churn\_pre\_decisionTree.R

John Mohd Bhat

Tue Jan 17 21:57:51 2017

#########################################################################  
# Churn prediction #  
# using decision Tree #  
  
  
setwd("C:/Users/admin/Desktop/R/programs/R-prc")  
  
churn\_data<-read.csv("churn.csv",stringsAsFactors=T)  
  
str(churn\_data)

## 'data.frame': 7043 obs. of 20 variables:  
## $ customerID : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 1003 4771 5605 4535 ...  
## $ gender : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...  
## $ SeniorCitizen : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Partner : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...  
## $ Dependents : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...  
## $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...  
## $ MultipleLines : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...  
## $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...  
## $ OnlineSecurity : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...  
## $ OnlineBackup : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...  
## $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...  
## $ TechSupport : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...  
## $ StreamingTV : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...  
## $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...  
## $ Contract : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...  
## $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...  
## $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...  
## $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ..

#head(churn\_data)  
#checking number of NA's  
  
sapply(churn\_data,function(x) sum(is.na(x)))

## customerID gender SeniorCitizen Partner   
## 0 0 0 0   
## Dependents PhoneService MultipleLines InternetService   
## 0 0 0 0   
## OnlineSecurity OnlineBackup DeviceProtection TechSupport   
## 0 0 0 0   
## StreamingTV StreamingMovies Contract PaperlessBilling   
## 0 0 0 0   
## PaymentMethod MonthlyCharges TotalCharges Churn   
## 0 0 11 0

summary(churn\_data)

## customerID gender SeniorCitizen Partner Dependents  
## 0002-ORFBO: 1 Female:3488 Min. :0.0000 No :3641 No :4933   
## 0003-MKNFE: 1 Male :3555 1st Qu.:0.0000 Yes:3402 Yes:2110   
## 0004-TLHLJ: 1 Median :0.0000   
## 0011-IGKFF: 1 Mean :0.1621   
## 0013-EXCHZ: 1 3rd Qu.:0.0000   
## 0013-MHZWF: 1 Max. :1.0000   
## (Other) :7037   
## PhoneService MultipleLines InternetService  
## No : 682 No :3390 DSL :2421   
## Yes:6361 No phone service: 682 Fiber optic:3096   
## Yes :2971 No :1526   
##   
##   
##   
##   
## OnlineSecurity OnlineBackup   
## No :3498 No :3088   
## No internet service:1526 No internet service:1526   
## Yes :2019 Yes :2429   
##   
##   
##   
##   
## DeviceProtection TechSupport   
## No :3095 No :3473   
## No internet service:1526 No internet service:1526   
## Yes :2422 Yes :2044   
##   
##   
##   
##   
## StreamingTV StreamingMovies  
## No :2810 No :2785   
## No internet service:1526 No internet service:1526   
## Yes :2707 Yes :2732   
##   
##   
##   
##   
## Contract PaperlessBilling PaymentMethod   
## Month-to-month:3875 No :2872 Bank transfer (automatic):1544   
## One year :1473 Yes:4171 Credit card (automatic) :1522   
## Two year :1695 Electronic check :2365   
## Mailed check :1612   
##   
##   
##   
## MonthlyCharges TotalCharges Churn   
## Min. : 18.25 Min. : 18.8 No :5174   
## 1st Qu.: 35.50 1st Qu.: 401.4 Yes:1869   
## Median : 70.35 Median :1397.5   
## Mean : 64.76 Mean :2283.3   
## 3rd Qu.: 89.85 3rd Qu.:3794.7   
## Max. :118.75 Max. :8684.8   
## NA's :11

#deleting the rows which are having NA's in total charges column #11 rows  
  
churn\_data<-churn\_data[-(which(is.na(churn\_data$TotalCharges)==TRUE)),]  
  
  
#converting Senior citizen column into factor  
churn\_data["SeniorCitizen"]<-as.factor(churn\_data[,"SeniorCitizen"])  
  
#levels(churn\_data$Churn)<-c()  
  
  
#deleting the Id column  
churn\_data<-churn\_data[,-1]  
  
#scaling total charges  
#churn\_data$TotalCharges<-scale(churn\_data$TotalCharges,0,scale=TRUE)  
  
minmax =function(x){  
 xnew <-(x - min(x))/(max(x)-min(x))  
}  
  
churn\_data[,17:18]<-apply(churn\_data[,17:18],2,minmax)  
  
  
#check the no. of positive and negative class  
table(churn\_data[,19])

