Final\_project\_no\_OT.R

Jennifer

Mon May 08 13:18:40 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

options(scipen = 999)  
   
setwd("C:/Users/Jennifer/Documents/ADM/Final")  
HR<-read.csv("WatsonHRnoOT.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 :944 Min. :18.00 Non-Travel :115 Min. : 102.0   
## Yes:110 1st Qu.:30.00 Travel\_Frequently:191 1st Qu.: 467.2   
## Median :35.00 Travel\_Rarely :748 Median : 803.0   
## Mean :36.76 Mean : 800.2   
## 3rd Qu.:42.00 3rd Qu.:1152.5   
## Max. :60.00 Max. :1499.0   
##   
## Department DistanceFromHome Education   
## Human Resources : 46 Min. : 1.000 Min. :1.000   
## Research & Development:690 1st Qu.: 2.000 1st Qu.:2.000   
## Sales :318 Median : 7.000 Median :3.000   
## Mean : 9.063 Mean :2.926   
## 3rd Qu.:13.000 3rd Qu.:4.000   
## Max. :29.000 Max. :5.000   
##   
## EducationField EnvironmentSatisfaction Gender   
## Human Resources : 19 Min. :1.000 Female:408   
## Life Sciences :439 1st Qu.:2.000 Male :646   
## Marketing :111 Median :3.000   
## Medical :332 Mean :2.674   
## Other : 55 3rd Qu.:4.000   
## Technical Degree: 98 Max. :4.000   
##   
## HourlyRate JobInvolvement JobLevel   
## Min. : 30.00 Min. :1.000 Min. :1.000   
## 1st Qu.: 48.00 1st Qu.:2.000 1st Qu.:1.000   
## Median : 66.00 Median :3.000 Median :2.000   
## Mean : 65.99 Mean :2.731 Mean :2.064   
## 3rd Qu.: 84.00 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :100.00 Max. :4.000 Max. :5.000   
##   
## JobRole JobSatisfaction MaritalStatus  
## Sales Executive :232 Min. :1.000 Divorced:228   
## Laboratory Technician :197 1st Qu.:2.000 Married :487   
## Research Scientist :195 Median :3.000 Single :339   
## Manufacturing Director :106 Mean :2.712   
## Healthcare Representative: 94 3rd Qu.:4.000   
## Manager : 75 Max. :4.000   
## (Other) :155   
## MonthlyIncome MonthlyRate NumCompaniesWorked PercentSalaryHike  
## Min. : 1051 Min. : 2094 Min. :0.000 Min. :11.00   
## 1st Qu.: 2931 1st Qu.: 7956 1st Qu.:1.000 1st Qu.:12.00   
## Median : 4880 Median :13963 Median :2.000 Median :14.00   
## Mean : 6485 Mean :14217 Mean :2.726 Mean :15.22   
## 3rd Qu.: 8368 3rd Qu.:20364 3rd Qu.:4.000 3rd Qu.:18.00   
## Max. :19999 Max. :26997 Max. :9.000 Max. :25.00   
##   
## PerformanceRating RelationshipSatisfaction StockOptionLevel  
## Min. :3.000 Min. :1.000 Min. :0.0000   
## 1st Qu.:3.000 1st Qu.:2.000 1st Qu.:0.0000   
## Median :3.000 Median :3.000 Median :1.0000   
## Mean :3.153 Mean :2.679 Mean :0.7941   
## 3rd Qu.:3.000 3rd Qu.:4.000 3rd Qu.:1.0000   
## Max. :4.000 Max. :4.000 Max. :3.0000   
##   
## TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany   
## Min. : 0.00 Min. :0.000 Min. :1.000 Min. : 0.000   
## 1st Qu.: 6.00 1st Qu.:2.000 1st Qu.:2.000 1st Qu.: 3.000   
## Median :10.00 Median :3.000 Median :3.000 Median : 5.000   
## Mean :11.22 Mean :2.863 Mean :2.773 Mean : 7.053   
## 3rd Qu.:15.00 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:10.000   
## Max. :40.00 Max. :6.000 Max. :4.000 Max. :40.000   
##   
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager  
## Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 2.000 1st Qu.: 0.000 1st Qu.: 2.000   
## Median : 3.000 Median : 1.000 Median : 3.000   
## Mean : 4.297 Mean : 2.213 Mean : 4.216   
## 3rd Qu.: 7.000 3rd Qu.: 3.000 3rd Qu.: 7.000   
## Max. :18.000 Max. :15.000 Max. :17.000   
##

dim(HR)

## [1] 1054 30

str(HR)

