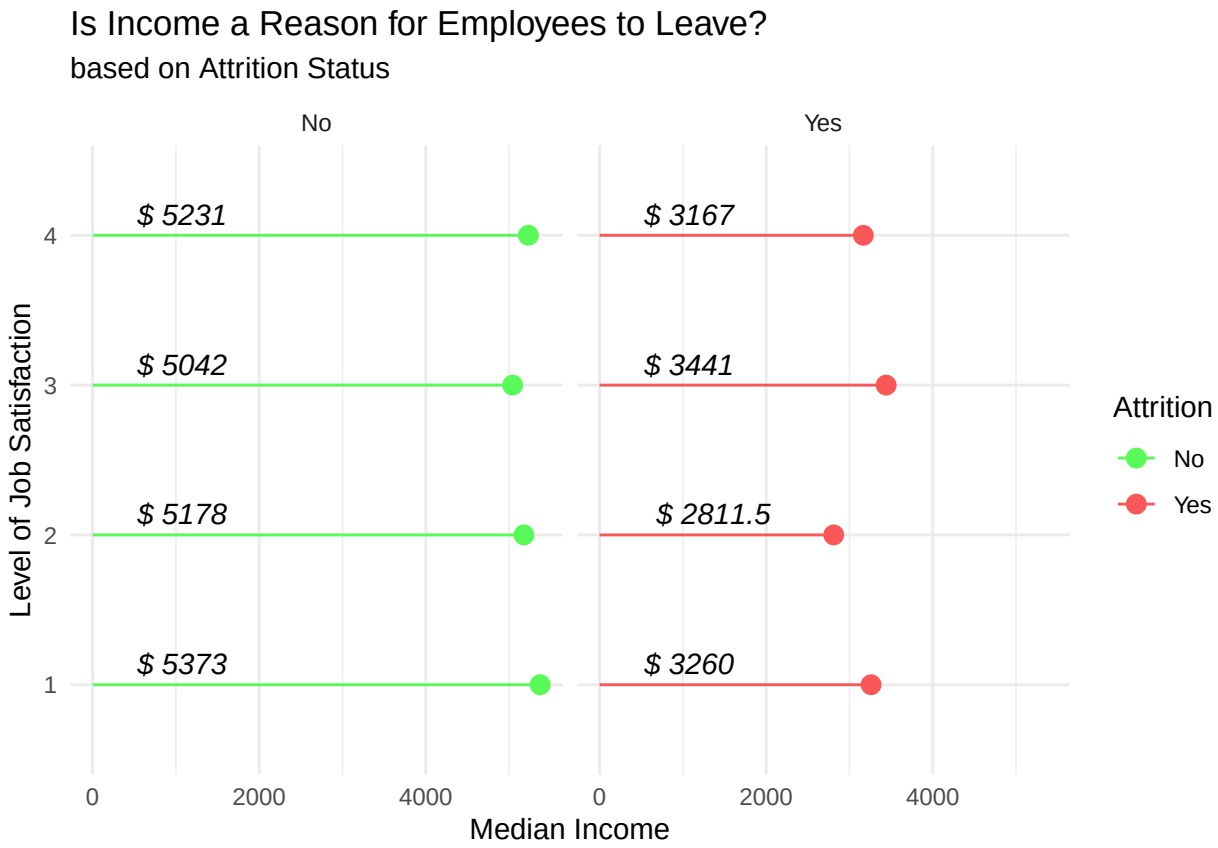
2. Distribution of Job Satisfaction

```
employee %>% select(Attrition, JobSatisfaction, Gender) %>%
  ggplot(aes(x=Attrition, y=JobSatisfaction, fill=Attrition)) +
  geom_boxplot(color="black") +
  theme_minimal() +
  facet_wrap(~Gender) +
  scale_fill_manual(values=c("#FA5858", "#9FF781"))
```



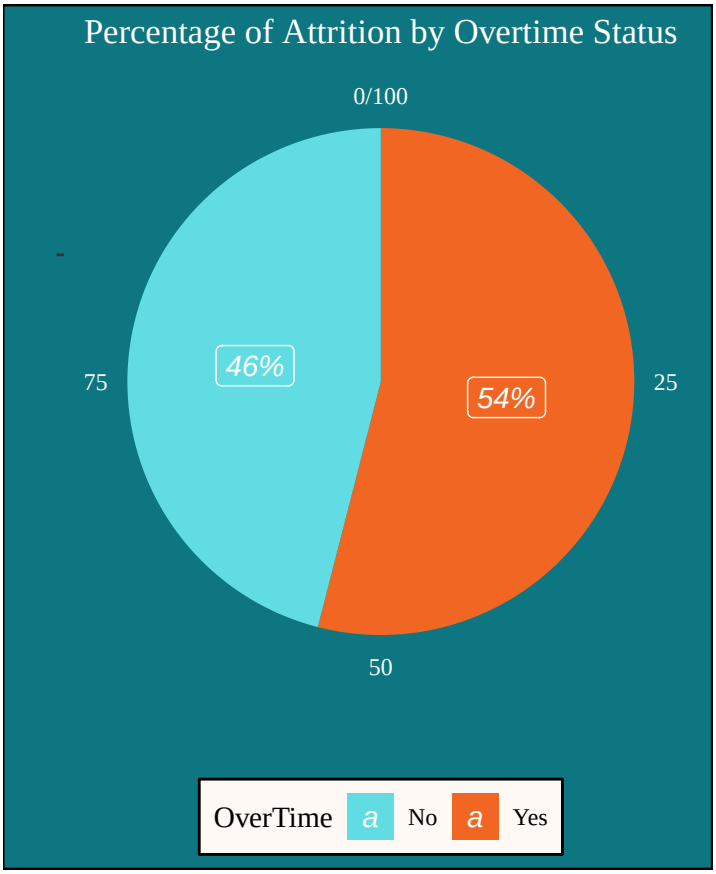
3. Is income a reason for employees to leave the organization?

```
employee$JobSatisfaction <- as.factor(employee$JobSatisfaction)

employee %>% select(JobSatisfaction, MonthlyIncome, Attrition) %>%
  group_by(JobSatisfaction, Attrition) %>%
  summarize(med=median(MonthlyIncome)) %>%
  ggplot(aes(x=JobSatisfaction, y=med, color=Attrition)) +
  geom_point(size=3) +
  geom_segment(aes(x=JobSatisfaction,
                   xend=JobSatisfaction,
                   y=0,
                   yend=med)) + facet_wrap(~Attrition)+
  labs(title="Is Income a Reason for Employees to Leave?",
       subtitle="based on Attrition Status",
       y="Median Income",
       x="Level of Job Satisfaction") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6),
        plot.title=element_text(hjust=0.5),
        strip.background = element_blank(),
        strip.text = element_blank()) +
  coord_flip() + theme_minimal() + scale_color_manual(values=c("#58FA58", "#FA5858")) +
  geom_text(aes(x=JobSatisfaction, y=0.01, label= paste0("$ ", round(med,2))),
            hjust=-0.5, vjust=-0.5, size=4,
            colour="black", fontface="italic",
            angle=360)
```



