projecthing

Michael Chen

2025-05-15

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

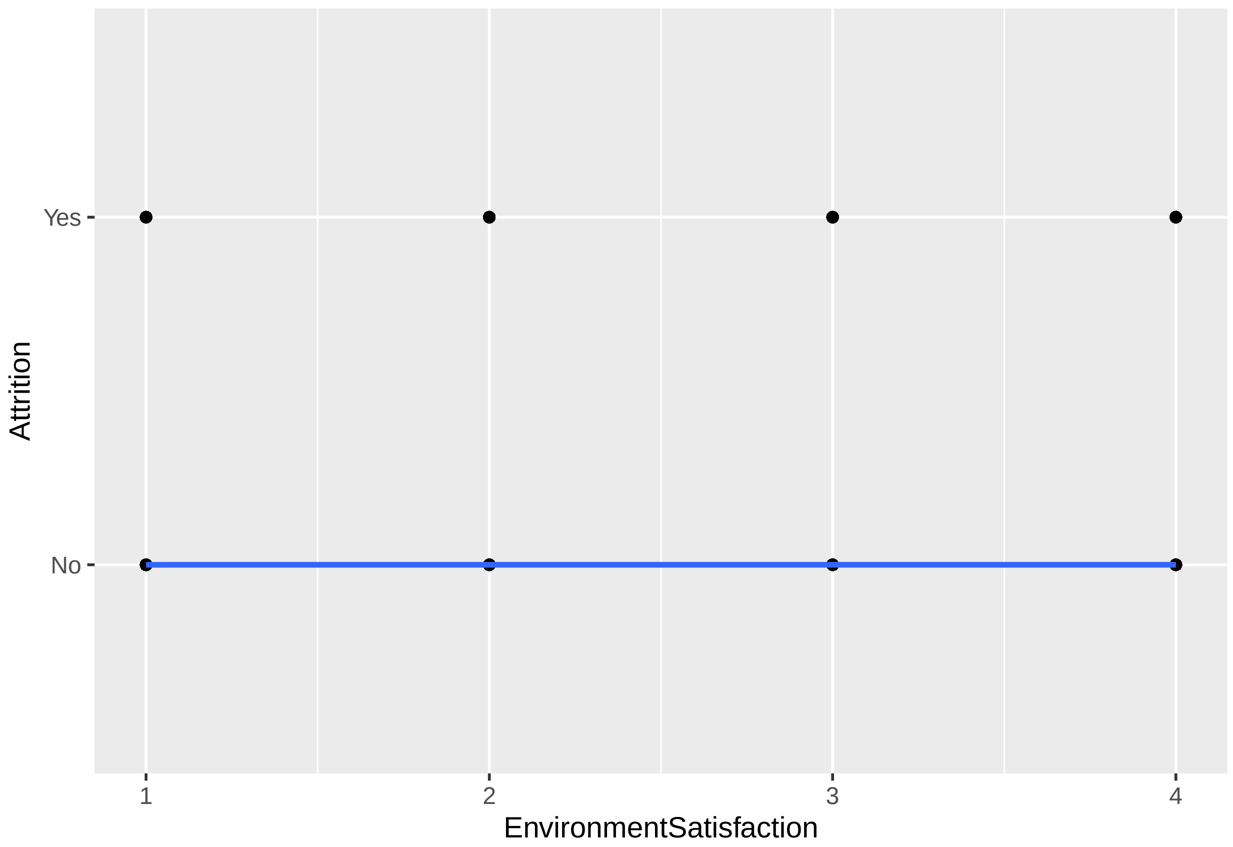
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
# binary response variable is 1, and 0 where 1 is yes and 0 is no for attrition.

pacman::p_load(tidyverse, MASS, car)