## 'data.frame': 1054 obs. of 30 variables:  
## $ Attrition : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Age : int 49 27 32 30 38 36 35 31 34 29 ...  
## $ BusinessTravel : Factor w/ 3 levels "Non-Travel","Travel\_Frequently",..: 2 3 2 3 2 3 3 3 3 3 ...  
## $ DailyRate : int 279 591 1005 1358 216 1299 809 670 1346 1389 ...  
## $ Department : Factor w/ 3 levels "Human Resources",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ DistanceFromHome : int 8 2 2 24 23 27 16 26 19 21 ...  
## $ Education : int 1 1 2 1 3 3 3 1 2 4 ...  
## $ EducationField : Factor w/ 6 levels "Human Resources",..: 2 4 2 2 2 4 4 2 4 2 ...  
## $ EnvironmentSatisfaction : int 3 1 4 4 4 3 1 1 2 2 ...  
## $ Gender : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...  
## $ HourlyRate : int 61 40 79 67 44 94 84 31 93 51 ...  
## $ JobInvolvement : int 2 3 3 3 2 3 4 3 3 4 ...  
## $ JobLevel : int 2 1 1 1 3 2 1 1 1 3 ...  
## $ JobRole : Factor w/ 9 levels "Healthcare Representative",..: 7 3 3 3 5 1 3 7 3 5 ...  
## $ JobSatisfaction : int 2 2 4 3 3 3 2 3 4 1 ...  
## $ MaritalStatus : Factor w/ 3 levels "Divorced","Married",..: 2 2 3 1 3 2 2 1 1 1 ...  
## $ MonthlyIncome : int 5130 3468 3068 2693 9526 5237 2426 2911 2661 9980 ...  
## $ MonthlyRate : int 24907 16632 11864 13335 8787 16577 16479 15170 8758 10195 ...  
## $ NumCompaniesWorked : int 1 9 0 1 0 6 0 1 0 1 ...  
## $ PercentSalaryHike : int 23 12 13 22 21 13 13 17 11 11 ...  
## $ PerformanceRating : int 4 3 3 4 4 3 3 3 3 3 ...  
## $ RelationshipSatisfaction: int 4 4 3 2 2 2 3 4 3 3 ...  
## $ StockOptionLevel : int 1 1 0 1 0 2 1 1 1 1 ...  
## $ TotalWorkingYears : int 10 6 8 1 10 17 6 5 3 10 ...  
## $ TrainingTimesLastYear : int 3 3 2 2 2 3 5 1 2 1 ...  
## $ WorkLifeBalance : int 3 3 2 3 3 2 3 2 3 3 ...  
## $ YearsAtCompany : int 10 2 7 1 9 7 5 5 2 10 ...  
## $ YearsInCurrentRole : int 7 2 7 0 7 7 4 2 2 9 ...  
## $ YearsSinceLastPromotion : int 1 2 3 0 1 7 0 4 1 8 ...  
## $ YearsWithCurrManager : int 7 2 6 0 8 7 3 3 2 8 ...

set.seed(123)  
HR\_rand <- HR[order(runif(1054)), ]  
1054\*.8

## [1] 843.2

HR\_train <- HR\_rand[1:843,]  
HR\_test <- HR\_rand[844:1054,]  
  
prop.table(table(HR\_train$Attrition))

##   
## No Yes   
## 0.8979834 0.1020166

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

##   
## No Yes   
## 0.8862559 0.1137441

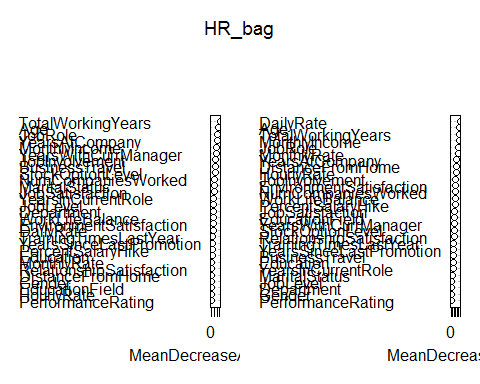
##################################################################  
#Random Forest  
  
set.seed(123)   
HR\_bag <- randomForest(Attrition~., data=HR\_train, mtry=29, na.action=na.omit, importance=TRUE)  
print(HR\_bag)

##   
## 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: 9.96%  
## Confusion matrix:  
## No Yes class.error  
## No 753 4 0.005284016  
## Yes 80 6 0.930232558

importance(HR\_bag)

## No Yes MeanDecreaseAccuracy  
## Age 10.7401582 7.03434443 12.9272521  
## BusinessTravel 6.1816370 5.67554927 7.5270728  
## DailyRate 2.6445321 -0.05880241 2.3643200  
## Department 3.4383769 0.64836520 3.6268898  
## DistanceFromHome -0.8033966 0.95527212 -0.3197930  
## Education 1.3626220 -1.47636231 0.7943075  
## EducationField -0.9864367 0.57135332 -0.7195664  
## EnvironmentSatisfaction -0.2273584 6.50435571 2.4350494  
## Gender -0.7262258 -0.14139156 -0.6714181  
## HourlyRate -0.4873323 -1.53421959 -1.0141956  
## JobInvolvement 4.7964258 8.13259191 7.6947963  
## JobLevel 3.4937162 -0.36357100 3.7679789  
## JobRole 10.2042593 3.42106031 10.9827463  
## JobSatisfaction 3.3008534 4.03143875 4.6149674  
## MaritalStatus 4.9985321 -3.61818977 4.7444547  
## MonthlyIncome 9.9047946 -1.83987670 9.4873036  
## MonthlyRate -0.5292658 2.18494380 0.3911473  
## NumCompaniesWorked 6.1859763 -1.03556840 5.6348379  
## PercentSalaryHike 2.2121191 -3.32103445 1.0808512  
## PerformanceRating -0.9703791 -0.79427247 -1.0771541  
## RelationshipSatisfaction 0.3459469 -0.22561992 0.2251069  
## StockOptionLevel 5.8619140 2.55854555 6.4937004  
## TotalWorkingYears 14.6561673 1.55783045 15.9723837  
## TrainingTimesLastYear 1.5099467 0.71426003 1.7006391  
## WorkLifeBalance 2.3141184 1.75921097 2.8467840  
## YearsAtCompany 9.5506723 3.46786379 10.5254929  
## YearsInCurrentRole 3.5112644 0.51464360 3.7832456  
## YearsSinceLastPromotion 2.4879628 -3.66603333 1.1754807  
## YearsWithCurrManager 8.4286363 0.05994444 8.5233730  
## MeanDecreaseGini  
## Age 11.3264565  
## BusinessTravel 3.4461418  
## DailyRate 12.1524384  
## Department 0.7395462  
## DistanceFromHome 7.0650293  
## Education 2.7280984  
## EducationField 4.6988998  
## EnvironmentSatisfaction 5.9419073  
## Gender 0.3437631  
## HourlyRate 6.7536944  
## JobInvolvement 6.2309319  
## JobLevel 1.0646380  
## JobRole 9.6103049  
## JobSatisfaction 4.9298970  
## MaritalStatus 1.2635753  
## MonthlyIncome 10.4049437  
## MonthlyRate 8.9761914  
## NumCompaniesWorked 5.8611063  
## PercentSalaryHike 5.1949083  
## PerformanceRating 0.3334062  
## RelationshipSatisfaction 3.7860736  
## StockOptionLevel 4.0625526  
## TotalWorkingYears 10.4438540  
## TrainingTimesLastYear 3.6829918  
## WorkLifeBalance 5.5051782  
## YearsAtCompany 7.3616705  
## YearsInCurrentRole 2.5082918  
## YearsSinceLastPromotion 3.4836188  
## YearsWithCurrManager 4.4925075

