HR Attrition Analysis

Veerasak Kritsanapraphan

4/25/2018

HR Attrition Analysis

```
hremployee <- read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv",
                       header=TRUE)
str(hremployee)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 2 3 2 3 2 3 3
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : int  2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : int  3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1
##  $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : int  3 4 3 3 3 3 4 4 4 3 ...
##  $ RelationshipSatisfaction: int  1 4 2 3 4 3 1 2 2 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : int  1 3 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

```
table(hremployee$Attrition)
```

```
##
##   No  Yes
```

```
## 1233  237
```

Sampling

```
library(fifer)
```

```
## Loading required package: MASS
```

```
yeshr <- stratified(hremployee, "Attrition", 230,
          select = list(Attrition = c("Yes")))
nohr <- stratified(hremployee, "Attrition", 230,
                   select = list(Attrition = c("No")))
hrsample <- rbind(yeshr,nohr)

set.seed(123)
ind <- sample(2, nrow(hrsample), replace=TRUE,
              prob=c(0.6,0.4))
trainData <- hrsample[ind==1,]
testData <- hrsample[ind==2,]
str(hrsample)
```

```
## 'data.frame':    460 obs. of  35 variables:
##  $ Age                     : int  26 30 34 28 58 37 40 45 26 22 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 1 3 3 3 3 3
##  $ DailyRate               : int  1330 740 967 1157 147 370 676 1449 1449 1368 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 2 3 2 2 2 2 2 3 2 2 ...
##  $ DistanceFromHome        : int  21 1 16 2 23 10 9 2 16 4 ...