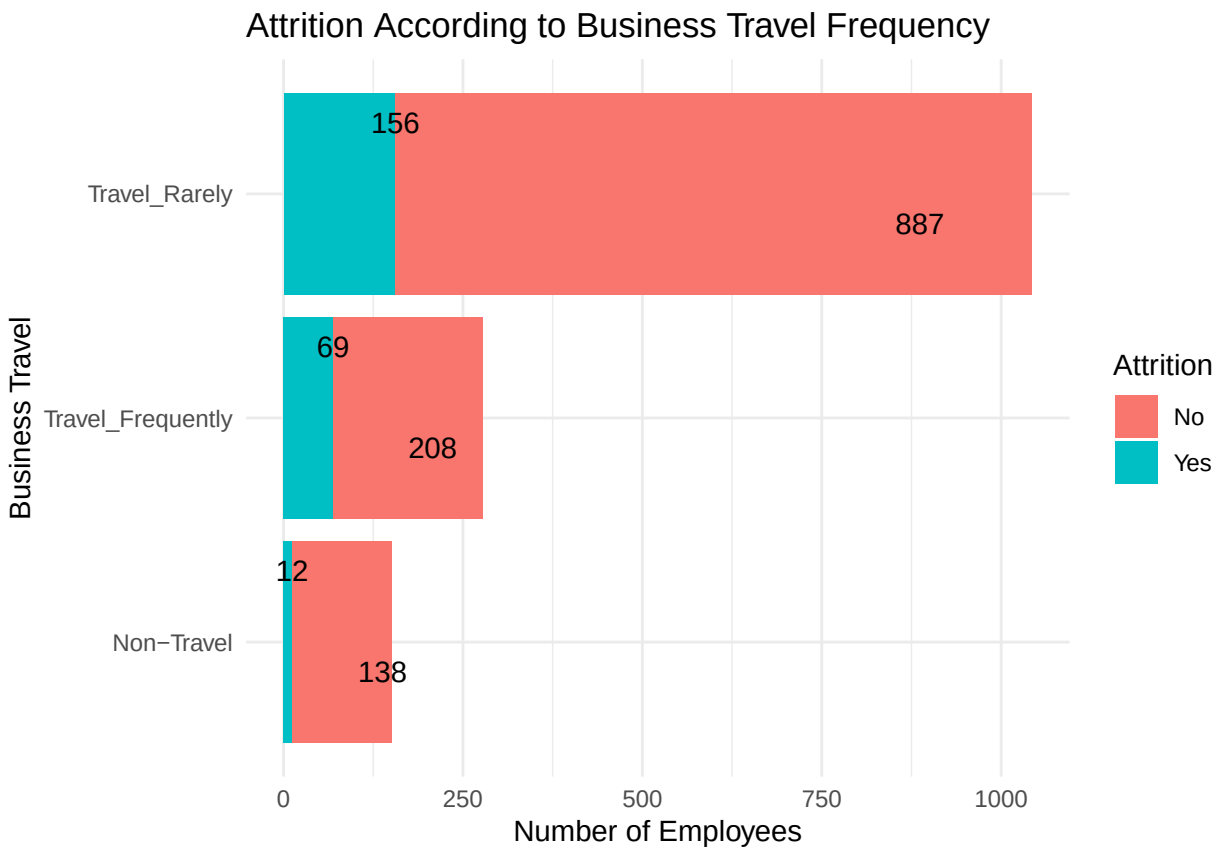
4. Percentage of attrition based on overtime status

```
employee %>% select(OverTime, Attrition) %>%
  filter(Attrition == "Yes") %>%
  group_by(Attrition, OverTime) %>%
  summarize(n=n()) %>%
  mutate(pct=round(prop.table(n),2) * 100) %>%
  ggplot(aes(x="", y=pct, fill=OverTime)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme_tufte() +
  scale_fill_manual(values=c("#62dce3", "#f26624")) +
  geom_label(aes(label = paste0(pct, "%")), position = position_stack(vjust = 0.5),
             colour = "white",  fontface = "italic")+
  theme(
    legend.position="bottom",
    strip.background = element_blank(),
    strip.text.x = element_blank(), plot.title=element_text(hjust=0.5, color="white"),
    plot.subtitle=element_text(color="white"),
    plot.background=element_rect(fill="#0D7680"),
    axis.text.x=element_text(color="white"),
    axis.text.y=element_text(color="white"),
    axis.title=element_text(color="white"),
    legend.background = element_rect(fill="#FFF9F5",
                                     size=0.5, linetype="solid", colour ="black")
    ) +
  labs(title="Percentage of Attrition by Overtime Status", x="", y="")
```



5. Is travel frequency a factor for attrition?

```
employee %>%
  group_by(BusinessTravel, Attrition) %>%
  tally() %>%
  ggplot(aes(x = BusinessTravel, y = n,fill = Attrition)) +
  geom_bar(stat = "identity") +
  theme_minimal()+
  coord_flip() +
  labs(x = "Business Travel", y = "Number of Employees")+
  ggtitle("Attrition According to Business Travel Frequency")+
  geom_text(aes(label = n), vjust = -0.5, position = position_dodge(0.9))
```



Data Cleaning

Check if there's any missing value in this dataset

```
sapply(employee, function(x) sum(is.na(x)))
```

```
##                     Age               Attrition          BusinessTravel
##                       0                       0                       0
##                DailyRate              Department        DistanceFromHome
##                       0                       0                       0
##                Education           EducationField           EmployeeCount
##                       0                       0                       0
##           EmployeeNumber  EnvironmentSatisfaction                  Gender
##                       0                       0                       0
##               HourlyRate          JobInvolvement                JobLevel
##                       0                       0                       0
##                  JobRole          JobSatisfaction           MaritalStatus
##                       0                       0                       0
##            MonthlyIncome             MonthlyRate       NumCompaniesWorked
##                       0                       0                       0
##                   Over18                OverTime        PercentSalaryHike
##                       0                       0                       0
##        PerformanceRating RelationshipSatisfaction           StandardHours
##                       0                       0                       0
##         StockOptionLevel       TotalWorkingYears    TrainingTimesLastYear
```

```
##                         0                         0                         0
##          WorkLifeBalance           YearsAtCompany       YearsInCurrentRole
##                         0                         0                         0
##  YearsSinceLastPromotion     YearsWithCurrManager
##                         0                         0
```

Fortunately, we don't have any missing values in this dataset.