Employee_Attrition <- read.csv("/Users/michaelchen/downloads/HR Employee Attrition.csv")



ggplot(Employee_Attrition, aes(EnvironmentSatisfaction,Attrition )) + geom_point() +
geom_smooth(method = "glm", method.args = list(family = "binomial"), se=F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: glm.fit: algorithm did not converge
```

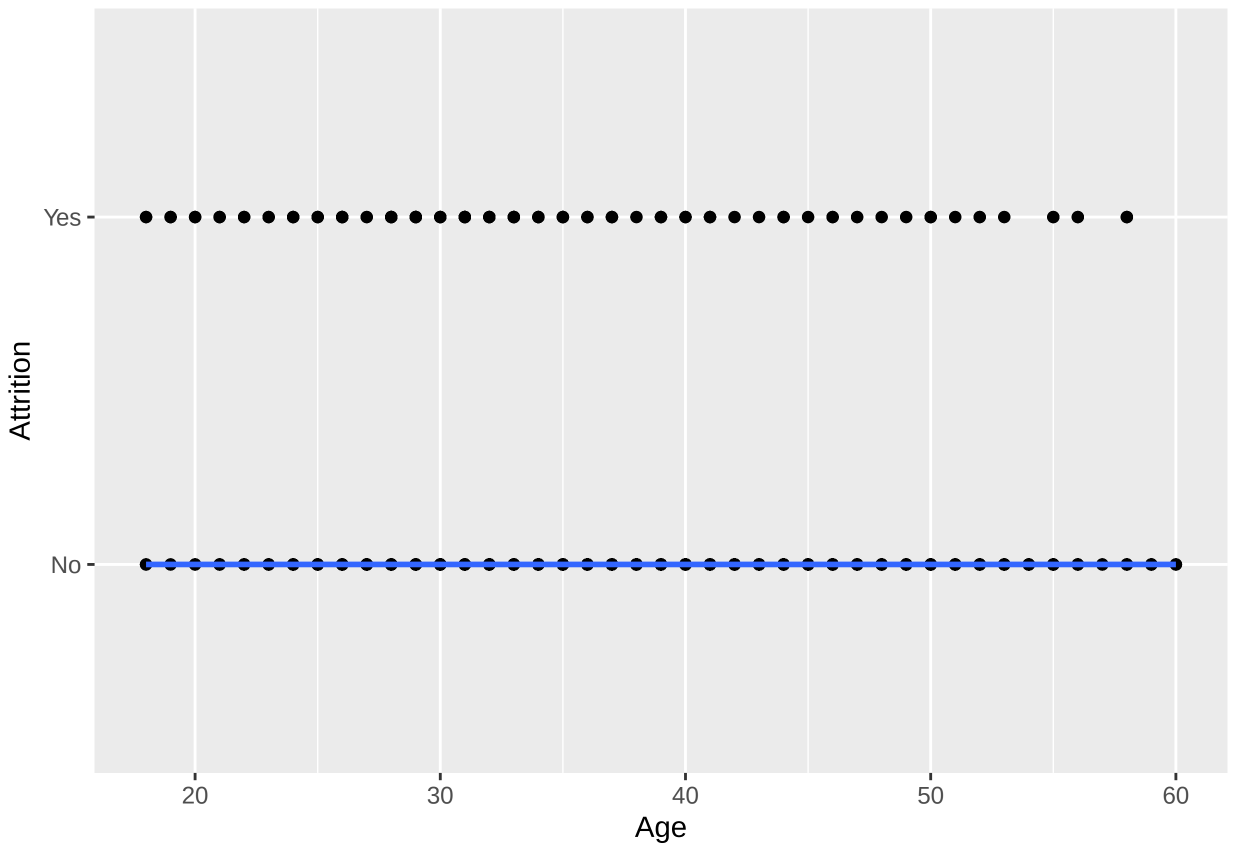
```
## Warning: Failed to fit group 2.
## Caused by error:
## ! y values must be 0 <= y <= 1
```



```
ggplot(Employee_Attrition, aes(Age,Attrition )) + geom_point() +
geom_smooth(method = "glm", method.args = list(family = "binomial"), se=F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

