Final\_project\_only\_OT.R

Jennifer

Mon May 08 13:19:08 2017

library(caret)

## Warning: package 'caret' was built under R version 3.3.3

## Loading required package: lattice

## Loading required package: ggplot2

library(mlbench)

## Warning: package 'mlbench' was built under R version 3.3.3

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.3.3

## randomForest 4.6-12

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

##   
## Attaching package: 'randomForest'

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

library(rpart)  
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 3.3.3

library(party)

## Warning: package 'party' was built under R version 3.3.3

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Warning: package 'strucchange' was built under R version 3.3.3

## Loading required package: zoo

##   
## Attaching package: 'zoo'

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

## Loading required package: sandwich

library(partykit)

## Warning: package 'partykit' was built under R version 3.3.3

##   
## Attaching package: 'partykit'

## The following objects are masked from 'package:party':  
##   
## cforest, ctree, ctree\_control, edge\_simple, mob, mob\_control,  
## node\_barplot, node\_bivplot, node\_boxplot, node\_inner,  
## node\_surv, node\_terminal

library(gmodels)

## Warning: package 'gmodels' was built under R version 3.3.3

library(e1071)

## Warning: package 'e1071' was built under R version 3.3.3

library(knitr)

## Warning: package 'knitr' was built under R version 3.3.3

options(scipen = 999)  
  
setwd("C:/Users/Jennifer/Documents/ADM/Final")  
HR<-read.csv("WatsonHRonlyOT.csv")  
HR<-HR[,c(2, 1, 3:22, 24:35)]  
HR<-HR[,c(1, 2:8, 11:21, 23:25, 27:34)]  
summary(HR)

## Attrition Age BusinessTravel DailyRate   
## No :289 Min. :18.00 Non-Travel : 35 Min. : 103.0   
## Yes:127 1st Qu.:30.00 Travel\_Frequently: 86 1st Qu.: 463.2   
## Median :36.00 Travel\_Rarely :295 Median : 799.0   
## Mean :37.33 Mean : 808.4   
## 3rd Qu.:44.00 3rd Qu.:1171.8   
## Max. :60.00 Max. :1498.0   
## Department DistanceFromHome Education   
## Human Resources : 17 Min. : 1.000 Min. :1.00   
## Research & Development:271 1st Qu.: 2.000 1st Qu.:2.00   
## Sales :128 Median : 7.000 Median :3.00   
## Mean : 9.522 Mean :2.88   
## 3rd Qu.:15.000 3rd Qu.:4.00   
## Max. :29.000 Max. :5.00   
## EducationField EnvironmentSatisfaction Gender   
## Human Resources : 8 Min. :1.000 Female:180   
## Life Sciences :167 1st Qu.:2.000 Male :236   
## Marketing : 48 Median :3.000   
## Medical :132 Mean :2.844   
## Other : 27 3rd Qu.:4.000   
## Technical Degree: 34 Max. :4.000   
## HourlyRate JobInvolvement JobLevel JobRole   
## Min. : 30.00 Min. :1.000 Min. :1.000 no :309   
## 1st Qu.: 48.75 1st Qu.:2.000 1st Qu.:1.000 yes:107   
## Median : 67.00 Median :3.000 Median :2.000   
## Mean : 65.64 Mean :2.726 Mean :2.065   
## 3rd Qu.: 82.00 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :100.00 Max. :4.000 Max. :5.000   
## JobSatisfaction MaritalStatus MonthlyIncome MonthlyRate   
## Min. :1.000 Divorced: 99 Min. : 1009 Min. : 2112   
## 1st Qu.:2.000 Married :186 1st Qu.: 2883 1st Qu.: 8443   
## Median :3.000 Single :131 Median : 5062 Median :14890   
## Mean :2.772 Mean : 6549 Mean :14556   
## 3rd Qu.:4.000 3rd Qu.: 8396 3rd Qu.:20745   
## Max. :4.000 Max. :19859 Max. :26999   
## NumCompaniesWorked PercentSalaryHike PerformanceRating  
## Min. :0.000 Min. :11.00 Min. :3.000   
## 1st Qu.:1.000 1st Qu.:12.00 1st Qu.:3.000   
## Median :2.000 Median :14.00 Median :3.000   
## Mean :2.611 Mean :15.18 Mean :3.156   
## 3rd Qu.:4.000 3rd Qu.:18.00 3rd Qu.:3.000   
## Max. :9.000 Max. :25.00 Max. :4.000   
## RelationshipSatisfaction StockOptionLevel TotalWorkingYears  
## Min. :1.000 Min. :0.0000 Min. : 0.00   
## 1st Qu.:2.000 1st Qu.:0.0000 1st Qu.: 6.00   
## Median :3.000 Median :1.0000 Median : 9.00   
## Mean :2.796 Mean :0.7933 Mean :11.44   
## 3rd Qu.:4.000 3rd Qu.:1.0000 3rd Qu.:16.00   
## Max. :4.000 Max. :3.0000 Max. :40.00   
## TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole  
## Min. :0.000 Min. :1.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.: 2.750 1st Qu.: 2.000   
## Median :2.000 Median :3.000 Median : 5.000 Median : 3.000   
## Mean :2.637 Mean :2.731 Mean : 6.894 Mean : 4.058   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.: 9.000 3rd Qu.: 7.000   
## Max. :6.000 Max. :4.000 Max. :36.000 Max. :16.000   
## YearsSinceLastPromotion YearsWithCurrManager  
## Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.: 2.000   
## Median : 1.000 Median : 3.000   
## Mean : 2.125 Mean : 3.887   
## 3rd Qu.: 2.000 3rd Qu.: 7.000   
## Max. :15.000 Max. :17.000

dim(HR)

## [1] 416 30

str(HR)