Remove unnecessary variables and make dummy variables

```
# Remove some unnecessary columns
employee$Over18 <- NULL
employee$EmployeeNumber <- NULL
employee$EmployeeCount <- NULL
employee$StandardHours <- NULL

# Change the class of variable from numeric to factor
cols <- c("Attrition","BusinessTravel", "Department","EducationField",
          "Gender","JobRole", "MaritalStatus", "OverTime")

employee[cols] <- lapply(employee[cols], factor)
employee$JobSatisfaction <- as.numeric(employee$JobSatisfaction)

## Change other variables to dummy variables
employee<-fastDummies::dummy_cols(employee, remove_first_dummy = T)

colnames(employee)[colnames(employee) == "Department_Research & Development"] <-
  "Department_Research_Development"
colnames(employee)[colnames(employee) == "EducationField_Life Sciences"] <-
  "EducationField_Life_Sciences"
colnames(employee)[colnames(employee) == "EducationField_Technical Degree"] <-
  "EducationField_Technical_Degree"
colnames(employee)[colnames(employee) == "JobRole_Human Resources"] <-
  "JobRole_Human_Resources"
colnames(employee)[colnames(employee) == "JobRole_Laboratory Technician"] <-
  "JobRole_Laboratory_Technician"
colnames(employee)[colnames(employee) == "JobRole_Manufacturing Director"] <-
  "JobRole_Manufacturing_Director"
colnames(employee)[colnames(employee) == "JobRole_Research Director"] <-
  "JobRole_Research_Director"
colnames(employee)[colnames(employee) == "JobRole_Research Scientist"] <-
  "JobRole_Research_Scientist"
colnames(employee)[colnames(employee) == "JobRole_Sales Executive"] <-
  "JobRole_Sales_Executive"
colnames(employee)[colnames(employee) == "JobRole_Sales Representative"] <-
  "JobRole_Sales_Representative"

employee[cols] <- NULL

# now we have 1470 observations / 45 columns
dim(employee)
```

```
## [1] 1470   45
```

```r
glimpse(employee)
```

```
## Observations: 1,470
## Variables: 45
## $ Age                                <dbl> 41, 49, 37, 33, 27, 32, 59, 3...
## $ DailyRate                          <dbl> 1102, 279, 1373, 1392, 591, 1...
## $ DistanceFromHome                   <dbl> 1, 8, 2, 3, 2, 2, 3, 24, 23, ...
## $ Education                          <dbl> 2, 1, 2, 4, 1, 2, 3, 1, 3, 3,...
## $ EnvironmentSatisfaction            <dbl> 2, 3, 4, 4, 1, 4, 3, 4, 4, 3,...
## $ HourlyRate                         <dbl> 94, 61, 92, 56, 40, 79, 81, 6...
## $ JobInvolvement                     <dbl> 3, 2, 2, 3, 3, 3, 4, 3, 2, 3,...
## $ JobLevel                           <dbl> 2, 2, 1, 1, 1, 1, 1, 1, 3, 2,...
## $ JobSatisfaction                    <dbl> 4, 2, 3, 3, 2, 4, 1, 3, 3, 3,...
## $ MonthlyIncome                      <dbl> 5993, 5130, 2090, 2909, 3468,...
## $ MonthlyRate                        <dbl> 19479, 24907, 2396, 23159, 16...
## $ NumCompaniesWorked                 <dbl> 8, 1, 6, 1, 9, 0, 4, 1, 0, 6,...
## $ PercentSalaryHike                  <dbl> 11, 23, 15, 11, 12, 13, 20, 2...
## $ PerformanceRating                  <dbl> 3, 4, 3, 3, 3, 3, 4, 4, 4, 3,...
## $ RelationshipSatisfaction           <dbl> 1, 4, 2, 3, 4, 3, 1, 2, 2, 2,...
## $ StockOptionLevel                   <dbl> 0, 1, 0, 0, 1, 0, 3, 1, 0, 2,...
## $ TotalWorkingYears                  <dbl> 8, 10, 7, 8, 6, 8, 12, 1, 10,...
## $ TrainingTimesLastYear              <dbl> 0, 3, 3, 3, 3, 2, 3, 2, 2, 3,...
## $ WorkLifeBalance                    <dbl> 1, 3, 3, 3, 3, 2, 2, 3, 3, 2,...
## $ YearsAtCompany                     <dbl> 6, 10, 0, 8, 2, 7, 1, 1, 9, 7...
## $ YearsInCurrentRole                 <dbl> 4, 7, 0, 7, 2, 7, 0, 0, 7, 7,...
## $ YearsSinceLastPromotion            <dbl> 0, 1, 0, 3, 2, 3, 0, 0, 1, 7,...
## $ YearsWithCurrManager               <dbl> 5, 7, 0, 0, 2, 6, 0, 0, 8, 7,...
## $ Attrition_Yes                      <int> 1, 0, 1, 0, 0, 0, 0, 0, 0, 0,...
## $ BusinessTravel_Travel_Frequently <int> 0, 1, 0, 1, 0, 1, 0, 0, 1, 0,...
## $ BusinessTravel_Travel_Rarely       <int> 1, 0, 1, 0, 1, 0, 1, 1, 0, 1,...
## $ Department_Research_Development    <int> 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ Department_Sales                   <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ EducationField_Life_Sciences       <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 0,...
## $ EducationField_Marketing           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ EducationField_Medical             <int> 0, 0, 0, 0, 1, 0, 1, 0, 0, 1,...
## $ EducationField_Other               <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...
## $ EducationField_Technical_Degree    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Gender_Male                        <int> 0, 1, 1, 0, 1, 1, 0, 1, 1, 1,...
## $ JobRole_Human_Resources            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ JobRole_Laboratory_Technician      <int> 0, 0, 1, 0, 1, 1, 1, 1, 0, 0,...
## $ JobRole_Manager                    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ JobRole_Manufacturing_Director     <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,...
## $ JobRole_Research_Director          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ JobRole_Research_Scientist         <int> 0, 1, 0, 1, 0, 0, 0, 0, 0, 0,...
## $ JobRole_Sales_Executive            <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ JobRole_Sales_Representative       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ MaritalStatus_Married              <int> 0, 1, 0, 1, 1, 0, 1, 0, 0, 1,...
## $ MaritalStatus_Single               <int> 1, 0, 1, 0, 0, 1, 0, 0, 1, 0,...
## $ OverTime_Yes                       <int> 1, 0, 1, 1, 0, 0, 1, 0, 0, 0,...
```

Split into the training and testing datasets

```r
set.seed(122333)

# Determine sample size
employee$train <- sample(c(0,1), nrow(employee), replace = TRUE, prob = c(0.2, 0.8))

# train dataset - 1166 observations
employee_train <- employee %>% filter(train == 1)
employee_train$train <- NULL

# test dataset - 304 observations
employee_test <- employee %>% filter(train == 0)
employee_test$train <- NULL
```

Linear Regression Model

```r
fl <- as.formula(Attrition_Yes ~.)
y_train <- employee_train$Attrition_Yes
y_test <- employee_test$Attrition_Yes

fit_lm <- lm(fl, data=employee_train)
sort(coef(fit_lm), decreasing = T)
```

```
##                       (Intercept)     JobRole_Sales_Representative
##                      4.274117e-01                      2.945214e-01
##          JobRole_Human_Resources                      OverTime_Yes
##                      2.337091e-01                      1.988552e-01
##  Department_Research_Development                Department_Sales
##                      1.630962e-01                      1.603811e-01
##    JobRole_Laboratory_Technician             MaritalStatus_Single
##                      1.483971e-01                      1.312356e-01
## BusinessTravel_Travel_Frequently          JobRole_Sales_Executive
##                      1.254205e-01                      1.201653e-01
##                   JobRole_Manager     BusinessTravel_Travel_Rarely
##                      8.095201e-02                      7.487971e-02
##       JobRole_Research_Scientist   JobRole_Manufacturing_Director
##                      4.878806e-02                      3.122136e-02
##                       Gender_Male                PerformanceRating
##                      3.119479e-02                      2.017963e-02
##        JobRole_Research_Director                NumCompaniesWorked
##                      1.840262e-02                      1.807907e-02
##            MaritalStatus_Married          YearsSinceLastPromotion
##                      1.245099e-02                      1.143890e-02
##                     YearsAtCompany                         Education
##                      6.777340e-03                      4.744595e-03
##                  DistanceFromHome                       MonthlyRate
##                      3.626020e-03                      5.124099e-07
##                     MonthlyIncome                         DailyRate
##                     -1.044230e-06                     -1.211588e-05
##                        HourlyRate                PercentSalaryHike
##                     -3.346293e-04                     -1.455349e-03
##                               Age                          JobLevel
```

```
##                    -1.959089e-03                    -3.198567e-03
##                 StockOptionLevel                TotalWorkingYears
##                    -4.246847e-03                    -4.989227e-03
##               YearsInCurrentRole             YearsWithCurrManager
##                    -8.796707e-03                    -9.199013e-03
##            TrainingTimesLastYear         RelationshipSatisfaction
##                    -1.253223e-02                    -1.559366e-02
##          EnvironmentSatisfaction                  WorkLifeBalance
##                    -2.948363e-02                    -3.233764e-02
##                  JobSatisfaction                   JobInvolvement
##                    -4.295167e-02                    -5.738001e-02
##  EducationField_Technical_Degree         EducationField_Marketing
##                    -6.872007e-02                    -1.433016e-01
##     EducationField_Life_Sciences             EducationField_Other
##                    -1.717470e-01                    -1.911866e-01
##           EducationField_Medical
##                    -1.959126e-01
```

```
yhat_train_lm <- predict(fit_lm)
mse_train_lm <- mean((y_train - yhat_train_lm) ^ 2)
mse_train_lm
```

```
## [1] 0.09873947
```

```
yhat_test_lm <- predict(fit_lm, employee_test)
mse_test_lm <- mean((y_test - yhat_test_lm) ^ 2)
mse_test_lm
```

```
## [1] 0.111862
```