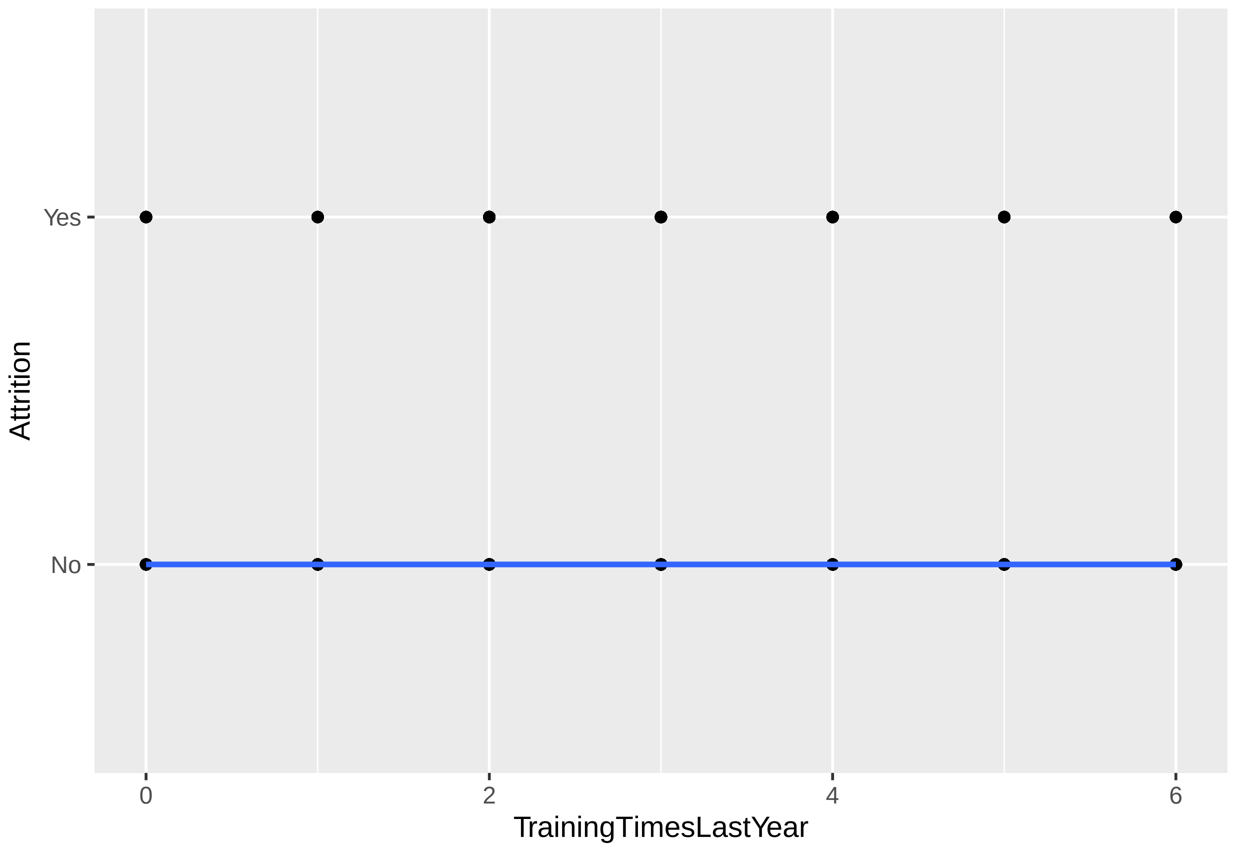
```
## Warning: glm.fit: algorithm did not converge
## Warning: Failed to fit group 2.
```



```
 ggplot(Employee_Attrition, aes(TrainingTimesLastYear,Attrition )) + geom_point() +
geom_smooth(method = "glm", method.args = list(family = "binomial"), se=F)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

```
## Warning: glm.fit: algorithm did not converge
## Warning: Failed to fit group 2.
```