## 'data.frame': 416 obs. of 30 variables:  
## $ Attrition : Factor w/ 2 levels "No","Yes": 2 2 1 1 1 2 1 1 1 2 ...  
## $ Age : int 41 37 33 59 29 28 32 22 38 32 ...  
## $ BusinessTravel : Factor w/ 3 levels "Non-Travel","Travel\_Frequently",..: 3 3 2 3 3 3 3 1 3 2 ...  
## $ DailyRate : int 1102 1373 1392 1324 153 103 334 1123 371 1125 ...  
## $ Department : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...  
## $ DistanceFromHome : int 1 2 3 3 15 24 5 16 2 16 ...  
## $ Education : int 2 2 4 3 2 3 2 2 3 1 ...  
## $ EducationField : Factor w/ 6 levels "Human Resources",..: 2 5 2 4 2 2 2 4 2 2 ...  
## $ EnvironmentSatisfaction : int 2 4 4 3 4 3 1 4 4 2 ...  
## $ Gender : Factor w/ 2 levels "Female","Male": 1 2 1 1 1 2 2 2 2 1 ...  
## $ HourlyRate : int 94 92 56 81 49 50 80 96 45 72 ...  
## $ JobInvolvement : int 3 2 3 4 2 2 4 4 3 1 ...  
## $ JobLevel : int 2 1 1 1 2 1 1 1 1 1 ...  
## $ JobRole : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ JobSatisfaction : int 4 3 3 1 3 3 2 4 4 1 ...  
## $ MaritalStatus : Factor w/ 3 levels "Divorced","Married",..: 3 3 2 2 3 3 1 1 3 3 ...  
## $ MonthlyIncome : int 5993 2090 2909 2670 4193 2028 3298 2935 3944 3919 ...  
## $ MonthlyRate : int 19479 2396 23159 9964 12682 12947 15053 7324 4306 4681 ...  
## $ NumCompaniesWorked : int 8 6 1 4 0 5 0 1 5 1 ...  
## $ PercentSalaryHike : int 11 15 11 20 12 14 12 13 11 22 ...  
## $ PerformanceRating : int 3 3 3 4 3 3 3 3 3 4 ...  
## $ RelationshipSatisfaction: int 1 2 3 1 4 2 4 2 3 2 ...  
## $ StockOptionLevel : int 0 0 0 3 0 0 2 2 0 0 ...  
## $ TotalWorkingYears : int 8 7 8 12 10 6 7 1 6 10 ...  
## $ TrainingTimesLastYear : int 0 3 3 3 3 4 5 2 3 5 ...  
## $ WorkLifeBalance : int 1 3 3 2 3 3 2 2 3 3 ...  
## $ YearsAtCompany : int 6 0 8 1 9 4 6 1 3 10 ...  
## $ YearsInCurrentRole : int 4 0 7 0 5 2 2 0 2 2 ...  
## $ YearsSinceLastPromotion : int 0 0 3 0 0 0 0 0 1 6 ...  
## $ YearsWithCurrManager : int 5 0 0 0 8 3 5 0 2 7 ...

set.seed(123)  
HR\_rand <- HR[order(runif(416)), ]  
416\*.8

## [1] 332.8

HR\_train <- HR\_rand[1:333,]  
HR\_test <- HR\_rand[334:416,]  
  
prop.table(table(HR\_train$Attrition))

##   
## No Yes   
## 0.6636637 0.3363363

prop.table(table(HR\_test$Attrition))

##   
## No Yes   
## 0.8192771 0.1807229

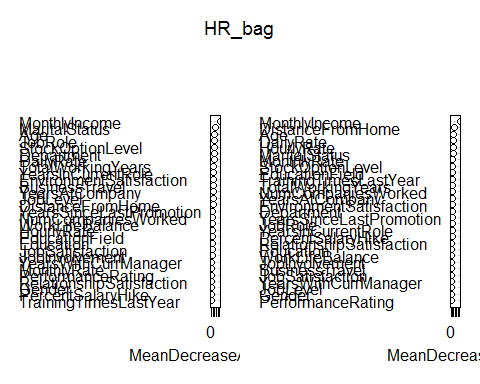
##################################################################  
#Random Forest  
  
set.seed(123)   
HR\_bag <- randomForest(Attrition~., data=HR\_train, mtry=29, na.action=na.omit, importance=TRUE)  
print(HR\_bag)

##   
## Call:  
## randomForest(formula = Attrition ~ ., data = HR\_train, mtry = 29, importance = TRUE, na.action = na.omit)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 29  
##   
## OOB estimate of error rate: 24.32%  
## Confusion matrix:  
## No Yes class.error  
## No 196 25 0.1131222  
## Yes 56 56 0.5000000

importance(HR\_bag)