Stepwise Regression

Forward Selection

```
fit_fw_min <- lm(Attrition_Yes ~ 1, data = employee_train)
fit_fw_max <- as.formula(lm(Attrition_Yes ~.,data = employee_train))
fw<-step(fit_fw_min, direction = "forward",scope = fit_fw_max)
```

```
sort(coef(fw), decreasing = T)
```

```
##                       (Intercept)                       OverTime_Yes
##                       0.517271117                        0.199923348
##     JobRole_Sales_Representative BusinessTravel_Travel_Frequently
##                       0.182571081                        0.132968276
##              MaritalStatus_Single    JobRole_Laboratory_Technician
##                       0.129878573                        0.115755826
##  EducationField_Technical_Degree     BusinessTravel_Travel_Rarely
##                       0.108011345                        0.078638051
##                       Gender_Male               NumCompaniesWorked
##                       0.031281525                        0.016184820
##          YearsSinceLastPromotion                  DistanceFromHome
##                       0.012519749                        0.003655603
##                TotalWorkingYears               YearsInCurrentRole
##                      -0.006714963                       -0.007430375
```

```
##            TrainingTimesLastYear         RelationshipSatisfaction
##                      -0.014515115                      -0.015105893
##          EnvironmentSatisfaction                  WorkLifeBalance
##                      -0.029338515                      -0.032939788
##                   JobSatisfaction                   JobInvolvement
##                      -0.042637771                      -0.061021194
##  Department_Research_Development
##                      -0.091691056
```

```
yhat_train_fw <- predict(fw)
mse_train_fw <- mean((y_train - yhat_train_fw) ^ 2)
mse_train_fw
```

```
## [1] 0.1004298
```

```
yhat_test_fw <- predict(fw, employee_test)
mse_test_fw <- mean((y_test - yhat_test_fw) ^ 2)
mse_test_fw
```

```
## [1] 0.1133327
```

Backward Selection

```
fit_bw_min <- as.formula(lm(Attrition_Yes ~ 1, data = employee_train))
fit_bw_max <- lm(Attrition_Yes ~.,data = employee_train)
bw<-step(fit_bw_max, direction = "backward",scope = fit_bw_min)
bw
```

```
sort(coef(bw), decreasing = T)
```

```
##                       (Intercept)                     OverTime_Yes
##                       0.554107817                      0.197395598
##     JobRole_Sales_Representative              MaritalStatus_Single
##                       0.176111424                      0.131032633
## BusinessTravel_Travel_Frequently     JobRole_Laboratory_Technician
##                       0.129427094                      0.111739347
##     BusinessTravel_Travel_Rarely                  Department_Sales
##                       0.076524773                      0.075648728
##                       Gender_Male               NumCompaniesWorked
##                       0.030617113                      0.017372578
##          YearsSinceLastPromotion                     YearsAtCompany
##                       0.011505385                      0.006730772
##                  DistanceFromHome                TotalWorkingYears
##                       0.003603010                     -0.008005535
##                YearsInCurrentRole              YearsWithCurrManager
##                      -0.008574365                     -0.008626961
##            TrainingTimesLastYear         RelationshipSatisfaction
##                      -0.013196173                     -0.014998898
##          EnvironmentSatisfaction                  WorkLifeBalance
##                      -0.029718738                     -0.032115995
##                   JobSatisfaction                   JobInvolvement
##                      -0.043387363                     -0.058809148
##         EducationField_Marketing     EducationField_Life_Sciences
##                      -0.085031856                     -0.115640541
##             EducationField_Other           EducationField_Medical
##                      -0.128090875                     -0.138085122
```

```
yhat_train_bw <- predict(bw)
mse_train_bw <- mean((y_train - yhat_train_bw) ^ 2)
mse_train_bw
```

```
## [1] 0.09952811
```

```
yhat_test_bw <- predict(bw, employee_test)
mse_test_bw <- mean((y_test - yhat_test_bw) ^ 2)
mse_test_bw
```

```
## [1] 0.1136218
```

Regularization Method

Ridge Regularization Method

```
set.seed(1234)

#Create the formula
fl

#Arrange data into matrices for glmnet
x_train <- model.matrix(fl, employee_train)[ , -1]
x_test <- model.matrix(fl, employee_test)[ , -1]

#Response Variables
y_train <- employee_train$Attrition_Yes
y_test <- employee_test$Attrition_Yes
```

```
cv_fit_ridge <- cv.glmnet(x_train, y_train, alpha = 0, nfolds = 5) # cross validation model
fit_ridge <- cv_fit_ridge$glmnet.fit # ridge model
```

```
best_lam_ridge <- cv_fit_ridge$lambda.min
coef(fit_ridge, s = best_lam_ridge, 50)
```

```
## 45 x 1 sparse Matrix of class "dgCMatrix"
##                                          1
## (Intercept)                   5.516889e-01
## Age                          -2.040150e-03
## DailyRate                    -1.598049e-05
## DistanceFromHome              3.163451e-03
## Education                     3.484291e-03
## EnvironmentSatisfaction      -2.547331e-02
## HourlyRate                   -3.247058e-04
## JobInvolvement               -5.194187e-02
## JobLevel                     -1.020463e-02
## JobSatisfaction              -3.749240e-02
## MonthlyIncome                -1.553884e-06
## MonthlyRate                   5.237257e-07
## NumCompaniesWorked            1.411533e-02
## PercentSalaryHike            -8.658500e-04
## PerformanceRating             1.322541e-02
## RelationshipSatisfaction     -1.284933e-02
```

```
## StockOptionLevel                 -1.237787e-02
## TotalWorkingYears                -3.030030e-03
## TrainingTimesLastYear            -1.124809e-02
## WorkLifeBalance                  -2.920190e-02
## YearsAtCompany                    2.744124e-03
## YearsInCurrentRole               -6.197305e-03
## YearsSinceLastPromotion           9.636140e-03
## YearsWithCurrManager             -6.400180e-03
## BusinessTravel_Travel_Frequently  9.132310e-02
## BusinessTravel_Travel_Rarely      4.428472e-02
## Department_Research_Development  -2.452908e-02
## Department_Sales                  3.597641e-02
## EducationField_Life_Sciences     -2.060724e-02
## EducationField_Marketing          1.309781e-02
## EducationField_Medical           -4.154977e-02
## EducationField_Other             -3.642342e-02
## EducationField_Technical_Degree   7.009248e-02
## Gender_Male                       2.595762e-02
## JobRole_Human_Resources           4.856473e-02
## JobRole_Laboratory_Technician     9.120790e-02
## JobRole_Manager                   1.166248e-02
## JobRole_Manufacturing_Director   -1.380296e-02
## JobRole_Research_Director        -1.387307e-02
## JobRole_Research_Scientist        1.606551e-03
## JobRole_Sales_Executive           1.235948e-02
## JobRole_Sales_Representative      1.628414e-01
## MaritalStatus_Married            -3.953122e-03
## MaritalStatus_Single              9.639242e-02
## OverTime_Yes                      1.742532e-01
```