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 187 23  
## Yes 0 1  
##   
## Accuracy : 0.891   
## 95% CI : (0.8409, 0.9296)  
## No Information Rate : 0.8863   
## P-Value [Acc > NIR] : 0.4679   
##   
## Kappa : 0.0716   
## Mcnemar's Test P-Value : 0.00000449   
##   
## Sensitivity : 0.041667   
## Specificity : 1.000000   
## Pos Pred Value : 1.000000   
## Neg Pred Value : 0.890476   
## Prevalence : 0.113744   
## Detection Rate : 0.004739   
## Detection Prevalence : 0.004739   
## Balanced Accuracy : 0.520833   
##   
## 'Positive' Class : Yes   
##

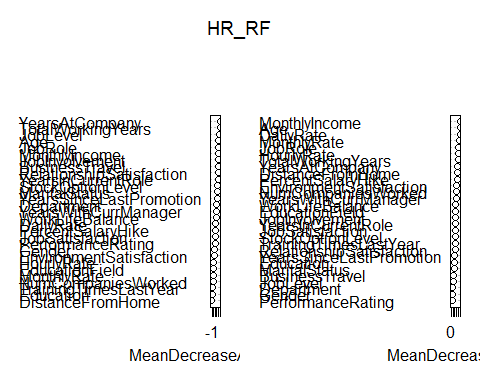
HR\_RF <- randomForest(Attrition~., data=HR\_train, mtry=3, ntree=100, na.action=na.omit, importance=TRUE)  
print(HR\_RF)

##   
## 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: 9.96%  
## Confusion matrix:  
## No Yes class.error  
## No 756 1 0.001321004  
## Yes 83 3 0.965116279

importance(HR\_RF)

## No Yes MeanDecreaseAccuracy  
## Age 2.81264532 2.31897048 3.27975176  
## BusinessTravel 2.10271017 2.71870115 2.79276887  
## DailyRate 1.16747023 1.02802248 1.31954915  
## Department 0.87298845 2.70955078 1.89639108  
## DistanceFromHome -1.72963013 -0.35202666 -1.72464179  
## Education -0.08153389 -2.16327389 -0.74273353  
## EducationField 0.10158812 -0.51692045 -0.15033565  
## EnvironmentSatisfaction -0.39906227 1.46826933 0.16144141  
## Gender 0.32699982 1.02536986 0.74420027  
## HourlyRate 0.60102833 -1.51135578 -0.06845163  
## JobInvolvement 2.51229827 2.59623577 3.18425850  
## JobLevel 3.01639234 1.86632032 3.47117240  
## JobRole 2.89669705 1.37503569 3.25648691  
## JobSatisfaction 0.66867887 0.69310378 0.95078208  
## MaritalStatus 1.72300989 2.81857491 2.48041508  
## MonthlyIncome 3.06806296 0.67462562 3.24517001  
## MonthlyRate 0.34249863 -1.57954005 -0.31221055  
## NumCompaniesWorked -0.25659028 -0.48072909 -0.39900180  
## PercentSalaryHike 1.29570913 -0.60975534 1.19217296  
## PerformanceRating 1.01640994 -0.59271661 0.80399774  
## RelationshipSatisfaction 3.07215657 -0.09199392 2.78053878  
## StockOptionLevel 1.93612607 2.78350890 2.58071879  
## TotalWorkingYears 3.85408009 0.36573281 4.01080393  
## TrainingTimesLastYear -0.84545136 0.16621311 -0.64406097  
## WorkLifeBalance 1.67154736 0.15770104 1.64298896  
## YearsAtCompany 3.38322798 3.26043507 4.26445605  
## YearsInCurrentRole 1.84971083 2.69477296 2.71693669  
## YearsSinceLastPromotion 1.64733072 2.13570145 2.16732706  
## YearsWithCurrManager 1.47384046 1.03897522 1.83188026  
## MeanDecreaseGini  
## Age 9.6751033  
## BusinessTravel 2.9525257  
## DailyRate 9.4955767  
## Department 1.8021906  
## DistanceFromHome 6.5411133  
## Education 3.4260463  
## EducationField 4.8428938  
## EnvironmentSatisfaction 5.7836784  
## Gender 1.4410504  
## HourlyRate 7.5123804  
## JobInvolvement 4.6589619  
## JobLevel 2.5805785  
## JobRole 7.6746641  
## JobSatisfaction 4.4322792  
## MaritalStatus 3.0610247  
## MonthlyIncome 9.9168510  
## MonthlyRate 8.9703360  
## NumCompaniesWorked 5.6442022  
## PercentSalaryHike 6.4378793  
## PerformanceRating 0.9939444  
## RelationshipSatisfaction 4.1026181  
## StockOptionLevel 4.1392044  
## TotalWorkingYears 7.4652435  
## TrainingTimesLastYear 4.1155546  
## WorkLifeBalance 4.8629461  
## YearsAtCompany 7.2911894  
## YearsInCurrentRole 4.6468638  
## YearsSinceLastPromotion 3.7764378  
## YearsWithCurrManager 5.5001827