## No Yes MeanDecreaseAccuracy  
## Age 6.0008771 5.72675760 8.59169623  
## BusinessTravel 4.9849185 3.66922792 5.71746376  
## DailyRate 8.4399145 1.58640781 7.24692240  
## Department 2.6469105 8.62502348 7.37234275  
## DistanceFromHome 3.5795684 2.71323328 4.33187206  
## Education -0.1419654 3.54584189 2.23282229  
## EducationField 2.5359177 0.79731246 2.52964021  
## EnvironmentSatisfaction 5.1851694 3.04116191 5.82779218  
## Gender -0.1161909 -1.03715817 -0.67738330  
## HourlyRate 4.1673446 -0.94757379 2.62362782  
## JobInvolvement 1.8678151 -0.21206975 1.38038779  
## JobLevel 3.6017229 3.79592149 5.48649293  
## JobRole 1.8069598 9.80100505 7.94316042  
## JobSatisfaction 1.5577519 1.04396386 1.75370389  
## MaritalStatus 9.0838398 8.14829933 11.22261511  
## MonthlyIncome 16.8558456 21.96342768 27.72252434  
## MonthlyRate -0.3807258 0.14062427 -0.06756628  
## NumCompaniesWorked 1.5373178 2.65691771 2.96181338  
## PercentSalaryHike 0.1454410 -1.98704459 -1.07763965  
## PerformanceRating -0.1681366 0.19251669 -0.09352125  
## RelationshipSatisfaction -1.2608909 1.19888389 -0.22204904  
## StockOptionLevel 5.9349051 5.64642117 7.79024842  
## TotalWorkingYears 9.1739164 -1.88582897 7.20374743  
## TrainingTimesLastYear -1.6941028 -0.60672685 -1.76890999  
## WorkLifeBalance 3.9948173 0.07062566 2.85448728  
## YearsAtCompany 5.0865551 1.94695554 5.62535813  
## YearsInCurrentRole 3.2484417 5.60791862 6.67510831  
## YearsSinceLastPromotion 3.6357806 1.46428234 3.68717074  
## YearsWithCurrManager -1.5987971 2.76943088 0.69479007  
## MeanDecreaseGini  
## Age 8.8035393  
## BusinessTravel 2.0942389  
## DailyRate 7.5006923  
## Department 3.8011627  
## DistanceFromHome 10.2760828  
## Education 2.6126641  
## EducationField 5.4158040  
## EnvironmentSatisfaction 3.8407018  
## Gender 0.6553069  
## HourlyRate 6.9391428  
## JobInvolvement 2.4610518  
## JobLevel 1.6715189  
## JobRole 3.1539549  
## JobSatisfaction 1.8963711  
## MaritalStatus 6.5390062  
## MonthlyIncome 33.0534057  
## MonthlyRate 6.2902677  
## NumCompaniesWorked 4.5326590  
## PercentSalaryHike 3.0832218  
## PerformanceRating 0.2273280  
## RelationshipSatisfaction 2.9508839  
## StockOptionLevel 5.7778468  
## TotalWorkingYears 4.6908392  
## TrainingTimesLastYear 5.0678665  
## WorkLifeBalance 2.5174463  
## YearsAtCompany 4.2470800  
## YearsInCurrentRole 3.0852916  
## YearsSinceLastPromotion 3.5111764  
## YearsWithCurrManager 1.8161215

varImpPlot(HR\_bag)



actualRF <- HR\_test$Attrition   
predictedRF <- predict(HR\_bag, HR\_test, type="class")   
HR\_bag\_matrix <- confusionMatrix(predictedRF, actualRF, positive="Yes")   
print("Bagged results")

## [1] "Bagged results"

print(HR\_bag\_matrix)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 55 5  
## Yes 13 10  
##   
## Accuracy : 0.7831   
## 95% CI : (0.6791, 0.8661)  
## No Information Rate : 0.8193   
## P-Value [Acc > NIR] : 0.84131   
##   
## Kappa : 0.3937   
## Mcnemar's Test P-Value : 0.09896   
##   
## Sensitivity : 0.6667   
## Specificity : 0.8088   
## Pos Pred Value : 0.4348   
## Neg Pred Value : 0.9167   
## Prevalence : 0.1807   
## Detection Rate : 0.1205   
## Detection Prevalence : 0.2771   
## Balanced Accuracy : 0.7377   
##   
## 'Positive' Class : Yes   
##

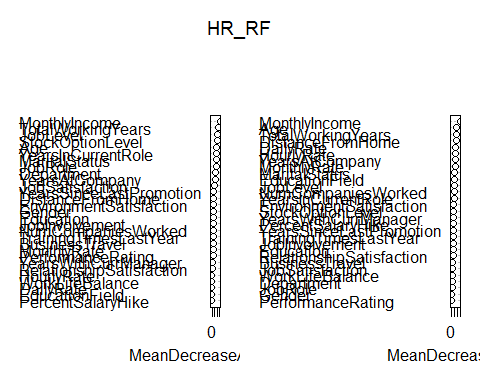
HR\_RF <- randomForest(Attrition~., data=HR\_train, mtry=3, ntree=100, na.action=na.omit, importance=TRUE)  
print(HR\_RF)

##   
## Call:  
## randomForest(formula = Attrition ~ ., data = HR\_train, mtry = 3, ntree = 100, importance = TRUE, na.action = na.omit)   
## Type of random forest: classification  
## Number of trees: 100  
## No. of variables tried at each split: 3  
##   
## OOB estimate of error rate: 24.02%  
## Confusion matrix:  
## No Yes class.error  
## No 205 16 0.07239819  
## Yes 64 48 0.57142857

importance(HR\_RF)

## No Yes MeanDecreaseAccuracy  
## Age 2.01111896 3.15054560 3.4763872  
## BusinessTravel 1.43521263 -0.72780493 0.6832796  
## DailyRate -1.23375671 0.68901192 -0.7331019  
## Department 1.00863958 3.08953774 2.8690568  
## DistanceFromHome 1.37584553 0.18860265 1.2199742  
## Education -0.60975043 1.89296152 0.8454003  
## EducationField -0.25170456 -0.92275716 -0.7558371  
## EnvironmentSatisfaction 1.51917338 -0.07954262 1.0893067  
## Gender 0.04713579 1.44162228 0.9033910  
## HourlyRate -1.13593320 0.66211775 -0.4412326  
## JobInvolvement 1.08438913 0.17865270 0.8426454  
## JobLevel 2.51121613 3.92177415 4.1994786  
## JobRole 2.02968920 2.47085011 3.0414266  
## JobSatisfaction 0.85282902 1.56658108 1.6020069  
## MaritalStatus 2.39698526 2.75385576 3.0724270  
## MonthlyIncome 3.61960900 6.37946920 6.5095091  
## MonthlyRate 0.45382747 0.46330109 0.5667634  
## NumCompaniesWorked 0.68157741 0.15707402 0.7959854  
## PercentSalaryHike -0.28792663 -1.19081293 -0.9448459  
## PerformanceRating 0.93333911 -0.20211337 0.5308399  
## RelationshipSatisfaction -0.44617117 0.12394381 -0.2405265  
## StockOptionLevel 3.00693688 3.36157346 4.1487094  
## TotalWorkingYears 3.93922042 2.63082917 5.0856782  
## TrainingTimesLastYear 1.40431333 -0.88437579 0.7613842  
## WorkLifeBalance -0.05367011 -0.78446258 -0.5142493  
## YearsAtCompany 2.11898800 2.27715607 2.8393936  
## YearsInCurrentRole 2.60845380 0.90675447 3.2821736  
## YearsSinceLastPromotion 0.68732443 1.13856801 1.2618219  
## YearsWithCurrManager 0.68851823 -0.23795926 0.3932294  
## MeanDecreaseGini  
## Age 9.5125410  
## BusinessTravel 2.9461062  
## DailyRate 7.1834070  
## Department 2.7334975  
## DistanceFromHome 8.1424304  
## Education 3.2528354  
## EducationField 5.7615304  
## EnvironmentSatisfaction 4.7514562  
## Gender 1.1717572  
## HourlyRate 6.8319005  
## JobInvolvement 3.2898498  
## JobLevel 5.4582290  
## JobRole 1.9535797  
## JobSatisfaction 2.9348510  
## MaritalStatus 6.2682963  
## MonthlyIncome 15.5934033  
## MonthlyRate 6.6349734  
## NumCompaniesWorked 5.1669846  
## PercentSalaryHike 4.3718291  
## PerformanceRating 0.7890649  
## RelationshipSatisfaction 3.2374459  
## StockOptionLevel 4.5467014  
## TotalWorkingYears 8.9304848  
## TrainingTimesLastYear 3.7771877  
## WorkLifeBalance 2.7650132  
## YearsAtCompany 6.7748657  
## YearsInCurrentRole 4.7693373  
## YearsSinceLastPromotion 4.0229217  
## YearsWithCurrManager 4.4804098