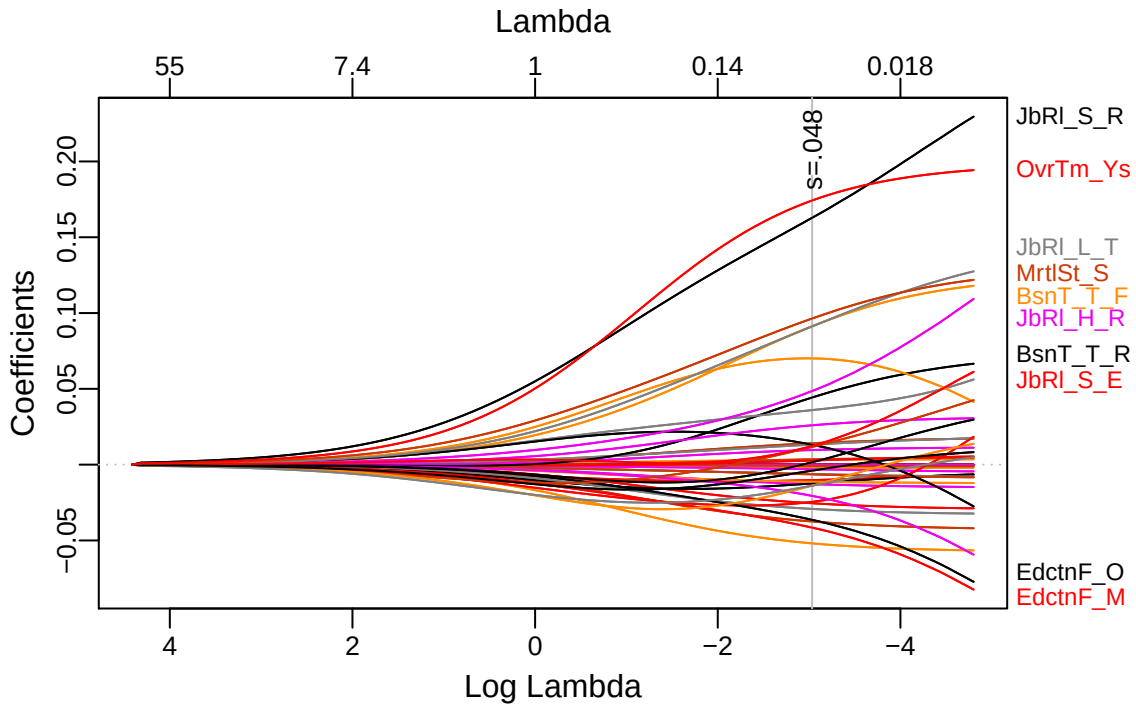
```
yhat_train_ridge <- predict(fit_ridge, x_train, s = best_lam_ridge)
mse_train_ridge <- mean((y_train - yhat_train_ridge) ^ 2)
mse_train_ridge
```

```
## [1] 0.09987582
```

```
yhat_test_ridge <- predict(fit_ridge, x_test, s = best_lam_ridge)
mse_test_ridge <- mean((y_test - yhat_test_ridge) ^ 2)
mse_test_ridge
```

```
## [1] 0.1117195
```

```
plot_glmnet(fit_ridge, s = best_lam_ridge)
```



Lasso Regularization Method

```
cv_fit_lasso <- cv.glmnet(x_train, y_train, alpha = 1, nfolds = 5) # cross validation model
fit_lasso <- cv_fit_lasso$glmnet.fit # lasso model
```

```
best_lam_lasso <- cv_fit_lasso$lambda.min
coef(fit_lasso, s = best_lam_lasso, 50)
```

```
## 45 x 1 sparse Matrix of class "dgCMatrix"
##                                          1
## (Intercept)                   5.232900e-01
## Age                          -1.859973e-03
## DailyRate                    -1.145494e-05
## DistanceFromHome              3.331362e-03
## Education                     2.044451e-03
## EnvironmentSatisfaction      -2.766121e-02
## HourlyRate                   -2.603288e-04
## JobInvolvement               -5.533326e-02
## JobLevel                     -6.258946e-03
## JobSatisfaction              -4.114530e-02
## MonthlyIncome                 .
## MonthlyRate                   3.378259e-07
## NumCompaniesWorked            1.591611e-02
## PercentSalaryHike             .
## PerformanceRating             2.562527e-03
## RelationshipSatisfaction     -1.308656e-02
## StockOptionLevel             -4.799077e-03
## TotalWorkingYears            -4.234291e-03
## TrainingTimesLastYear        -1.117479e-02
## WorkLifeBalance              -3.015239e-02
## YearsAtCompany                3.636082e-03
```

```
## YearsInCurrentRole               -6.735524e-03
## YearsSinceLastPromotion           1.076202e-02
## YearsWithCurrManager             -7.010625e-03
## BusinessTravel_Travel_Frequently  1.073252e-01
## BusinessTravel_Travel_Rarely      5.638716e-02
## Department_Research_Development  -5.068802e-03
## Department_Sales                  5.930988e-02
## EducationField_Life_Sciences     -2.777991e-02
## EducationField_Marketing          .
## EducationField_Medical           -5.050204e-02
## EducationField_Other             -3.969153e-02
## EducationField_Technical_Degree   6.580203e-02
## Gender_Male                       2.687064e-02
## JobRole_Human_Resources           7.395069e-02
## JobRole_Laboratory_Technician     1.120576e-01
## JobRole_Manager                   1.895535e-03
## JobRole_Manufacturing_Director    .
## JobRole_Research_Director        -1.272129e-02
## JobRole_Research_Scientist        1.219145e-02
## JobRole_Sales_Executive           1.561266e-02
## JobRole_Sales_Representative      1.864574e-01
## MaritalStatus_Married             1.459048e-03
## MaritalStatus_Single              1.180465e-01
## OverTime_Yes                      1.939451e-01
```

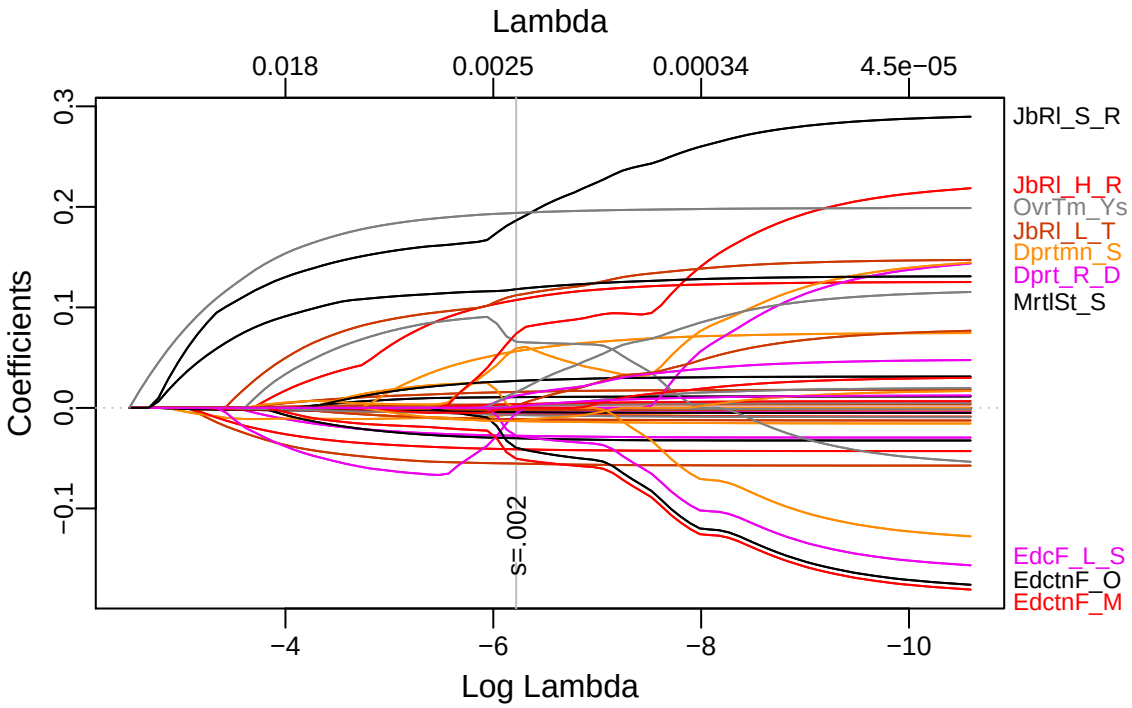
```
yhat_train_lasso <- predict(fit_lasso, x_train, s = best_lam_lasso)
mse_train_lasso <- mean((y_train - yhat_train_ridge) ^ 2)
mse_train_lasso
```

```
## [1] 0.09987582
```

```
yhat_test_lasso <- predict(fit_lasso, x_test, s = best_lam_lasso)
mse_test_lasso <- mean((y_test - yhat_test_lasso) ^ 2)
mse_test_lasso
```