##  $ Education               : int  3 3 4 4 4 4 4 3 4 1 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 4 2 6 4 4 4 2 3 4 6 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  1107 1562 1905 440 165 1809 1534 1277 394 593 ...
##  $ EnvironmentSatisfaction : int  1 2 4 1 4 4 4 1 1 3 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 2 2 1 2 2 ...
##  $ HourlyRate              : int  37 64 85 84 94 58 86 94 45 99 ...
##  $ JobInvolvement          : int  3 2 1 1 3 3 3 1 3 2 ...
##  $ JobLevel                : int  1 2 1 1 3 2 1 5 1 1 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 3 8 7 7 1 5 3 4 3 3
##  $ JobSatisfaction         : int  3 1 1 4 4 1 1 2 2 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 1 2 2 2 2 3 3 3 1 3 ...
##  $ MonthlyIncome           : int  2377 9714 2307 3464 10312 4213 2018 18824 2373 3894 ...
##  $ MonthlyRate             : int  19373 5323 14460 24737 3465 4992 21831 2493 14180 9129 ...
##  $ NumCompaniesWorked      : int  1 1 1 5 1 1 3 2 2 5 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 1 1 2 2 1 ...
##  $ PercentSalaryHike       : int  20 11 23 13 12 15 14 16 13 16 ...
##  $ PerformanceRating       : int  4 3 4 3 3 3 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  3 4 2 4 4 2 2 1 4 3 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  1 1 1 0 1 0 0 0 1 0 ...
##  $ TotalWorkingYears       : int  1 10 5 5 40 10 15 26 5 4 ...
##  $ TrainingTimesLastYear   : int  0 4 2 4 3 4 3 2 2 3 ...
##  $ WorkLifeBalance         : int  2 3 3 2 2 1 1 3 3 3 ...
##  $ YearsAtCompany          : int  1 10 5 3 40 10 5 24 3 2 ...
```

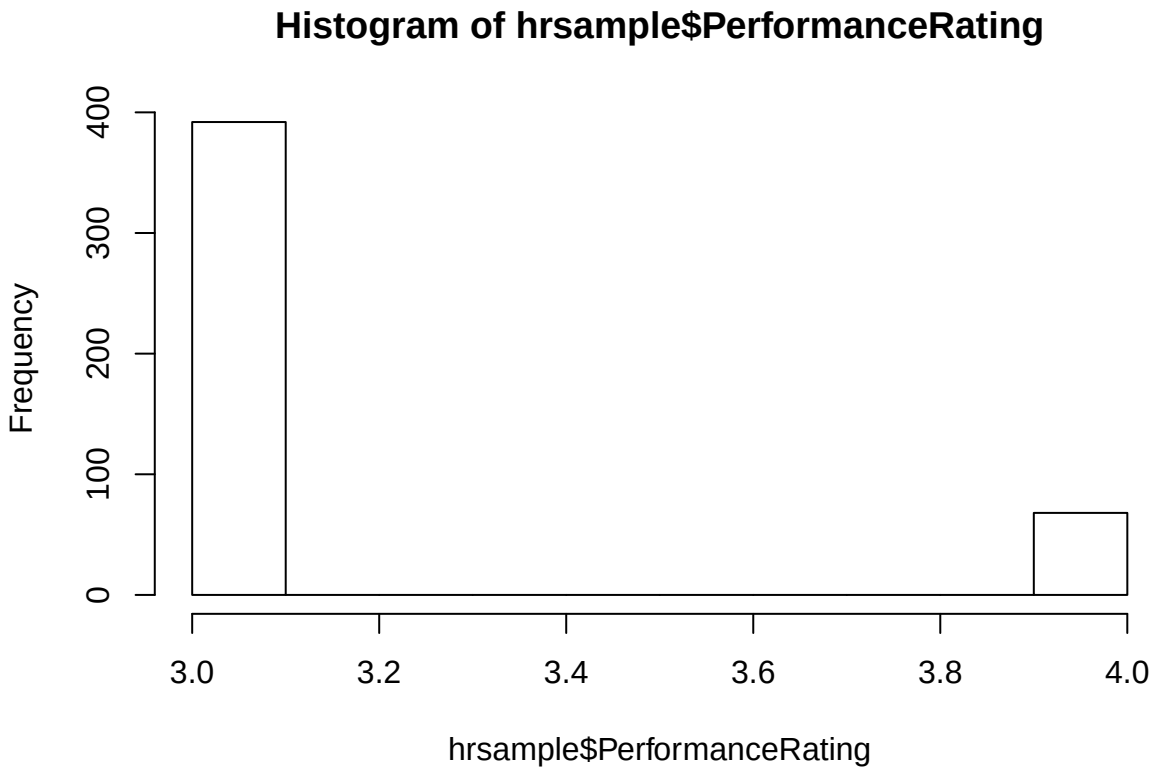
```
##  $ YearsInCurrentRole      : int  1 8 2 2 10 3 4 10 2 2 ...
##  $ YearsSinceLastPromotion : int  0 6 3 2 15 0 1 1 0 1 ...
##  $ YearsWithCurrManager    : int  0 7 0 2 6 8 0 11 2 2 ...
```