varImpPlot(HR\_RF)



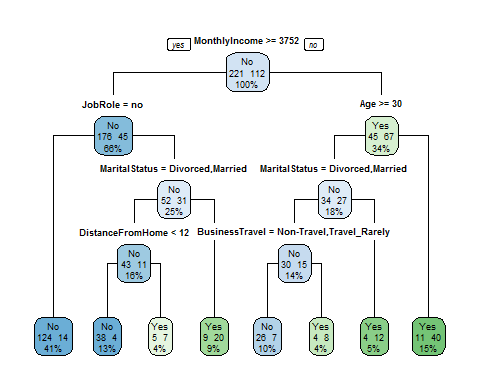
actualRF2 <- HR\_test$Attrition   
predictedRF2 <- predict(HR\_RF, HR\_test, type="class")   
HR\_RF\_matrix <- confusionMatrix(predictedRF2, actualRF2, positive="Yes")   
print("RF results")

## [1] "RF results"

print(HR\_RF\_matrix)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 56 7  
## Yes 12 8  
##   
## Accuracy : 0.7711   
## 95% CI : (0.6658, 0.8562)  
## No Information Rate : 0.8193   
## P-Value [Acc > NIR] : 0.8976   
##   
## Kappa : 0.3158   
## Mcnemar's Test P-Value : 0.3588   
##   
## Sensitivity : 0.53333   
## Specificity : 0.82353   
## Pos Pred Value : 0.40000   
## Neg Pred Value : 0.88889   
## Prevalence : 0.18072   
## Detection Rate : 0.09639   
## Detection Prevalence : 0.24096   
## Balanced Accuracy : 0.67843   
##   
## 'Positive' Class : Yes   
##

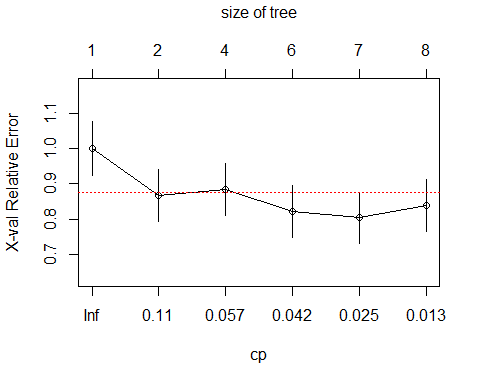
##################################################################  
#Decision Tree  
  
set.seed(123)  
HR\_DT <- rpart(HR\_train$Attrition~., method="class", parms = list(split="gini"), data=HR\_train)  
rpart.plot(HR\_DT, type=1, extra=101)



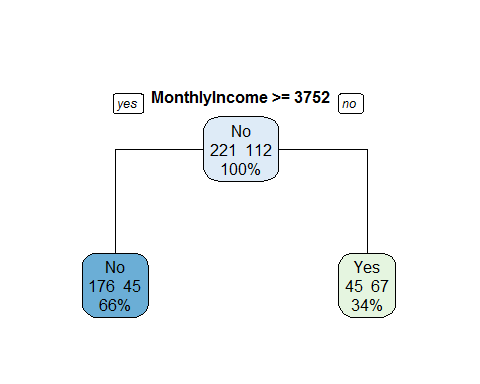
set.seed(123)  
cptable<-printcp(HR\_DT)

##   
## Classification tree:  
## rpart(formula = HR\_train$Attrition ~ ., data = HR\_train, method = "class",   
## parms = list(split = "gini"))  
##   
## Variables actually used in tree construction:  
## [1] Age BusinessTravel DistanceFromHome JobRole   
## [5] MaritalStatus MonthlyIncome   
##   
## Root node error: 112/333 = 0.33634  
##   
## n= 333   
##   
## CP nsplit rel error xerror xstd  
## 1 0.196429 0 1.00000 1.00000 0.076978  
## 2 0.066964 1 0.80357 0.86607 0.074029  
## 3 0.049107 3 0.66964 0.88393 0.074471  
## 4 0.035714 5 0.57143 0.82143 0.072855  
## 5 0.017857 6 0.53571 0.80357 0.072358  
## 6 0.010000 7 0.51786 0.83929 0.073337