varImpPlot(HR\_RF)



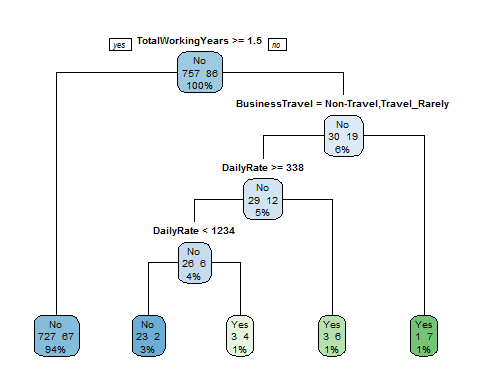
actualRF2 <- HR\_test$Attrition   
predictedRF2 <- predict(HR\_RF, HR\_test, type="class")   
HR\_RF\_matrix <- confusionMatrix(predictedRF2, actualRF2, positive="Yes")   
print("RF results")

## [1] "RF results"

print(HR\_RF\_matrix)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 187 24  
## Yes 0 0  
##   
## Accuracy : 0.8863   
## 95% CI : (0.8355, 0.9257)  
## No Information Rate : 0.8863   
## P-Value [Acc > NIR] : 0.5541   
##   
## Kappa : 0   
## Mcnemar's Test P-Value : 0.000002668   
##   
## Sensitivity : 0.0000   
## Specificity : 1.0000   
## Pos Pred Value : NaN   
## Neg Pred Value : 0.8863   
## Prevalence : 0.1137   
## Detection Rate : 0.0000   
## Detection Prevalence : 0.0000   
## Balanced Accuracy : 0.5000   
##   
## 'Positive' Class : Yes   
##

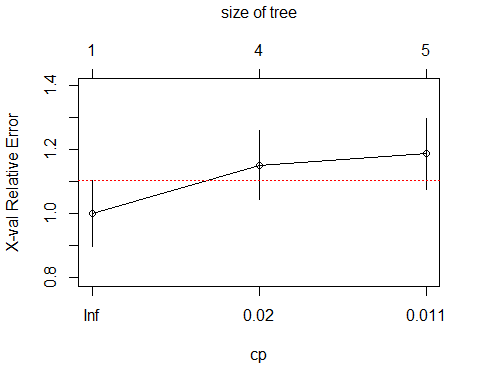
##################################################################  
#Decision Tree  
  
set.seed(123)  
HR\_DT <- rpart(HR\_train$Attrition~., method="class", parms = list(split="gini"), data=HR\_train)  
rpart.plot(HR\_DT, type=1, extra=101)



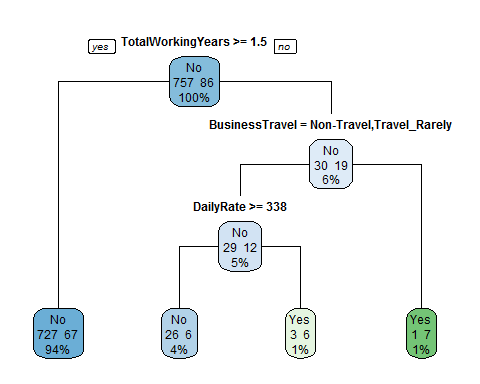
set.seed(123)  
cptable<-printcp(HR\_DT)

##   
## Classification tree:  
## rpart(formula = HR\_train$Attrition ~ ., data = HR\_train, method = "class",   
## parms = list(split = "gini"))  
##   
## Variables actually used in tree construction:  
## [1] BusinessTravel DailyRate TotalWorkingYears  
##   
## Root node error: 86/843 = 0.10202  
##   
## n= 843   
##   
## CP nsplit rel error xerror xstd  
## 1 0.034884 0 1.00000 1.0000 0.10218  
## 2 0.011628 3 0.89535 1.1512 0.10869  
## 3 0.010000 4 0.88372 1.1860 0.11010

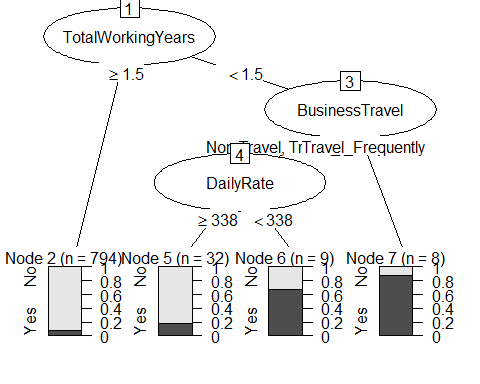
plotcp(HR\_DT, minline=TRUE, col="red")