```
Employee_Attrition$Attrition <- ifelse(Employee_Attrition$Attrition == "Yes",1,0)

Employee_Attrition$BusinessTravel <- as.factor(Employee_Attrition$BusinessTravel)

Employee_Attrition$Department <- as.factor(Employee_Attrition$Department)

Employee_Attrition$EducationField <- as.factor(Employee_Attrition$EducationField)

Employee_Attrition$Gender <- as.factor(Employee_Attrition$Gender)

Employee_Attrition$JobRole <- as.factor(Employee_Attrition$JobRole)

Employee_Attrition$MaritalStatus <- as.factor(Employee_Attrition$MaritalStatus)

Employee_Attrition$Over18 <- as.factor(Employee_Attrition$Over18)

Employee_Attrition$OverTime<- as.factor(Employee_Attrition$OverTime)

## dim(Employee_Attrition)

glm1 <- glm(Attrition ~ Age + BusinessTravel  +  DistanceFromHome +
            EducationField + EnvironmentSatisfaction + Gender + HourlyRate + JobInvolvement +
            JobLevel + JobRole + JobSatisfaction + MaritalStatus + MonthlyIncome +
            MonthlyRate + NumCompaniesWorked + OverTime + PercentSalaryHike +
            PerformanceRating + RelationshipSatisfaction + StockOptionLevel +
            TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
            YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
            YearsWithCurrManager,
            data = Employee_Attrition, family = binomial)

library(rpart)

glm3 <- glm(Attrition ~ Age + BusinessTravel + DistanceFromHome + EnvironmentSatisfaction + JobInvolveme
    OverTime + RelationshipSatisfaction +
    TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
    YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
    YearsWithCurrManager, family = binomial, data = Employee_Attrition)
glm4 <- step(glm3, trace=0)

glm4 <- step(glm3, trace=0)

set.seed(999)

n <- nrow(Employee_Attrition); n
```

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

```
floor(0.7*n) # 70% of data used for training, 30 % is used for prediction
```

```
## [1] 1029
```

```
train <- sample(1:n,1029)

glm_train <-glm ( Attrition ~ Age + BusinessTravel + DistanceFromHome +
    EnvironmentSatisfaction + JobInvolvement + JobSatisfaction +
    NumCompaniesWorked + OverTime + RelationshipSatisfaction +
    TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
    YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
    YearsWithCurrManager,data=Employee_Attrition,subset= train, family= binomial)

Employee_Attrition_test<- Employee_Attrition[-train, ]

probs_test <- predict(glm_train, newdata = Employee_Attrition_test,
type = "response")


length(probs_test)
```

```
## [1] 441
```

```
preds_test <- rep(0, 441)

preds_test[probs_test > 0.5] <- 1

head(probs_test)
```

```
##          1         12         15         16         18         20
## 0.47342759 0.12418196 0.63345144 0.09478628 0.12055382 0.12325812
```

```
tb <- table(prediction = preds_test,
actual = Employee_Attrition_test$Attrition)

addmargins(tb)
```

```
##           actual
## prediction   0   1 Sum
##        0   364  39 403
##        1    12  26  38
##        Sum 376  65 441
```