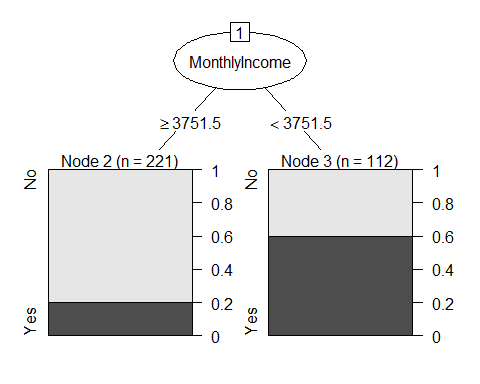
plotcp(HR\_DT, minline=TRUE, col="red")



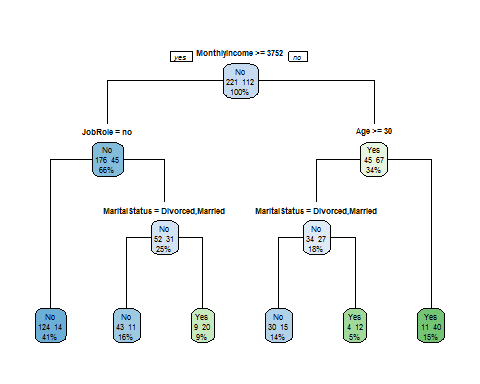
set.seed(123)  
Pruned\_HR\_DT <-prune(HR\_DT,cp=.1, minsplit=10, minbucket=round(minsplit/3))   
rpart.plot(Pruned\_HR\_DT, type=1, extra=101)



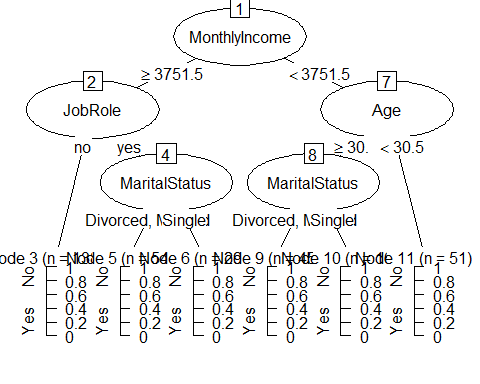
Pruned\_HR\_DT\_party<-as.party(Pruned\_HR\_DT)  
plot(Pruned\_HR\_DT\_party)



set.seed(123)  
Pruned\_HR\_DT2 <-prune(HR\_DT,cp=.04, minsplit=10, minbucket=round(minsplit/3))   
rpart.plot(Pruned\_HR\_DT2, type=1, extra=101)



Pruned\_HR\_DT2\_party2<-as.party(Pruned\_HR\_DT2)  
plot(Pruned\_HR\_DT2\_party2)



summary(Pruned\_HR\_DT2)