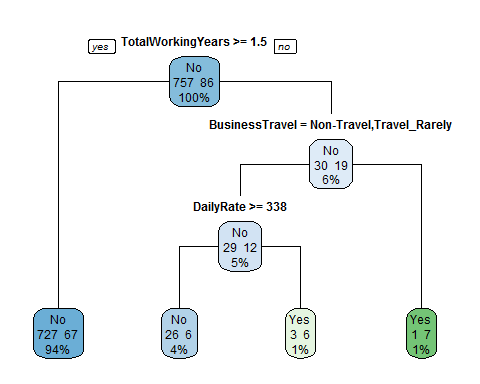
set.seed(123)  
Pruned\_HR\_DT <-prune(HR\_DT,cp=.012, minsplit=10, minbucket=round(minsplit/3))   
rpart.plot(Pruned\_HR\_DT, type=1, extra=101)



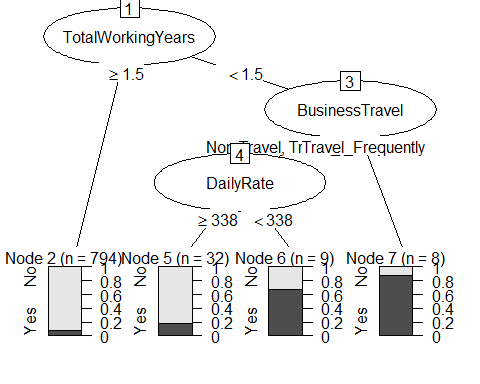
Pruned\_HR\_DT\_party<-as.party(Pruned\_HR\_DT)  
plot(Pruned\_HR\_DT\_party)



set.seed(123)  
Pruned\_HR\_DT2 <-prune(HR\_DT,cp=.025, 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)



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 183 21  
## Yes 4 3  
##   
## Accuracy : 0.8815   
## 95% CI : (0.8301, 0.9218)  
## No Information Rate : 0.8863   
## P-Value [Acc > NIR] : 0.636850   
##   
## Kappa : 0.1499   
## Mcnemar's Test P-Value : 0.001374   
##   
## Sensitivity : 0.12500   
## Specificity : 0.97861   
## Pos Pred Value : 0.42857   
## Neg Pred Value : 0.89706   
## Prevalence : 0.11374   
## Detection Rate : 0.01422   
## Detection Prevalence : 0.03318   
## Balanced Accuracy : 0.55180   
##   
## '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 187 22  
## Yes 0 2  
##   
## Accuracy : 0.8957   
## 95% CI : (0.8464, 0.9335)  
## No Information Rate : 0.8863   
## P-Value [Acc > NIR] : 0.3822   
##   
## Kappa : 0.1388   
## Mcnemar's Test P-Value : 0.000007562   
##   
## Sensitivity : 0.083333   
## Specificity : 1.000000   
## Pos Pred Value : 1.000000   
## Neg Pred Value : 0.894737   
## Prevalence : 0.113744   
## Detection Rate : 0.009479   
## Detection Prevalence : 0.009479   
## Balanced Accuracy : 0.541667   
##   
## '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 187 22  
## Yes 0 2  
##   
## Accuracy : 0.8957   
## 95% CI : (0.8464, 0.9335)  
## No Information Rate : 0.8863   
## P-Value [Acc > NIR] : 0.3822   
##   
## Kappa : 0.1388   
## Mcnemar's Test P-Value : 0.000007562   
##   
## Sensitivity : 0.083333   
## Specificity : 1.000000   
## Pos Pred Value : 1.000000   
## Neg Pred Value : 0.894737   
## Prevalence : 0.113744   
## Detection Rate : 0.009479   
## Detection Prevalence : 0.009479   
## Balanced Accuracy : 0.541667   
##   
## 'Positive' Class : Yes   
##

#########################################################################################  
#Logistic Regression  
  
set.seed(123)  
HR\_train\_logit<-HR\_train  
HR\_train\_logit$TotalWorkingYears<-ifelse(HR\_train\_logit$TotalWorkingYears>=1.5,"no","yes")  
HR.logit <- glm(HR\_train\_logit$Attrition~., data=HR\_train\_logit, family=binomial())  
summary(HR.logit)