```
(tb[1,1] + tb[2,2]) / 441
```

```
## [1] 0.8843537
```

```
tb[2,2] / 70
```

```
## [1] 0.3714286
```

```
tb[1,1] / 371
```

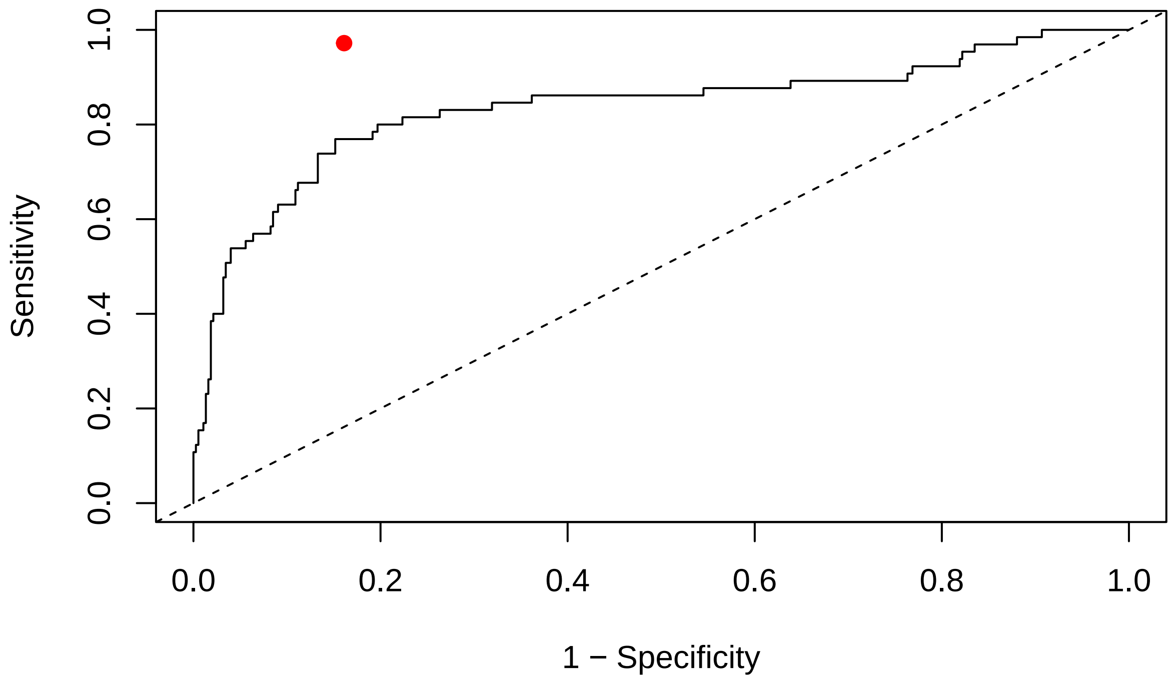
```
## [1] 0.9811321
```

```
roc_obj <- roc(Employee_Attrition_test$Attrition, probs_test)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(1 - roc_obj$specificities, roc_obj$sensitivities, type="l",
xlab = "1 - Specificity", ylab = "Sensitivity")
# plot red point corresponding to 0.5 threshold:
 points(x = 24/149, y = 763/785, col="red", pch=19)
 abline(0, 1, lty=2) # 1-1 line
```



```
auc(roc_obj)
```

```
## Area under the curve: 0.8346
```

```
library(randomForest)
```

```
## randomForest 4.7-1.2
```

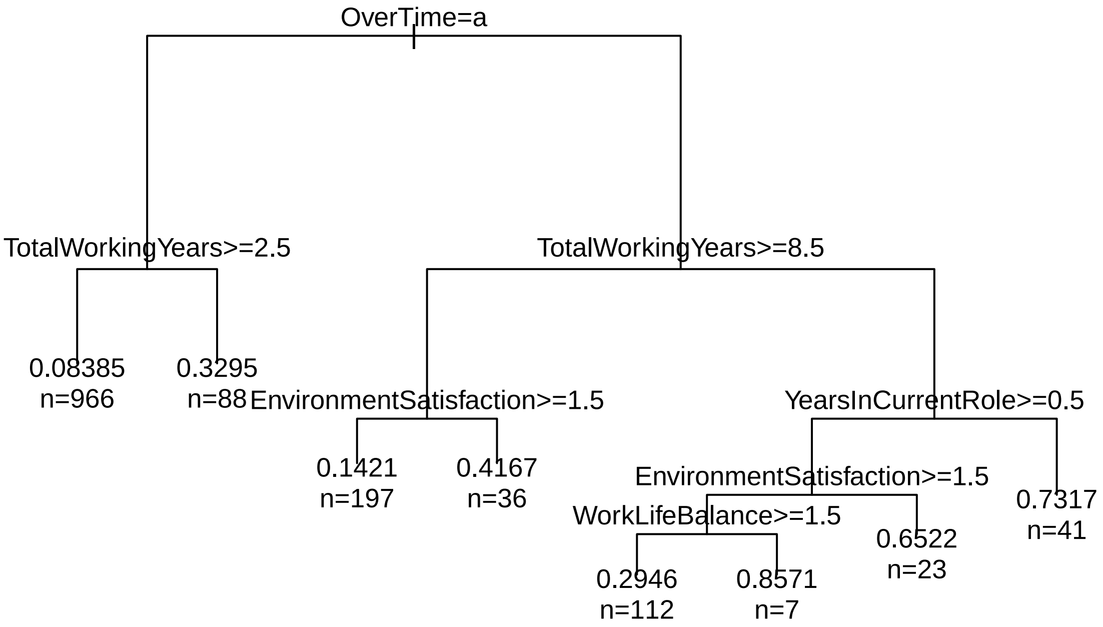
```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