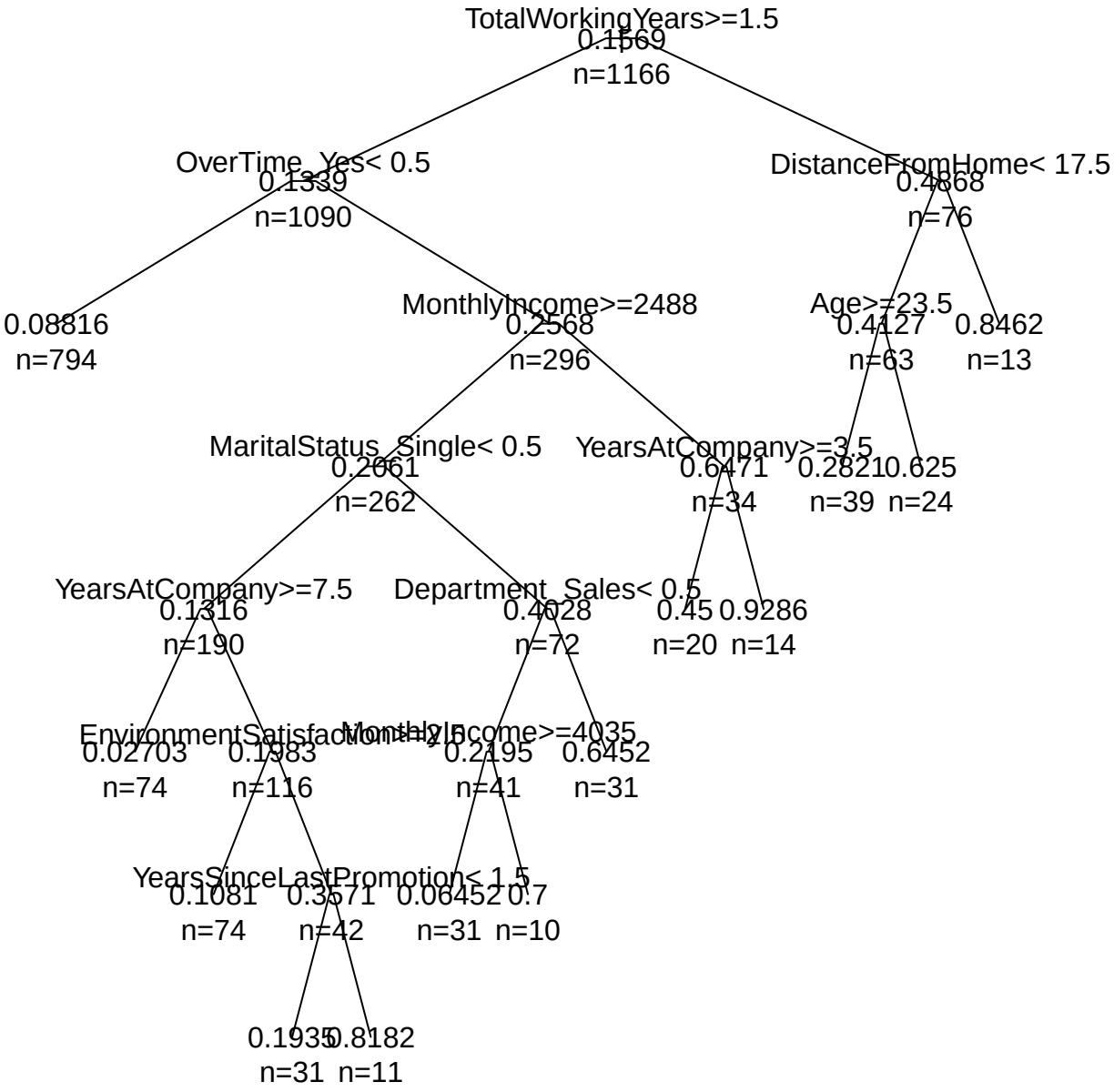
```
## [1] 0.111514
```

```
plot_glmnet(fit_lasso, s = best_lam_lasso)
```



Decision Tree Model

```
rpart.tree <- rpart(Attrition_Yes ~ ., data = employee_train)
plot(rpart.tree, uniform = TRUE, branch = 0.05, margin = 0.08, cex=0.5)
text(rpart.tree, all = TRUE, use.n = TRUE)
title("Training Set's Decision Tree Model")
```




```
# find train MSE
yhat_dt_train <- predict(rpart.tree, employee_train)
mse_dt_train <- mean((yhat_dt_train - y_train) ^ 2)
print(mse_dt_train)
```

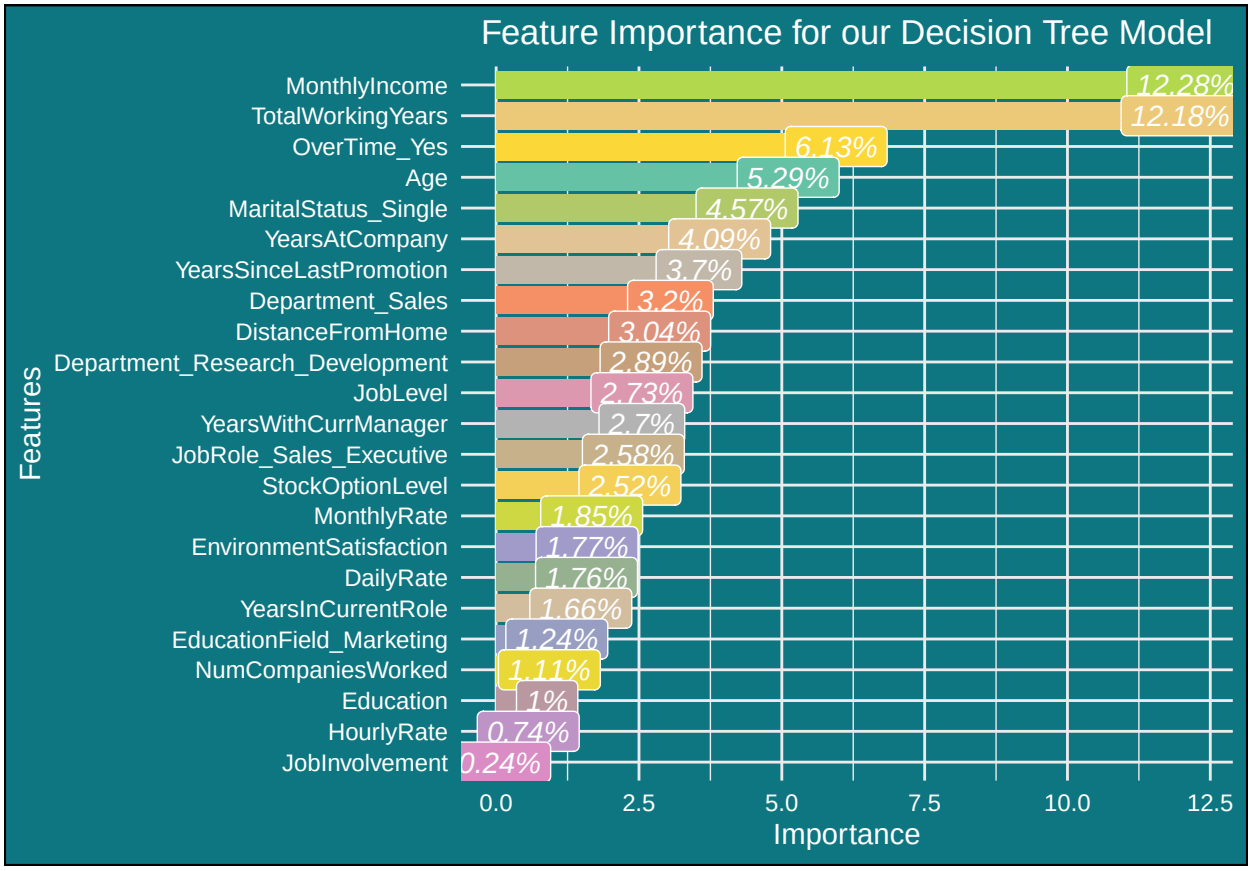
```
## [1] 0.09566546
```

```
# find test MSE
yhat_dt_test <- predict(rpart.tree, employee_test)
mse_dt_test <- mean((yhat_dt_test - y_test) ^ 2)
print(mse_dt_test)
```

```
## [1] 0.1202312
```