```
table(hrsample$JobSatisfaction)
```

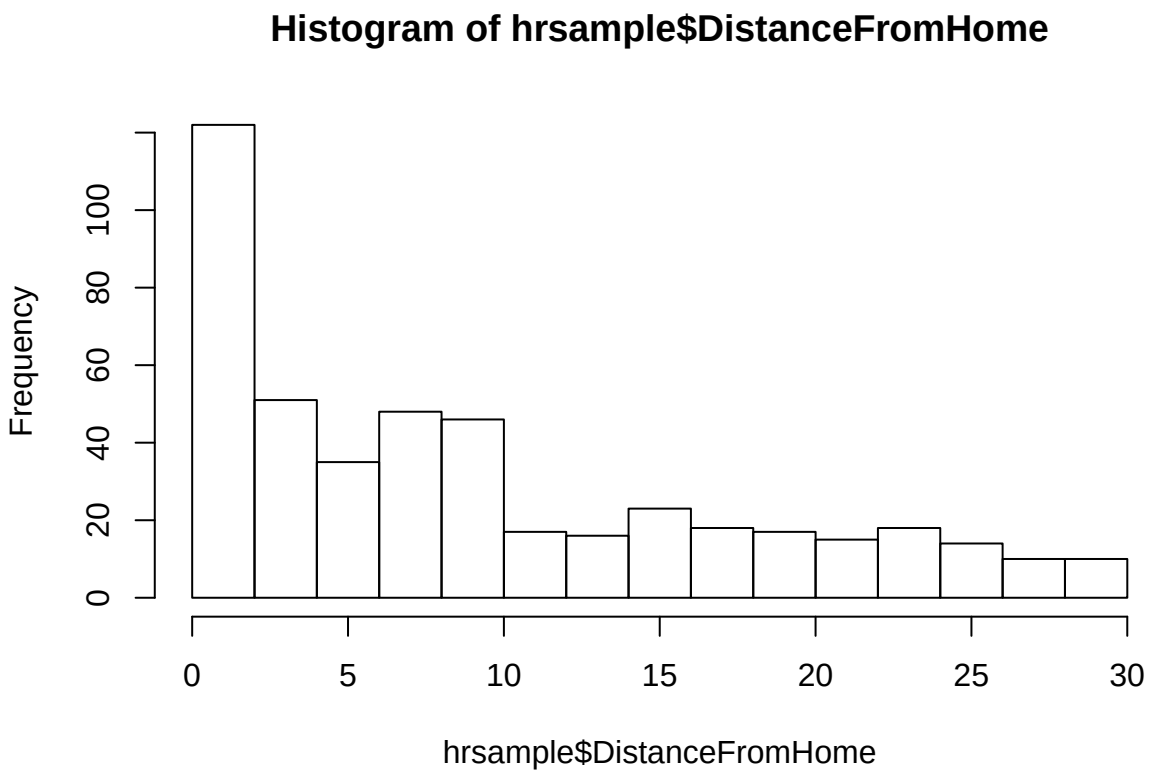
```
##
##   1   2   3   4
## 110  81 142 127
```