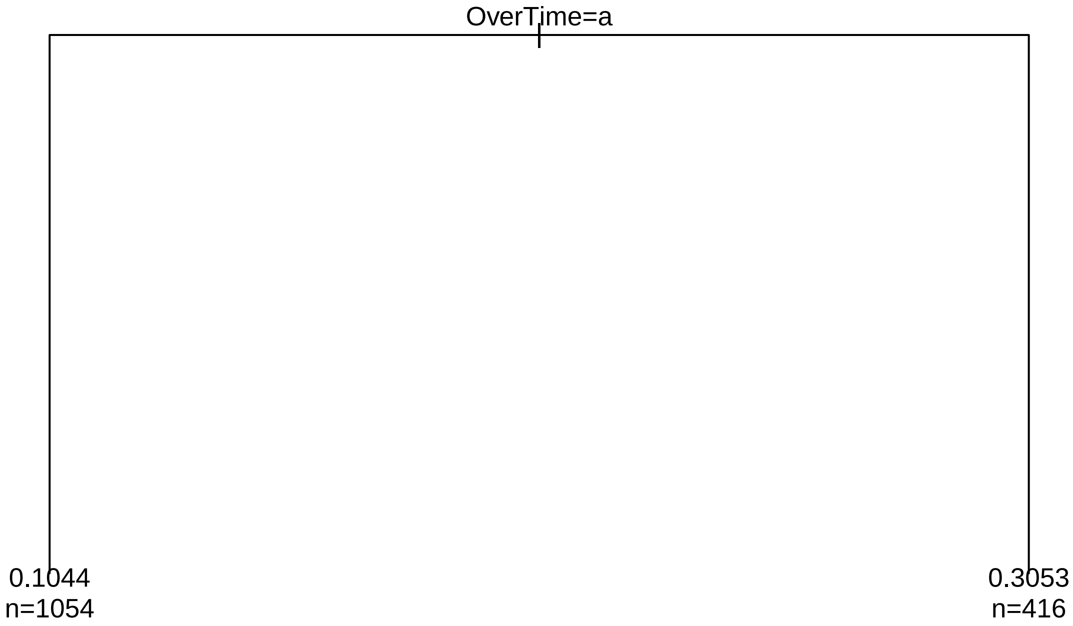
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