##   
## Call:  
## glm(formula = HR\_train\_logit$Attrition ~ ., family = binomial(),   
## data = HR\_train\_logit)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6950 -0.4003 -0.2269 -0.0843 3.3313   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) 1.14058257 2451.86502346 0.000  
## Age -0.03675603 0.01904791 -1.930  
## BusinessTravelTravel\_Frequently 2.19400754 0.70309553 3.120  
## BusinessTravelTravel\_Rarely 1.34340704 0.66462785 2.021  
## DailyRate -0.00037832 0.00035276 -1.072  
## DepartmentResearch & Development 1.11044714 2451.86401589 0.000  
## DepartmentSales 16.46143361 2266.53393627 0.007  
## DistanceFromHome 0.03500199 0.01628249 2.150  
## Education 0.03038184 0.14249374 0.213  
## EducationFieldLife Sciences -1.45971201 1.16630446 -1.252  
## EducationFieldMarketing -0.64618145 1.23814490 -0.522  
## EducationFieldMedical -1.32126411 1.15458323 -1.144  
## EducationFieldOther -1.41216579 1.26753768 -1.114  
## EducationFieldTechnical Degree -0.39686658 1.18042693 -0.336  
## EnvironmentSatisfaction -0.49040982 0.13154084 -3.728  
## GenderMale 0.44119766 0.29629989 1.489  
## HourlyRate 0.00168440 0.00707045 0.238  
## JobInvolvement -0.57239662 0.19277523 -2.969  
## JobLevel -0.10975129 0.46784870 -0.235  
## JobRoleHuman Resources 1.94397310 2451.86408847 0.001  
## JobRoleLaboratory Technician 1.00792949 0.66845576 1.508  
## JobRoleManager -16.90581901 935.12742092 -0.018  
## JobRoleManufacturing Director -0.43041022 0.79053242 -0.544  
## JobRoleResearch Director -16.59596847 916.90332638 -0.018  
## JobRoleResearch Scientist 0.28871949 0.71503083 0.404  
## JobRoleSales Executive -15.00360789 935.12782919 -0.016  
## JobRoleSales Representative -13.94995896 935.12797736 -0.015  
## JobSatisfaction -0.36537637 0.12605961 -2.898  
## MaritalStatusMarried -0.17602452 0.44161670 -0.399  
## MaritalStatusSingle 0.11488168 0.56482239 0.203  
## MonthlyIncome 0.00009987 0.00012421 0.804  
## MonthlyRate 0.00001391 0.00001989 0.699  
## NumCompaniesWorked 0.18937737 0.05759868 3.288  
## PercentSalaryHike -0.02234398 0.06051853 -0.369  
## PerformanceRating 0.12052054 0.61035257 0.197  
## RelationshipSatisfaction -0.12986659 0.13070998 -0.994  
## StockOptionLevel -0.57642688 0.28186386 -2.045  
## TotalWorkingYearsyes 1.96769857 0.50618616 3.887  
## TrainingTimesLastYear -0.14452541 0.10954385 -1.319  
## WorkLifeBalance -0.33434243 0.19165010 -1.745  
## YearsAtCompany 0.09392137 0.05298113 1.773  
## YearsInCurrentRole -0.11736540 0.07007599 -1.675  
## YearsSinceLastPromotion 0.13020570 0.06245815 2.085  
## YearsWithCurrManager -0.10622779 0.07108283 -1.494  
## Pr(>|z|)   
## (Intercept) 0.999629   
## Age 0.053649 .   
## BusinessTravelTravel\_Frequently 0.001805 \*\*   
## BusinessTravelTravel\_Rarely 0.043250 \*   
## DailyRate 0.283507   
## DepartmentResearch & Development 0.999639   
## DepartmentSales 0.994205   
## DistanceFromHome 0.031581 \*   
## Education 0.831159   
## EducationFieldLife Sciences 0.210726   
## EducationFieldMarketing 0.601744   
## EducationFieldMedical 0.252472   
## EducationFieldOther 0.265236   
## EducationFieldTechnical Degree 0.736716   
## EnvironmentSatisfaction 0.000193 \*\*\*  
## GenderMale 0.136481   
## HourlyRate 0.811702   
## JobInvolvement 0.002985 \*\*   
## JobLevel 0.814529   
## JobRoleHuman Resources 0.999367   
## JobRoleLaboratory Technician 0.131594   
## JobRoleManager 0.985576   
## JobRoleManufacturing Director 0.586128   
## JobRoleResearch Director 0.985559   
## JobRoleResearch Scientist 0.686370   
## JobRoleSales Executive 0.987199   
## JobRoleSales Representative 0.988098   
## JobSatisfaction 0.003750 \*\*   
## MaritalStatusMarried 0.690194   
## MaritalStatusSingle 0.838827   
## MonthlyIncome 0.421401   
## MonthlyRate 0.484334   
## NumCompaniesWorked 0.001009 \*\*   
## PercentSalaryHike 0.711972   
## PerformanceRating 0.843467   
## RelationshipSatisfaction 0.320443   
## StockOptionLevel 0.040850 \*   
## TotalWorkingYearsyes 0.000101 \*\*\*  
## TrainingTimesLastYear 0.187056   
## WorkLifeBalance 0.081064 .   
## YearsAtCompany 0.076273 .   
## YearsInCurrentRole 0.093967 .   
## YearsSinceLastPromotion 0.037098 \*   
## YearsWithCurrManager 0.135065   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 555.52 on 842 degrees of freedom  
## Residual deviance: 378.37 on 799 degrees of freedom  
## AIC: 466.37  
##   
## Number of Fisher Scoring iterations: 17

names(HR\_train\_logit)

## [1] "Attrition" "Age"   
## [3] "BusinessTravel" "DailyRate"   
## [5] "Department" "DistanceFromHome"   
## [7] "Education" "EducationField"   
## [9] "EnvironmentSatisfaction" "Gender"   
## [11] "HourlyRate" "JobInvolvement"   
## [13] "JobLevel" "JobRole"   
## [15] "JobSatisfaction" "MaritalStatus"   
## [17] "MonthlyIncome" "MonthlyRate"   
## [19] "NumCompaniesWorked" "PercentSalaryHike"   
## [21] "PerformanceRating" "RelationshipSatisfaction"  
## [23] "StockOptionLevel" "TotalWorkingYears"   
## [25] "TrainingTimesLastYear" "WorkLifeBalance"   
## [27] "YearsAtCompany" "YearsInCurrentRole"   
## [29] "YearsSinceLastPromotion" "YearsWithCurrManager"