Variable Importance of Decision Trees

```
var_imp <- data.frame(rpart.tree$variable.importance)
var_imp$features <- rownames(var_imp)
var_imp <- var_imp[, c(2, 1)]
var_imp$importance <- round(var_imp$rpart.tree.variable.importance, 2)
var_imp$rpart.tree.variable.importance <- NULL

colorCount <- length(unique(var_imp$features))
feature_importance <- var_imp %>%
  ggplot(aes(x = reorder(features, importance), y = importance, fill = features)) +
  geom_bar(stat = 'identity') +
  coord_flip() +
  theme_minimal() +
  theme(
    legend.position = "none",
    strip.background = element_blank(),
    strip.text.x = element_blank(),
    plot.title=element_text(hjust = 0.5, color = "white"),
    plot.subtitle = element_text(color = "white"),
    plot.background=element_rect(fill = "#0D7680"),
    axis.text.x=element_text(color = "white"),
    axis.text.y=element_text(color = "white"),
    axis.title=element_text(color = "white"),
    legend.background = element_rect(fill = "#FFF9F5",
                                     size = 0.5, linetype = "solid",
                                     color ="black")
    ) +
  scale_fill_manual(values = colorRampPalette(brewer.pal(8, "Set2"))(colorCount)) +
  geom_label(aes(label=paste0(importance, "%")),
             color = "white", fontface = "italic", hjust = 0.6) +
  labs(title="Feature Importance for our Decision Tree Model",
       x = "Features", y = "Importance")

feature_importance
```



Random forest

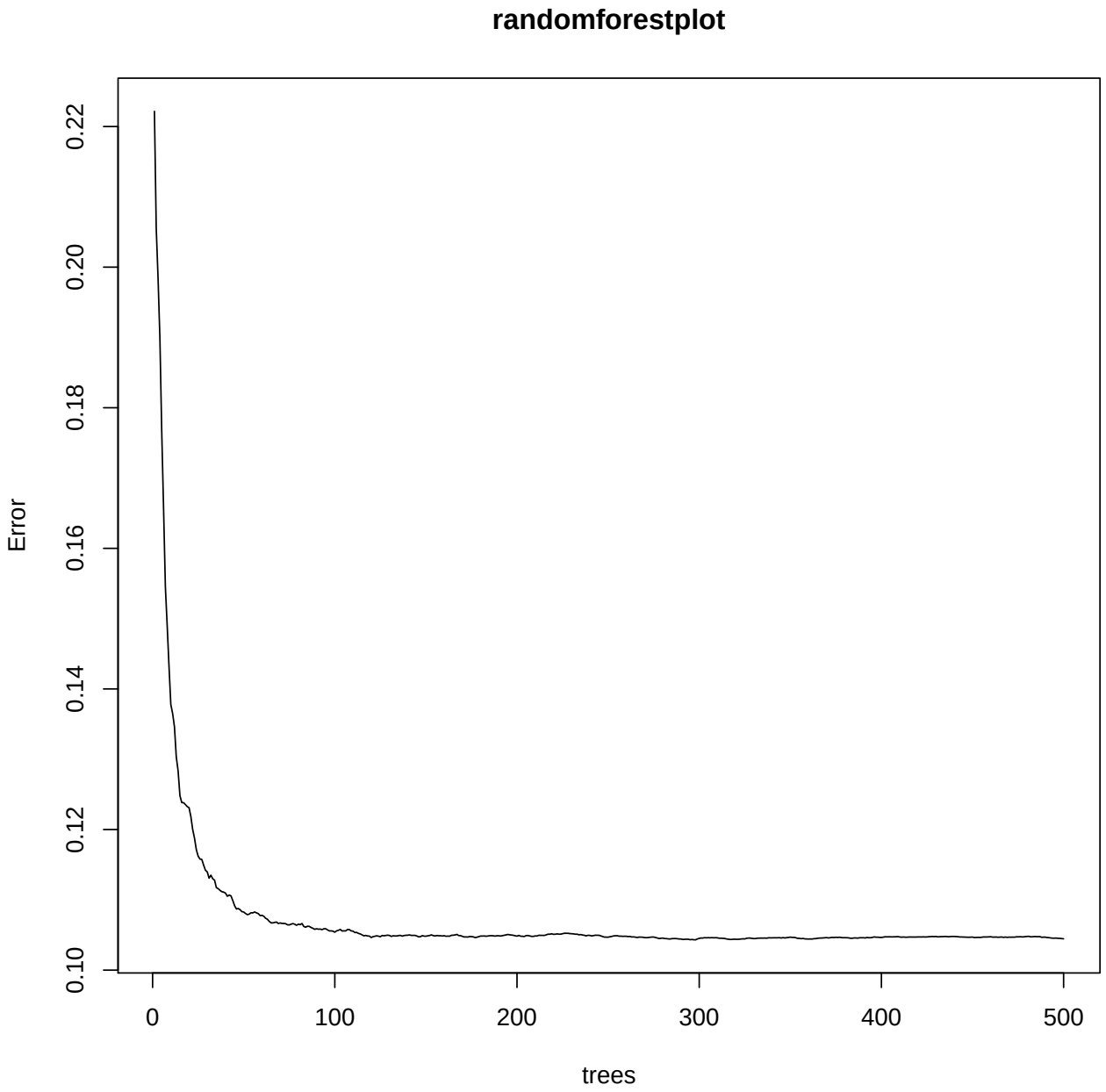
```
set.seed(654321)
rf <- as.formula(Attrition_Yes~., employee_train)
randomforestplot <- randomForest(rf,
                                 employee_train,
                                 ntree=500,
                                 mtry=8,
                                 do.trace=F)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

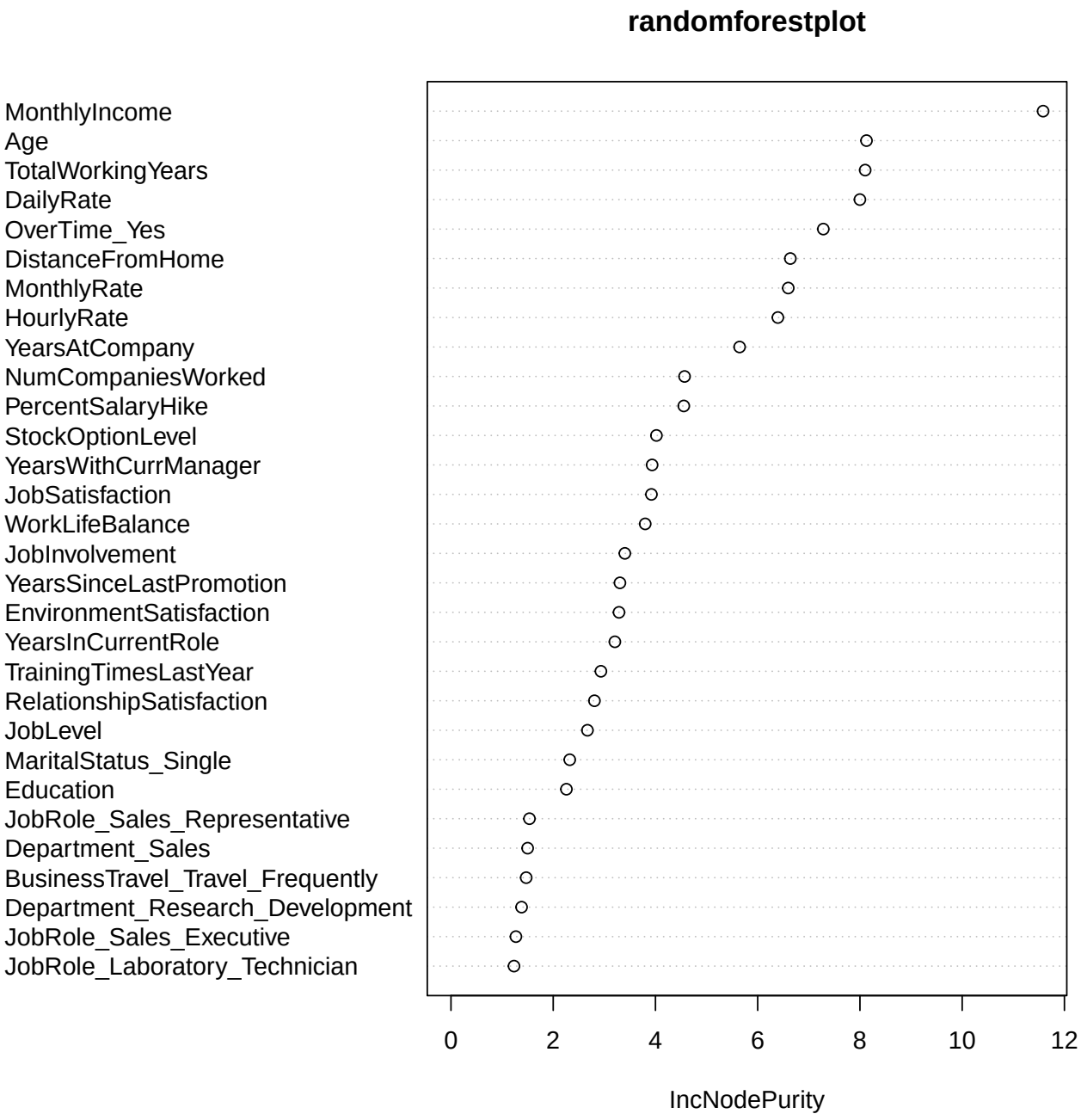
```
randomforestplot
```