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

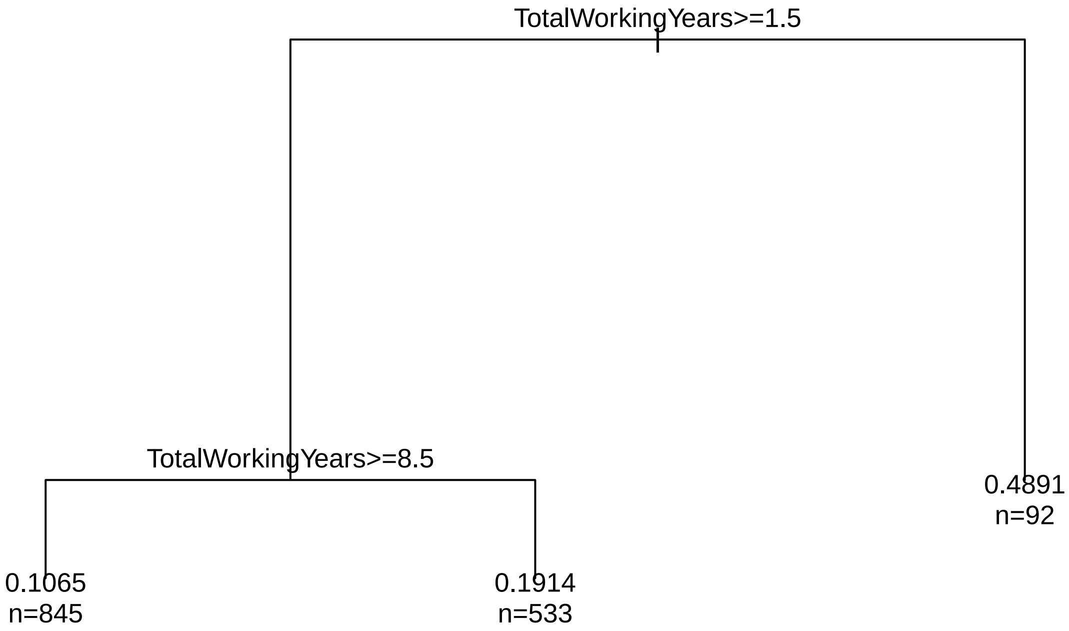
```
set.seed(999)


Employee_Attrition1 <- factor(Employee_Attrition$Attrition,
levels = c(1,0), labels=c("Yes", "No"))

#table(Employee_Attrition1)

t1 <- rpart(Attrition ~  Age  + EnvironmentSatisfaction + JobInvolvement + JobSatisfaction  +
    OverTime  +
    TotalWorkingYears + WorkLifeBalance +
    YearsAtCompany + YearsInCurrentRole, data=Employee_Attrition)
 par(cex=0.8, xpd=NA)
 plot(t1)

 text(t1, use.n=T)
```



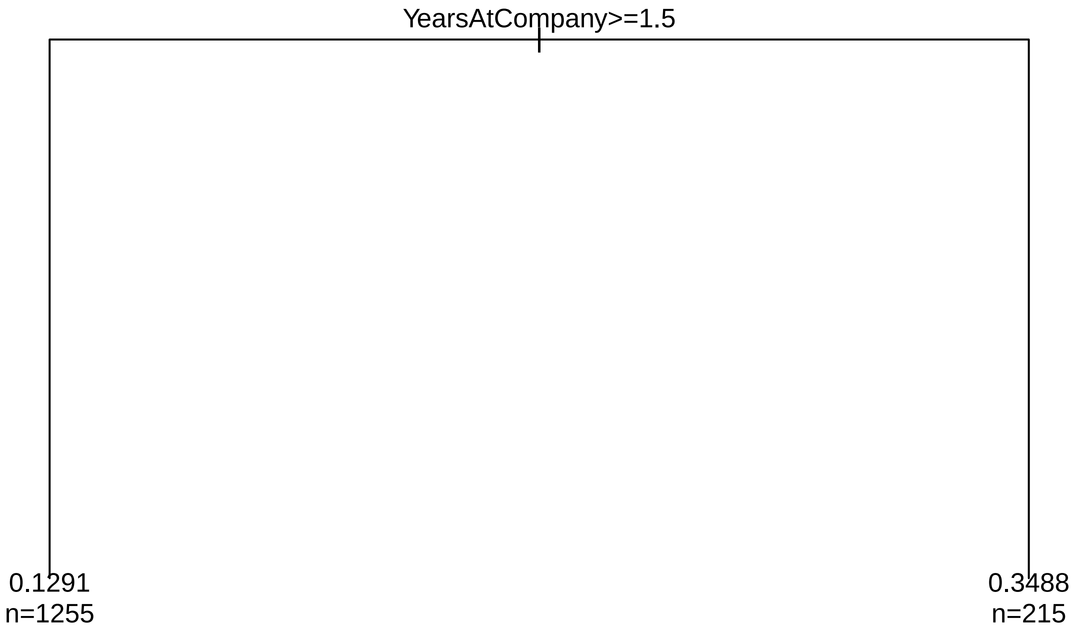
```
#Overtime seems to be very important, left is yes, right is no
t2 <- rpart(Attrition ~
    OverTime, data=Employee_Attrition)
 par(cex=0.8, xpd=NA)
 plot(t2)

 text(t2, use.n=T)
```