HR\_train\_logit2<-HR\_train\_logit[,c(1,3, 6, 9, 12, 15, 19, 23:24)]  
HR.logit2 <- glm(HR\_train\_logit2$Attrition~., data=HR\_train\_logit2, family=binomial())  
summary(HR.logit2)

##   
## Call:  
## glm(formula = HR\_train\_logit2$Attrition ~ ., family = binomial(),   
## data = HR\_train\_logit2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.2684 -0.4456 -0.3093 -0.1930 3.0299   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -0.47916 0.88400 -0.542 0.587792  
## BusinessTravelTravel\_Frequently 2.04913 0.65254 3.140 0.001688  
## BusinessTravelTravel\_Rarely 1.17021 0.62573 1.870 0.061462  
## DistanceFromHome 0.03196 0.01470 2.173 0.029751  
## EnvironmentSatisfaction -0.41581 0.11339 -3.667 0.000245  
## JobInvolvement -0.62491 0.17417 -3.588 0.000333  
## JobSatisfaction -0.35022 0.11466 -3.055 0.002254  
## NumCompaniesWorked 0.13837 0.05002 2.767 0.005666  
## StockOptionLevel -0.61891 0.18198 -3.401 0.000671  
## TotalWorkingYearsyes 2.45024 0.39661 6.178 0.000000000649  
##   
## (Intercept)   
## BusinessTravelTravel\_Frequently \*\*   
## BusinessTravelTravel\_Rarely .   
## DistanceFromHome \*   
## EnvironmentSatisfaction \*\*\*  
## JobInvolvement \*\*\*  
## JobSatisfaction \*\*   
## NumCompaniesWorked \*\*   
## StockOptionLevel \*\*\*  
## TotalWorkingYearsyes \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 555.52 on 842 degrees of freedom  
## Residual deviance: 445.13 on 833 degrees of freedom  
## AIC: 465.13  
##   
## Number of Fisher Scoring iterations: 6

HR\_train\_logit3<-HR\_train\_logit[,c(1, 12, 14, 17, 24)]  
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   
## -1.3963 -0.4778 -0.3650 -0.2192 2.8626   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.96587411 0.80655435 -2.437 0.014794  
## JobInvolvement -0.54627114 0.16023477 -3.409 0.000652  
## JobRoleHuman Resources 1.25825847 0.66592263 1.889 0.058825  
## JobRoleLaboratory Technician 0.97647739 0.57766419 1.690 0.090953  
## JobRoleManager -2.85104350 1.26436383 -2.255 0.024138  
## JobRoleManufacturing Director -0.81619064 0.73267123 -1.114 0.265282  
## JobRoleResearch Director -16.14978192 601.74607350 -0.027 0.978589  
## JobRoleResearch Scientist 0.21135258 0.62588424 0.338 0.735600  
## JobRoleSales Executive 0.41753247 0.49743724 0.839 0.401263  
## JobRoleSales Representative 1.67917670 0.65024959 2.582 0.009813  
## MonthlyIncome 0.00011910 0.00006444 1.848 0.064592  
## TotalWorkingYearsyes 1.73758897 0.37218259 4.669 0.00000303  
##   
## (Intercept) \*   
## JobInvolvement \*\*\*  
## JobRoleHuman Resources .   
## JobRoleLaboratory Technician .   
## JobRoleManager \*   
## JobRoleManufacturing Director   
## JobRoleResearch Director   
## JobRoleResearch Scientist   
## JobRoleSales Executive   
## JobRoleSales Representative \*\*   
## MonthlyIncome .   
## TotalWorkingYearsyes \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 555.52 on 842 degrees of freedom  
## Residual deviance: 475.82 on 831 degrees of freedom  
## AIC: 499.82  
##   
## Number of Fisher Scoring iterations: 16

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

## Odds\_Ratio  
## (Intercept) 0.14003342864965  
## JobInvolvement 0.57910519148840  
## JobRoleHuman Resources 3.51928719843915  
## JobRoleLaboratory Technician 2.65508690998028  
## JobRoleManager 0.05778399193294  
## JobRoleManufacturing Director 0.44211261630589  
## JobRoleResearch Director 0.00000009688105  
## JobRoleResearch Scientist 1.23534783233335  
## JobRoleSales Executive 1.51821069249722  
## JobRoleSales Representative 5.36114030505939  
## MonthlyIncome 1.00011910357905  
## TotalWorkingYearsyes 5.68362350445205

prob<-odds/(1+odds)  
prob

## Odds\_Ratio  
## (Intercept) 0.12283273905004  
## JobInvolvement 0.36672996492562  
## JobRoleHuman Resources 0.77872616718287  
## JobRoleLaboratory Technician 0.72640869434062  
## JobRoleManager 0.05462740254496  
## JobRoleManufacturing Director 0.30657287878002  
## JobRoleResearch Director 0.00000009688104  
## JobRoleResearch Scientist 0.55264232906601  
## JobRoleSales Executive 0.60289264000848  
## JobRoleSales Representative 0.84279548130629  
## MonthlyIncome 0.50002977412166  
## TotalWorkingYearsyes 0.85038056088379

anova(HR.logit2,test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: HR\_train\_logit2$Attrition  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)  
## NULL 842 555.52   
## BusinessTravel 2 14.616 840 540.91 0.0006703  
## DistanceFromHome 1 2.415 839 538.49 0.1201648  
## EnvironmentSatisfaction 1 13.225 838 525.27 0.0002762  
## JobInvolvement 1 15.061 837 510.21 0.0001041  
## JobSatisfaction 1 10.027 836 500.18 0.0015426  
## NumCompaniesWorked 1 1.853 835 498.33 0.1734796  
## StockOptionLevel 1 17.643 834 480.68 0.000026653003  
## TotalWorkingYears 1 35.555 833 445.13 0.000000002479  
##   
## NULL   
## BusinessTravel \*\*\*  
## DistanceFromHome   
## EnvironmentSatisfaction \*\*\*  
## JobInvolvement \*\*\*  
## JobSatisfaction \*\*   
## NumCompaniesWorked   
## StockOptionLevel \*\*\*  
## TotalWorkingYears \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