## Call:  
## rpart(formula = HR\_train$Attrition ~ ., data = HR\_train, method = "class",   
## parms = list(split = "gini"))  
## n= 333   
##   
## CP nsplit rel error xerror xstd  
## 1 0.19642857 0 1.0000000 1.0000000 0.07697771  
## 2 0.06696429 1 0.8035714 0.8660714 0.07402897  
## 3 0.04910714 3 0.6696429 0.8839286 0.07447069  
## 4 0.04000000 5 0.5714286 0.8214286 0.07285549  
##   
## Variable importance  
## MonthlyIncome JobLevel MaritalStatus   
## 21 16 12   
## Age TotalWorkingYears StockOptionLevel   
## 10 10 8   
## JobRole Department EducationField   
## 7 6 2   
## YearsWithCurrManager YearsAtCompany DistanceFromHome   
## 2 2 1   
## MonthlyRate HourlyRate TrainingTimesLastYear   
## 1 1 1   
##   
## Node number 1: 333 observations, complexity param=0.1964286  
## predicted class=No expected loss=0.3363363 P(node) =1  
## class counts: 221 112  
## probabilities: 0.664 0.336   
## left son=2 (221 obs) right son=3 (112 obs)  
## Primary splits:  
## MonthlyIncome < 3751.5 to the right, improve=23.14717, (0 missing)  
## JobLevel < 1.5 to the right, improve=20.57330, (0 missing)  
## TotalWorkingYears < 8.5 to the right, improve=14.85924, (0 missing)  
## StockOptionLevel < 0.5 to the right, improve=13.18539, (0 missing)  
## MaritalStatus splits as LLR, improve=13.08146, (0 missing)  
## Surrogate splits:  
## JobLevel < 1.5 to the right, agree=0.946, adj=0.839, (0 split)  
## TotalWorkingYears < 8.5 to the right, agree=0.793, adj=0.384, (0 split)  
## Age < 28.5 to the right, agree=0.730, adj=0.196, (0 split)  
## YearsAtCompany < 1.5 to the right, agree=0.697, adj=0.098, (0 split)  
## YearsWithCurrManager < 0.5 to the right, agree=0.685, adj=0.062, (0 split)  
##   
## Node number 2: 221 observations, complexity param=0.04910714  
## predicted class=No expected loss=0.2036199 P(node) =0.6636637  
## class counts: 176 45  
## probabilities: 0.796 0.204   
## left son=4 (138 obs) right son=5 (83 obs)  
## Primary splits:  
## JobRole splits as LR, improve=7.671414, (0 missing)  
## Department splits as LLR, improve=6.462486, (0 missing)  
## MaritalStatus splits as LLR, improve=5.774892, (0 missing)  
## StockOptionLevel < 0.5 to the right, improve=5.604272, (0 missing)  
## DistanceFromHome < 11.5 to the left, improve=4.246539, (0 missing)  
## Surrogate splits:  
## Department splits as RLR, agree=0.946, adj=0.855, (0 split)  
## EducationField splits as RLRLLL, agree=0.751, adj=0.337, (0 split)  
## TotalWorkingYears < 8.5 to the right, agree=0.656, adj=0.084, (0 split)  
## Age < 28.5 to the right, agree=0.652, adj=0.072, (0 split)  
## DailyRate < 1469.5 to the left, agree=0.652, adj=0.072, (0 split)  
##   
## Node number 3: 112 observations, complexity param=0.06696429  
## predicted class=Yes expected loss=0.4017857 P(node) =0.3363363  
## class counts: 45 67  
## probabilities: 0.402 0.598   
## left son=6 (61 obs) right son=7 (51 obs)  
## Primary splits:  
## Age < 30.5 to the right, improve=6.486023, (0 missing)  
## StockOptionLevel < 0.5 to the right, improve=6.188502, (0 missing)  
## MaritalStatus splits as LLR, improve=5.913750, (0 missing)  
## MonthlyIncome < 2475 to the right, improve=5.408890, (0 missing)  
## NumCompaniesWorked < 0.5 to the left, improve=4.404298, (0 missing)  
## Surrogate splits:  
## TotalWorkingYears < 2.5 to the right, agree=0.670, adj=0.275, (0 split)  
## MonthlyIncome < 2572.5 to the right, agree=0.661, adj=0.255, (0 split)  
## YearsWithCurrManager < 1.5 to the right, agree=0.616, adj=0.157, (0 split)  
## MaritalStatus splits as LLR, agree=0.607, adj=0.137, (0 split)  
## StockOptionLevel < 0.5 to the right, agree=0.607, adj=0.137, (0 split)  
##   
## Node number 4: 138 observations  
## predicted class=No expected loss=0.1014493 P(node) =0.4144144  
## class counts: 124 14  
## probabilities: 0.899 0.101   
##   
## Node number 5: 83 observations, complexity param=0.04910714  
## predicted class=No expected loss=0.373494 P(node) =0.2492492  
## class counts: 52 31  
## probabilities: 0.627 0.373   
## left son=10 (54 obs) right son=11 (29 obs)  
## Primary splits:  
## MaritalStatus splits as LLR, improve=8.911062, (0 missing)  
## DistanceFromHome < 11 to the left, improve=8.726971, (0 missing)  
## StockOptionLevel < 0.5 to the right, improve=6.880670, (0 missing)  
## YearsAtCompany < 2.5 to the right, improve=4.827989, (0 missing)  
## YearsSinceLastPromotion < 1.5 to the left, improve=4.820732, (0 missing)  
## Surrogate splits:  
## StockOptionLevel < 0.5 to the right, agree=0.880, adj=0.655, (0 split)  
## HourlyRate < 79 to the left, agree=0.699, adj=0.138, (0 split)  
## DistanceFromHome < 18.5 to the left, agree=0.687, adj=0.103, (0 split)  
## MonthlyRate < 10419.5 to the right, agree=0.675, adj=0.069, (0 split)  
## TrainingTimesLastYear < 0.5 to the right, agree=0.675, adj=0.069, (0 split)  
##   
## Node number 6: 61 observations, complexity param=0.06696429  
## predicted class=No expected loss=0.442623 P(node) =0.1831832  
## class counts: 34 27  
## probabilities: 0.557 0.443   
## left son=12 (45 obs) right son=13 (16 obs)  
## Primary splits:  
## MaritalStatus splits as LLR, improve=4.098361, (0 missing)  
## BusinessTravel splits as RRL, improve=3.942263, (0 missing)  
## StockOptionLevel < 0.5 to the right, improve=3.242525, (0 missing)  
## DailyRate < 1374 to the right, improve=3.098361, (0 missing)  
## EnvironmentSatisfaction < 1.5 to the right, improve=3.055223, (0 missing)  
## Surrogate splits:  
## StockOptionLevel < 0.5 to the right, agree=0.885, adj=0.562, (0 split)  
## Age < 33.5 to the right, agree=0.787, adj=0.187, (0 split)  
## MonthlyRate < 3574.5 to the right, agree=0.787, adj=0.187, (0 split)  
## DistanceFromHome < 1.5 to the right, agree=0.770, adj=0.125, (0 split)  
## YearsInCurrentRole < 7.5 to the left, agree=0.754, adj=0.062, (0 split)  
##   
## Node number 7: 51 observations  
## predicted class=Yes expected loss=0.2156863 P(node) =0.1531532  
## class counts: 11 40  
## probabilities: 0.216 0.784   
##   
## Node number 10: 54 observations  
## predicted class=No expected loss=0.2037037 P(node) =0.1621622  
## class counts: 43 11  
## probabilities: 0.796 0.204   
##   
## Node number 11: 29 observations  
## predicted class=Yes expected loss=0.3103448 P(node) =0.08708709  
## class counts: 9 20  
## probabilities: 0.310 0.690   
##   
## Node number 12: 45 observations  
## predicted class=No expected loss=0.3333333 P(node) =0.1351351  
## class counts: 30 15  
## probabilities: 0.667 0.333   
##   
## Node number 13: 16 observations  
## predicted class=Yes expected loss=0.25 P(node) =0.04804805  
## class counts: 4 12  
## probabilities: 0.250 0.750

actualFullDT <- HR\_test$Attrition  
predictedFullDT <- predict(HR\_DT, HR\_test, type="class")  
results.matrix <- confusionMatrix(predictedFullDT, actualFullDT, positive="Yes")  
print(results.matrix)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 49 6  
## Yes 19 9  
##   
## Accuracy : 0.6988   
## 95% CI : (0.5882, 0.7947)  
## No Information Rate : 0.8193   
## P-Value [Acc > NIR] : 0.9976   
##   
## Kappa : 0.2396   
## Mcnemar's Test P-Value : 0.0164   
##   
## Sensitivity : 0.6000   
## Specificity : 0.7206   
## Pos Pred Value : 0.3214   
## Neg Pred Value : 0.8909   
## Prevalence : 0.1807   
## Detection Rate : 0.1084   
## Detection Prevalence : 0.3373   
## Balanced Accuracy : 0.6603   
##   
## 'Positive' Class : Yes   
##