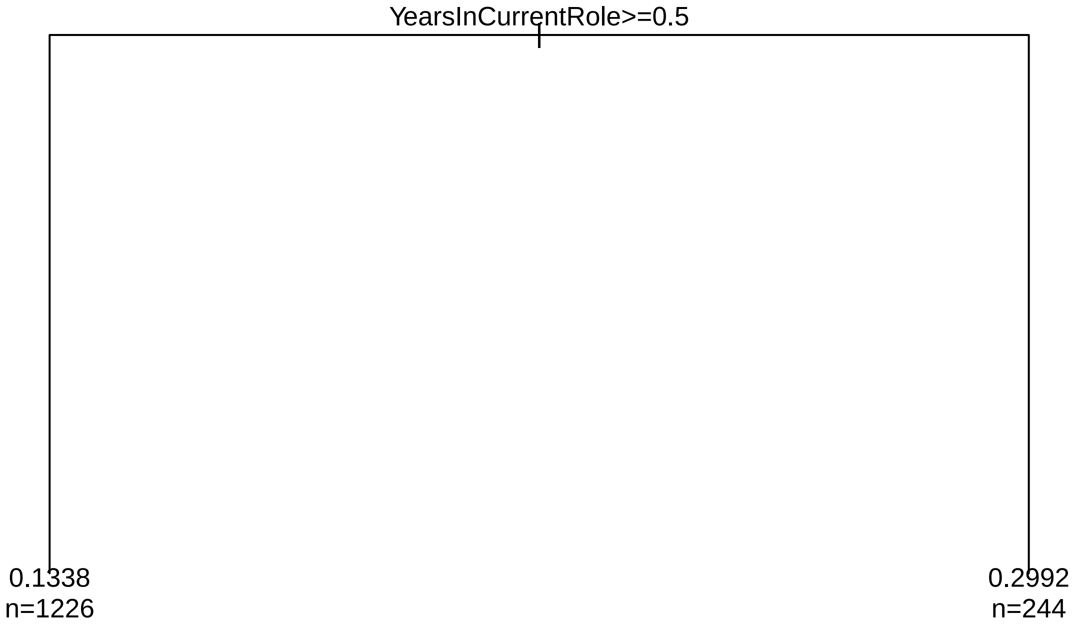
```
#Relationship Satisfaction not important
t3<- rpart(Attrition ~
   TotalWorkingYears, data=Employee_Attrition)
par(cex=0.8, xpd=NA)
plot(t3)

text(t3, use.n=T)
```



```
 # NumCompaniesWorked not important

 # WorkLifeBalance not important

t4<- rpart(Attrition ~
    YearsAtCompany, data=Employee_Attrition)
 par(cex=0.8, xpd=NA)
 plot(t4)

 text(t4, use.n=T)
```



```
t5<- rpart(Attrition ~
    YearsInCurrentRole, data=Employee_Attrition)
 par(cex=0.8, xpd=NA)
 plot(t5)

 text(t5, use.n=T)
```



```
# YearsSinceLastPromotion not important
```