HR\_test\_logit<-HR\_test[,c(1,3, 6, 9, 12, 15, 19, 23:24)]  
HR\_test\_logit$TotalWorkingYears<-ifelse(HR\_test\_logit$TotalWorkingYears>=1.5,"no","yes")  
HR\_test\_logit$predict.Attrition<-predict(HR.logit2, newdata=HR\_test\_logit,type = "response")  
  
HR\_test\_logit\_CI<-cbind(HR\_test\_logit,predict(HR.logit2, newdata=HR\_test\_logit,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))  
 })   
  
summary(HR\_test\_logit\_CI)

## Attrition BusinessTravel DistanceFromHome  
## No :187 Non-Travel : 29 Min. : 1.000   
## Yes: 24 Travel\_Frequently: 34 1st Qu.: 2.000   
## Travel\_Rarely :148 Median : 7.000   
## Mean : 9.128   
## 3rd Qu.:13.000   
## Max. :29.000   
## EnvironmentSatisfaction JobInvolvement JobSatisfaction  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.000   
## Median :3.000 Median :3.000 Median :3.000   
## Mean :2.739 Mean :2.711 Mean :2.768   
## 3rd Qu.:4.000 3rd Qu.:3.000 3rd Qu.:4.000   
## Max. :4.000 Max. :4.000 Max. :4.000   
## NumCompaniesWorked StockOptionLevel TotalWorkingYears   
## Min. :0.000 Min. :0.0000 Length:211   
## 1st Qu.:1.000 1st Qu.:0.0000 Class :character   
## Median :2.000 Median :1.0000 Mode :character   
## Mean :2.739 Mean :0.7014   
## 3rd Qu.:4.000 3rd Qu.:1.0000   
## Max. :9.000 Max. :3.0000   
## predict.Attrition fit se.fit residual.scale  
## Min. :0.0008061 Min. :-7.1225 Min. :0.2234 Min. :1   
## 1st Qu.:0.0281378 1st Qu.:-3.5422 1st Qu.:0.3260 1st Qu.:1   
## Median :0.0644539 Median :-2.6752 Median :0.3641 Median :1   
## Mean :0.0990073 Mean :-2.7314 Mean :0.4117 Mean :1   
## 3rd Qu.:0.1273402 3rd Qu.:-1.9247 3rd Qu.:0.4392 3rd Qu.:1   
## Max. :0.6795641 Max. : 0.7518 Max. :0.8539 Max. :1   
## UL LL PredictedProb   
## Min. :0.004283 Min. :0.0001513 Min. :0.0008061   
## 1st Qu.:0.065663 1st Qu.:0.0123585 1st Qu.:0.0281378   
## Median :0.122803 Median :0.0327513 Median :0.0644539   
## Mean :0.174118 Mean :0.0533037 Mean :0.0990073   
## 3rd Qu.:0.215534 3rd Qu.:0.0683976 3rd Qu.:0.1273402   
## Max. :0.829861 Max. :0.4797356 Max. :0.6795641

dim(HR\_test\_logit\_CI)

## [1] 211 16

HR\_test\_logit\_CI$predict.Attrition<-ifelse(HR\_test\_logit\_CI$predict.Attrition>.5, "yes", "no")  
probs<-HR\_test\_logit\_CI[,c(1,10)]   
  
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: 211   
##   
##   
## | probs$predict.Attrition   
## probs$Attrition | no | yes | Row Total |   
## ----------------|-----------|-----------|-----------|  
## No | 187 | 0 | 187 |   
## | 0.017 | 1.773 | |   
## | 1.000 | 0.000 | 0.886 |   
## | 0.895 | 0.000 | |   
## | 0.886 | 0.000 | |   
## ----------------|-----------|-----------|-----------|  
## Yes | 22 | 2 | 24 |   
## | 0.132 | 13.811 | |   
## | 0.917 | 0.083 | 0.114 |   
## | 0.105 | 1.000 | |   
## | 0.104 | 0.009 | |   
## ----------------|-----------|-----------|-----------|  
## Column Total | 209 | 2 | 211 |   
## | 0.991 | 0.009 | |   
## ----------------|-----------|-----------|-----------|  
##   
##

TP = 2  
TN = 187  
FP = 0  
FN = 22  
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))  
library(knitr)

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

kable(table)

|  |  |
| --- | --- |
| Measure | Value |
| True Positive | 2 |
| True Negative | 187 |
| False Positive | 0 |
| False Negative | 22 |
| Sensitivity=TP/(TN+FP) | 0.083 |
| Specificity=TN/(TN+TP) | 1 |
| Precision=TP/(TP+FP) | 1 |
| Accuracy=(TP+TN)/total | 0.896 |