```
hist(hrsample$Age)
```

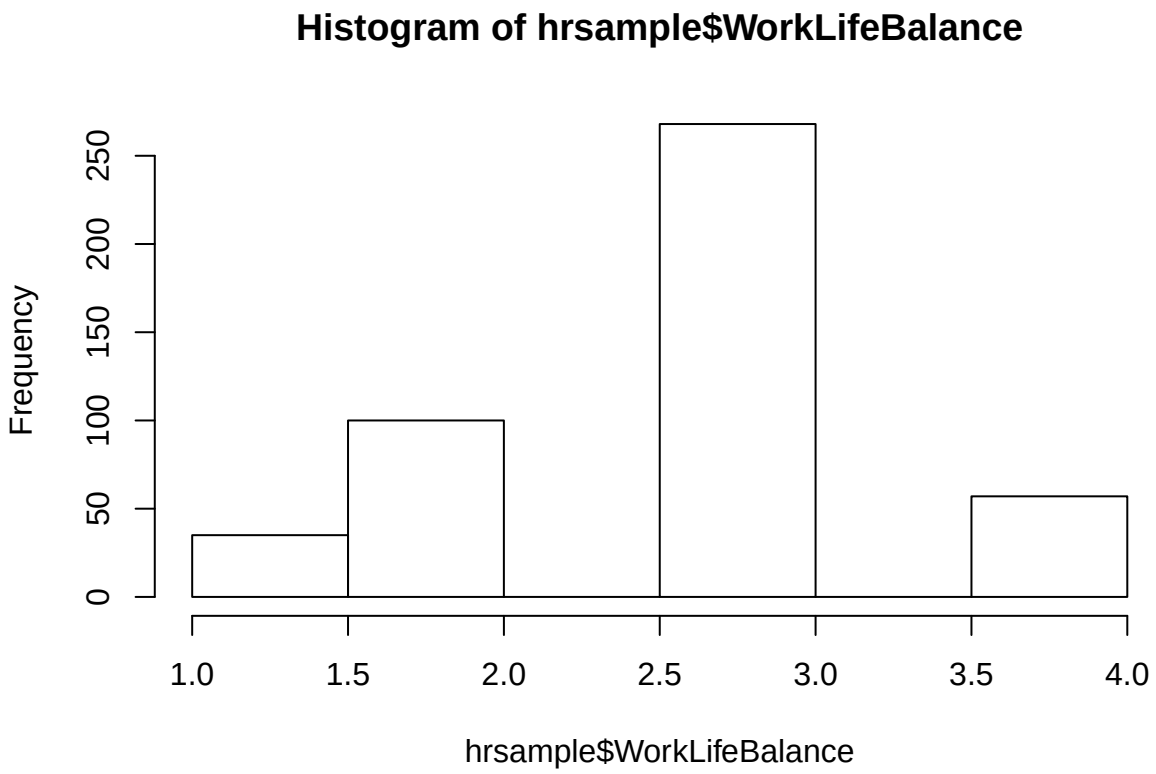

```
hist(hrsample$PerformanceRating)
```



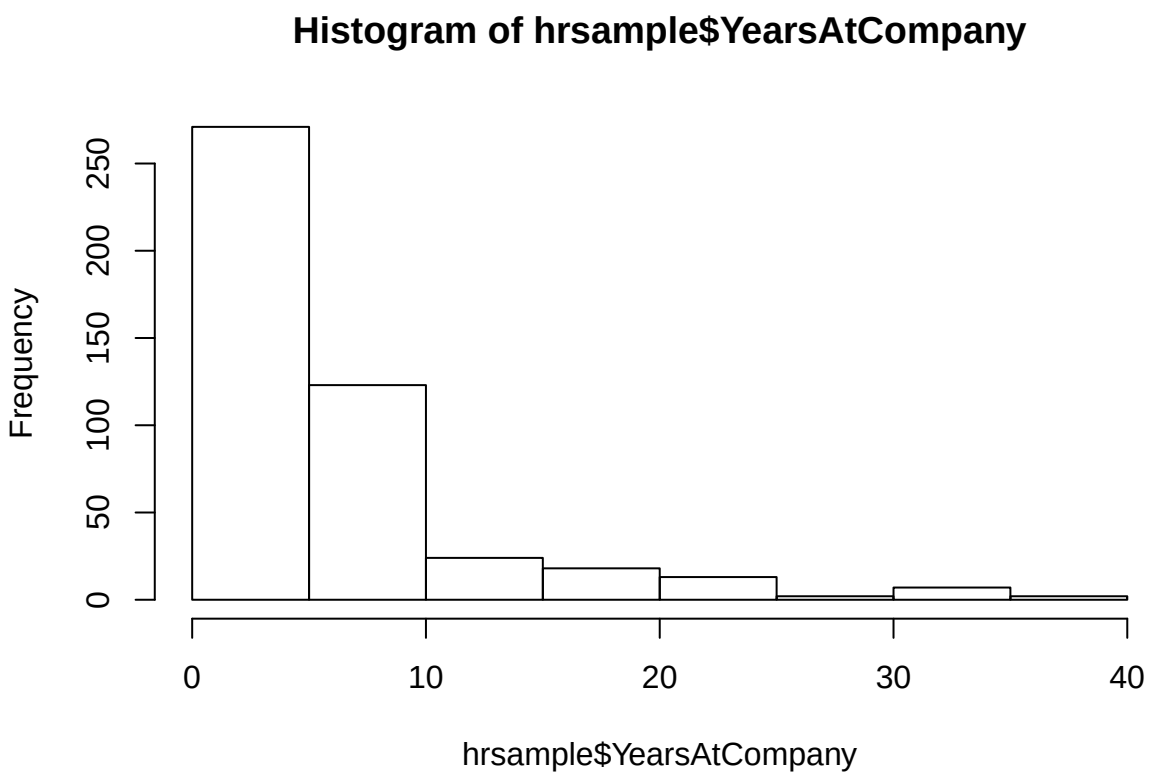

```
hist(hrsample$DistanceFromHome)
```




```
hist(hrsample$WorkLifeBalance)
```




```r
hist(hrsample$YearsAtCompany)
```




```r
#myformula <- Attrition ~ .
myformula <- Attrition ~ JobSatisfaction +
                         Age + PerformanceRating +
```

```
                        DistanceFromHome +
                        WorkLifeBalance +
                        YearsAtCompany
table(trainData$Attrition)
```