actualDT <- HR\_test$Attrition  
predictedDT <- predict(Pruned\_HR\_DT, HR\_test, type="class")  
DT.matrix <- confusionMatrix(predictedDT, actualDT, positive="Yes")  
print(DT.matrix)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 48 4  
## Yes 20 11  
##   
## Accuracy : 0.7108   
## 95% CI : (0.6009, 0.8052)  
## No Information Rate : 0.8193   
## P-Value [Acc > NIR] : 0.9948   
##   
## Kappa : 0.3102   
## Mcnemar's Test P-Value : 0.0022   
##   
## Sensitivity : 0.7333   
## Specificity : 0.7059   
## Pos Pred Value : 0.3548   
## Neg Pred Value : 0.9231   
## Prevalence : 0.1807   
## Detection Rate : 0.1325   
## Detection Prevalence : 0.3735   
## Balanced Accuracy : 0.7196   
##   
## 'Positive' Class : Yes   
##

actualDT2 <- HR\_test$Attrition  
predictedDT2 <- predict(Pruned\_HR\_DT2, HR\_test, type="class")  
DT.matrix2 <- confusionMatrix(predictedDT2, actualDT2, positive="Yes")  
print(DT.matrix2)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 53 7  
## Yes 15 8  
##   
## Accuracy : 0.7349   
## 95% CI : (0.6266, 0.8258)  
## No Information Rate : 0.8193   
## P-Value [Acc > NIR] : 0.9800   
##   
## Kappa : 0.2589   
## Mcnemar's Test P-Value : 0.1356   
##   
## Sensitivity : 0.53333   
## Specificity : 0.77941   
## Pos Pred Value : 0.34783   
## Neg Pred Value : 0.88333   
## Prevalence : 0.18072   
## Detection Rate : 0.09639   
## Detection Prevalence : 0.27711   
## Balanced Accuracy : 0.65637   
##   
## 'Positive' Class : Yes   
##

#########################################################################################  
#Logistic Regression  
  
set.seed(123)  
HR\_train\_logit<-HR\_train  
HR\_train\_logit$MonthlyIncome<-ifelse(HR\_train\_logit$MonthlyIncome>=3752,"no","yes")  
HR\_train\_logit$Age<-ifelse(HR\_train\_logit$Age>=30.5,"no","yes")  
HR\_train\_logit$MaritalStatus<-ifelse(HR\_train\_logit$MaritalStatus=="Single","yes","no")  
  
#HR.logit <- glm(HR\_train\_logit$Attrition~., data=HR\_train\_logit, family=binomial())  
#summary(HR.logit)   
#names(HR\_train\_logit)  
  
#HR\_train\_logit2<-HR\_train\_logit[,c(1,2:3, 6, 9, 12, 16:17, 19, 22, 25)]  
#HR.logit2 <- glm(HR\_train\_logit2$Attrition~., data=HR\_train\_logit2, family=binomial())  
#summary(HR.logit2)   
  
HR\_train\_logit3<-HR\_train\_logit[,c(1,2:3, 6, 9, 12, 16:17, 19, 22,25)]  
HR.logit3 <- glm(HR\_train\_logit3$Attrition~., data=HR\_train\_logit3, family=binomial())  
summary(HR.logit3)

##   
## Call:  
## glm(formula = HR\_train\_logit3$Attrition ~ ., family = binomial(),   
## data = HR\_train\_logit3)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1891 -0.6165 -0.2985 0.5447 2.8566   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -0.15473 1.02861 -0.150  
## Ageyes 1.00813 0.36966 2.727  
## BusinessTravelTravel\_Frequently 1.76337 0.69030 2.555  
## BusinessTravelTravel\_Rarely 1.11202 0.64472 1.725  
## DistanceFromHome 0.10086 0.02106 4.789  
## EnvironmentSatisfaction -0.44399 0.14347 -3.095  
## JobInvolvement -0.72527 0.21751 -3.334  
## MaritalStatusyes 1.83080 0.33810 5.415  
## MonthlyIncomeyes 2.39361 0.35789 6.688  
## NumCompaniesWorked 0.19422 0.06135 3.166  
## RelationshipSatisfaction -0.40843 0.14816 -2.757  
## TrainingTimesLastYear -0.30378 0.13482 -2.253  
## Pr(>|z|)   
## (Intercept) 0.880430   
## Ageyes 0.006388 \*\*   
## BusinessTravelTravel\_Frequently 0.010634 \*   
## BusinessTravelTravel\_Rarely 0.084562 .   
## DistanceFromHome 0.0000016786183 \*\*\*  
## EnvironmentSatisfaction 0.001970 \*\*   
## JobInvolvement 0.000855 \*\*\*  
## MaritalStatusyes 0.0000000613033 \*\*\*  
## MonthlyIncomeyes 0.0000000000226 \*\*\*  
## NumCompaniesWorked 0.001546 \*\*   
## RelationshipSatisfaction 0.005839 \*\*   
## TrainingTimesLastYear 0.024246 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 425.29 on 332 degrees of freedom  
## Residual deviance: 274.29 on 321 degrees of freedom  
## AIC: 298.29  
##   
## Number of Fisher Scoring iterations: 5

odds<-exp(cbind(Odds\_Ratio=coef(HR.logit3)))  
odds

## Odds\_Ratio  
## (Intercept) 0.8566475  
## Ageyes 2.7404654  
## BusinessTravelTravel\_Frequently 5.8320578  
## BusinessTravelTravel\_Rarely 3.0404951  
## DistanceFromHome 1.1061234  
## EnvironmentSatisfaction 0.6414696  
## JobInvolvement 0.4841953  
## MaritalStatusyes 6.2388711  
## MonthlyIncomeyes 10.9529564  
## NumCompaniesWorked 1.2143579  
## RelationshipSatisfaction 0.6646941  
## TrainingTimesLastYear 0.7380212