```
##
## Call:
##  randomForest(formula = rf, data = employee_train, ntree = 500,      mtry = 8, do.trace = F)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 8
##
##           Mean of squared residuals: 0.104454
##                     % Var explained: 21.06
```

```r
plot(randomforestplot)
```




```r
## We can check which variables are most predictive using a variable importance plot
varImpPlot(randomforestplot)
```




```
# find train MSE
yhat_rf_train <- predict(randomforestplot, employee_train)
mse_rf_train <- mean((yhat_rf_train - y_train) ^ 2)
print(mse_rf_train)
```

```
## [1] 0.02398722
```

```
# find test MSE
yhat_rf_test <- predict(randomforestplot, employee_test)
mse_rf_test <- mean((yhat_rf_test - y_test) ^ 2)
print(mse_rf_test)
```

```
## [1] 0.1145158
```

Boosting

```
fit_btree <- gbm(rf,
                 data = employee_train,
                 distribution = "gaussian",
                 n.trees = 500,
                 interaction.depth = 6,
                 shrinkage = 0.001, cv.folds = 10)

## We can check which variables are most predictive as follows
relative.influence(fit_btree)
```

```
## n.trees not given. Using 500 trees.
```

```
##                              Age                        DailyRate
##                       295.593762                       150.382581
##                 DistanceFromHome                        Education
##                       222.467799                        16.624899
##          EnvironmentSatisfaction                       HourlyRate
##                        90.442570                       130.952893
##                   JobInvolvement                         JobLevel
##                       110.653925                       171.494067
##                  JobSatisfaction                    MonthlyIncome
##                       204.363537                      1287.267258
##                      MonthlyRate               NumCompaniesWorked
##                        91.657311                       220.281706
##                PercentSalaryHike                PerformanceRating
##                        45.352131                         0.000000
##         RelationshipSatisfaction                 StockOptionLevel
##                        32.444283                       355.748555
##                TotalWorkingYears            TrainingTimesLastYear
##                       965.229668                        12.617834
##                  WorkLifeBalance                   YearsAtCompany
##                       104.502166                       236.176043
##               YearsInCurrentRole          YearsSinceLastPromotion
##                        34.255374                        61.169509
##             YearsWithCurrManager BusinessTravel_Travel_Frequently
##                       178.452701                        25.446792
##     BusinessTravel_Travel_Rarely  Department_Research_Development
##                         2.919044                        69.253709
##                 Department_Sales     EducationField_Life_Sciences
##                       178.148984                         0.979363
##         EducationField_Marketing           EducationField_Medical
##                        15.056427                        59.524900
##             EducationField_Other  EducationField_Technical_Degree
##                         0.000000                        24.050442
##                      Gender_Male          JobRole_Human_Resources
##                        29.373460                         0.000000
##    JobRole_Laboratory_Technician                   JobRole_Manager
##                        18.800351                         0.000000
##   JobRole_Manufacturing_Director        JobRole_Research_Director
##                         0.000000                         0.000000
##       JobRole_Research_Scientist          JobRole_Sales_Executive
##                        98.127239                       140.252224
```

```
##     JobRole_Sales_Representative            MaritalStatus_Married
##                        91.451647                        27.657505
##             MaritalStatus_Single                      OverTime_Yes
##                       429.466171                      1358.049478
```

```
# find train MSE
yhat_btree_train<- predict(fit_btree, employee_train, n.trees = 100)
mse_btree_train <- mean((yhat_btree_train - y_train) ^ 2)
print(mse_btree_train)
```

```
## [1] 0.12761
```

```
# find test MSE
yhat_btree_test<- predict(fit_btree, employee_test, n.trees = 100)
mse_btree_test <- mean((yhat_btree_test - y_test) ^ 2)
print(mse_btree_test)
```

```
## [1] 0.1422179
```

Conclusion

1. Lasso Regression Model is the best ML model for our dataset.
2. According to our model, we believe that overtime, marital status(single), job role as sales representatives, age and monthly income are the factors that impact the employee attrition the most. Some other minor factors may include total working years, job role as lab technician and frequent business travels.