```
##
##  No Yes
## 143 140
```

```
table(testData$Attrition)
```

```
##
##  No Yes
##  87  90
```

```
library(party)
```

```
## Loading required package: grid
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
## Loading required package: strucchange
```

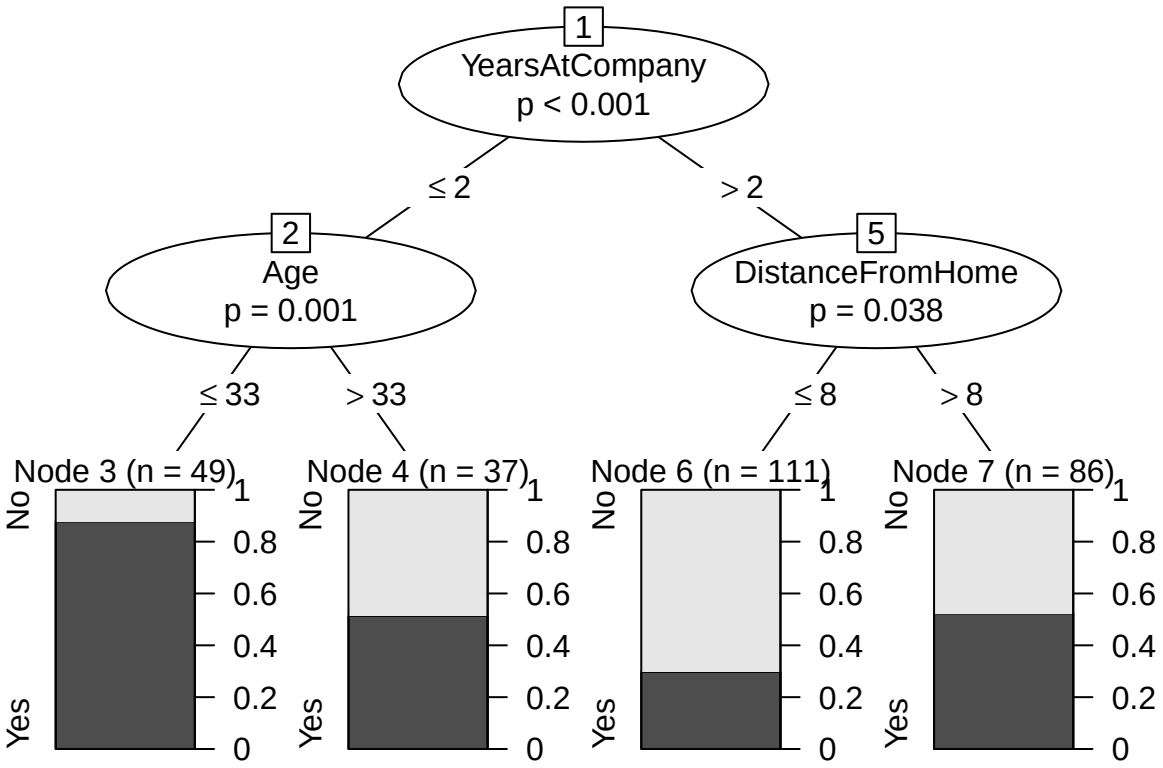
```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
ctree_model <- ctree(myformula, data=trainData)
plot(ctree_model)
```




```
testpred <- predict(ctree_model,newdata=testData)
table(testpred,
      testData$Attrition)
```

```
##
## testpred No Yes
##      No  45  34
##      Yes 42  56
```

```
library(caret)
```

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

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

```
confusionMatrix(testpred, testData$Attrition)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  45  34
##        Yes 42  56
##
##                Accuracy : 0.5706
##                  95% CI : (0.4942, 0.6446)
##     No Information Rate : 0.5085
##     P-Value [Acc > NIR] : 0.057
##
##                   Kappa : 0.1397
##
##  Mcnemar's Test P-Value : 0.422
```

```
##
##             Sensitivity : 0.5172
##             Specificity : 0.6222
##          Pos Pred Value : 0.5696
##          Neg Pred Value : 0.5714
##              Prevalence : 0.4915
##          Detection Rate : 0.2542
##    Detection Prevalence : 0.4463
##       Balanced Accuracy : 0.5697
##
##        'Positive' Class : No
##
```