prob<-odds/(1+odds)  
prob

## Odds\_Ratio  
## (Intercept) 0.4613948  
## Ageyes 0.7326536  
## BusinessTravelTravel\_Frequently 0.8536312  
## BusinessTravelTravel\_Rarely 0.7525056  
## DistanceFromHome 0.5251940  
## EnvironmentSatisfaction 0.3907898  
## JobInvolvement 0.3262342  
## MaritalStatusyes 0.8618569  
## MonthlyIncomeyes 0.9163387  
## NumCompaniesWorked 0.5484018  
## RelationshipSatisfaction 0.3992890  
## TrainingTimesLastYear 0.4246330

anova(HR.logit3,test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: HR\_train\_logit3$Attrition  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev  
## NULL 332 425.29  
## Age 1 22.607 331 402.68  
## BusinessTravel 2 5.395 329 397.29  
## DistanceFromHome 1 9.503 328 387.79  
## EnvironmentSatisfaction 1 9.543 327 378.24  
## JobInvolvement 1 7.720 326 370.52  
## MaritalStatus 1 26.023 325 344.50  
## MonthlyIncome 1 47.425 324 297.08  
## NumCompaniesWorked 1 8.238 323 288.84  
## RelationshipSatisfaction 1 9.244 322 279.59  
## TrainingTimesLastYear 1 5.303 321 274.29  
## Pr(>Chi)   
## NULL   
## Age 0.000001987636637 \*\*\*  
## BusinessTravel 0.067368 .   
## DistanceFromHome 0.002051 \*\*   
## EnvironmentSatisfaction 0.002008 \*\*   
## JobInvolvement 0.005462 \*\*   
## MaritalStatus 0.000000337456969 \*\*\*  
## MonthlyIncome 0.000000000005716 \*\*\*  
## NumCompaniesWorked 0.004102 \*\*   
## RelationshipSatisfaction 0.002363 \*\*   
## TrainingTimesLastYear 0.021289 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

HR\_test\_logit3<-HR\_test[,c(1,2:3, 6, 9, 12, 16:17, 19, 22,25)]  
HR\_test\_logit3$MonthlyIncome<-ifelse(HR\_test\_logit3$MonthlyIncome>=3752,"no","yes")  
HR\_test\_logit3$Age<-ifelse(HR\_test\_logit3$Age>=30.5,"no","yes")  
HR\_test\_logit3$MaritalStatus<-ifelse(HR\_test\_logit3$MaritalStatus=="Single","yes","no")  
HR\_test\_logit3$predict.Attrition<-predict(HR.logit3, newdata=HR\_test\_logit3,type = "response")  
  
  
HR\_test\_logit\_CI<-cbind(HR\_test\_logit3,predict(HR.logit3, newdata=HR\_test\_logit3,type="link",se=TRUE))  
  
HR\_test\_logit\_CI <- within(HR\_test\_logit\_CI,   
 {  
 PredictedProb <- plogis(fit)  
 LL <- plogis(fit - (1.96 \* se.fit))  
 UL <- plogis(fit + (1.96 \* se.fit))  
 })   
  
names(HR\_test\_logit\_CI)

## [1] "Attrition" "Age"   
## [3] "BusinessTravel" "DistanceFromHome"   
## [5] "EnvironmentSatisfaction" "JobInvolvement"   
## [7] "MaritalStatus" "MonthlyIncome"   
## [9] "NumCompaniesWorked" "RelationshipSatisfaction"  
## [11] "TrainingTimesLastYear" "predict.Attrition"   
## [13] "fit" "se.fit"   
## [15] "residual.scale" "UL"   
## [17] "LL" "PredictedProb"

dim(HR\_test\_logit\_CI)

## [1] 83 18

HR\_test\_logit\_CI$predict.Attrition<-ifelse(HR\_test\_logit\_CI$predict.Attrition>.5, "yes", "no")  
probs<-HR\_test\_logit\_CI[,c(1,12)]   
  
CrossTable(x=probs$Attrition, y=probs$predict.Attrition, prob.chisq=FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 83   
##   
##   
## | probs$predict.Attrition   
## probs$Attrition | no | yes | Row Total |   
## ----------------|-----------|-----------|-----------|  
## No | 55 | 13 | 68 |   
## | 1.178 | 2.733 | |   
## | 0.809 | 0.191 | 0.819 |   
## | 0.948 | 0.520 | |   
## | 0.663 | 0.157 | |   
## ----------------|-----------|-----------|-----------|  
## Yes | 3 | 12 | 15 |   
## | 5.341 | 12.390 | |   
## | 0.200 | 0.800 | 0.181 |   
## | 0.052 | 0.480 | |   
## | 0.036 | 0.145 | |   
## ----------------|-----------|-----------|-----------|  
## Column Total | 58 | 25 | 83 |   
## | 0.699 | 0.301 | |   
## ----------------|-----------|-----------|-----------|  
##   
##

TP = 12  
TN = 55  
FP = 13  
FN = 3  
Sensitivity = TP/(TP+FN) #true positive rate; recall; TP/(TP+FN)  
Specificity = TN/(TN+FP) #how often is the prediction negative when actual is negative?  
  
Precision = TP/(TP+FP) #how often is prediction positive when actual is positive?  
Accuracy = (TP+TN)/(TP+TN+FP+FN) #how often is classifier correct  
Value<-round(c(TP,TN,FP,FN,Sensitivity,Specificity,Precision,Accuracy),digits=3)  
Measure<-c("True Positive","True Negative","False Positive","False Negative","Sensitivity=TP/(TN+FP)",  
 "Specificity=TN/(TN+TP)","Precision=TP/(TP+FP)","Accuracy=(TP+TN)/total")  
table<-as.data.frame(cbind(Measure,Value))  
kable(table)

|  |  |
| --- | --- |
| Measure | Value |
| True Positive | 12 |
| True Negative | 55 |
| False Positive | 13 |
| False Negative | 3 |
| Sensitivity=TP/(TN+FP) | 0.8 |
| Specificity=TN/(TN+TP) | 0.809 |
| Precision=TP/(TP+FP) | 0.48 |
| Accuracy=(TP+TN)/total | 0.807 |