```
myformula <- Attrition ~ .
table(trainData$Attrition)
```

```
##
##  No Yes
## 143 140
```

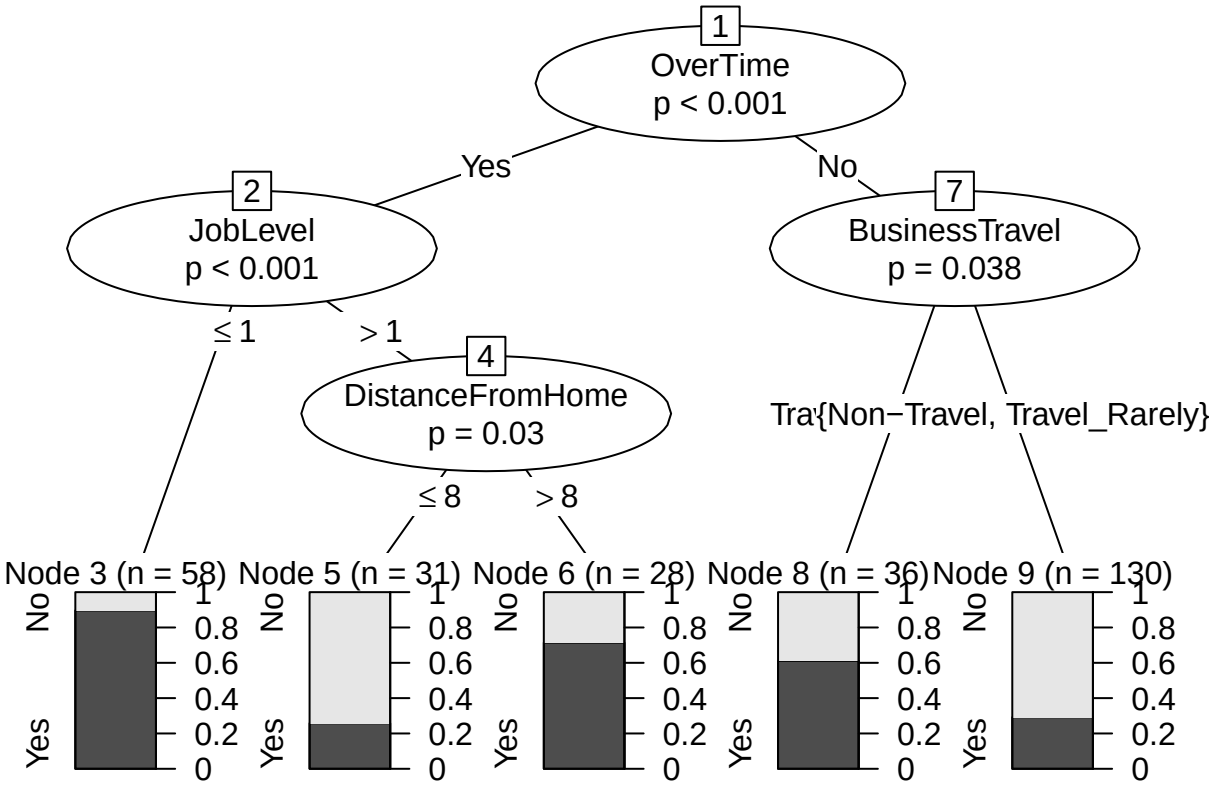
```
table(testData$Attrition)
```

```
##
##  No Yes
##  87  90
```

```
library(party)
ctree_model <- ctree(myformula, data=trainData)
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```
plot(ctree_model)
```



```
testpred <- predict(ctree_model,newdata=testData)
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```
table(testpred,
      testData$Attrition)
```

```
##
## testpred No Yes
##      No  57  47
##      Yes 30  43
```

```
library(caret)

confusionMatrix(testpred, testData$Attrition)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  57  47
##        Yes 30  43
##
##                Accuracy : 0.565
##                  95% CI : (0.4885, 0.6392)
##     No Information Rate : 0.5085
##     P-Value [Acc > NIR] : 0.07643
##
##                   Kappa : 0.1325
##
##  Mcnemar's Test P-Value : 0.06825
##
##             Sensitivity : 0.6552
##             Specificity : 0.4778
##          Pos Pred Value : 0.5481
##          Neg Pred Value : 0.5890
##              Prevalence : 0.4915
##          Detection Rate : 0.3220
##    Detection Prevalence : 0.5876
##       Balanced Accuracy : 0.5665
##
##        'Positive' Class : No
##
```