##   
## No Yes   
## 5163 1869

#data Partition  
  
#subset the data based on classes  
class1<-subset(churn\_data,Churn=="Yes")  
class0<-subset(churn\_data,Churn=="No")  
set.seed(234)  
s1<-sample(nrow(class1),0.7\*nrow(class1))  
s0<-sample(nrow(class0),0.7\*nrow(class0))  
train1<-class1[s1,]  
test1<-class1[-s1,]  
  
train0<-class0[s0,]  
test0<-class0[-s0,]  
  
train<-rbind(train1,train0)  
test<-rbind(test1,test0)  
  
  
################### MODEL DEVELOPMENT####################################  
#decision tree Algorith   
#rpart pkg  
  
library(rpart)  
  
fit= rpart(Churn ~ gender+SeniorCitizen+Partner+Dependents+  
 PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup+  
 DeviceProtection+TechSupport+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+ PaymentMethod+MonthlyCharges+TotalCharges,  
 method = "class",control=rpart.control(minsplit = 10), cp=0.01000000,train)  
  
  
#fit= rpart(Churn ~.,method = "class",control=rpart.control(minsplit = #2),train)  
  
fit

## n= 4922   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 4922 1308 No (0.73425437 0.26574563)   
## 2) Contract=One year,Two year 2210 147 No (0.93348416 0.06651584) \*  
## 3) Contract=Month-to-month 2712 1161 No (0.57190265 0.42809735)   
## 6) InternetService=DSL,No 1223 341 No (0.72117743 0.27882257) \*  
## 7) InternetService=Fiber optic 1489 669 Yes (0.44929483 0.55070517)   
## 14) TotalCharges>=0.1749798 728 296 No (0.59340659 0.40659341)   
## 28) PaymentMethod=Bank transfer (automatic),Credit card (automatic),Mailed check 327 106 No (0.67584098 0.32415902) \*  
## 29) PaymentMethod=Electronic check 401 190 No (0.52618454 0.47381546)   
## 58) TotalCharges>=0.353658 219 84 No (0.61643836 0.38356164) \*  
## 59) TotalCharges< 0.353658 182 76 Yes (0.41758242 0.58241758) \*  
## 15) TotalCharges< 0.1749798 761 237 Yes (0.31143233 0.68856767) \*

#rpart.plot(fit)  
fit$cptable

## CP nsplit rel error xerror xstd  
## 1 0.05772171 0 1.0000000 1.0000000 0.02369296  
## 2 0.01146789 3 0.7805810 0.7966361 0.02191145  
## 3 0.01000000 5 0.7576453 0.7951070 0.02189606

fit$variable.importance

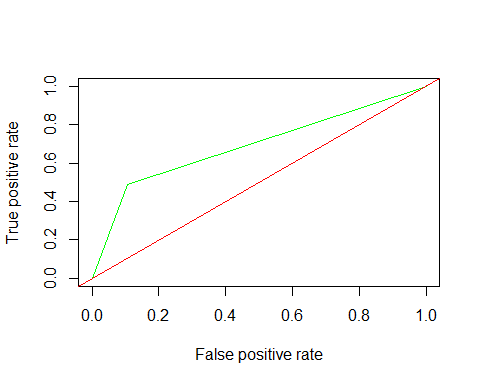
## Contract TotalCharges TechSupport OnlineSecurity   
## 318.4071186 166.4389239 145.5106913 135.8450270   
## MonthlyCharges InternetService DeviceProtection PaymentMethod   
## 117.7715037 99.2714365 97.8696279 89.9032429   
## MultipleLines OnlineBackup StreamingTV Partner   
## 51.9750179 45.5344445 13.8042419 13.7349035   
## StreamingMovies   
## 0.6661892

p<-predict(fit,test,type="class")  
head(p)

## 3 19 23 56 96 98   
## No No No Yes Yes No   
## Levels: No Yes

library(ROCR)

pr<-prediction(as.numeric(p),as.numeric(test$Churn))  
  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf,col="green")  
abline(0,1,col="red")  
  
  
library(caret)



confusionMatrix(p,test$Churn)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1383 287  
## Yes 166 274  
##   
## Accuracy : 0.7853   
## 95% CI : (0.7672, 0.8027)  
## No Information Rate : 0.7341   
## P-Value [Acc > NIR] : 3.089e-08   
##   
## Kappa : 0.4094   
## Mcnemar's Test P-Value : 1.719e-08   
##   
## Sensitivity : 0.8928   
## Specificity : 0.4884   
## Pos Pred Value : 0.8281   
## Neg Pred Value : 0.6227   
## Prevalence : 0.7341   
## Detection Rate : 0.6555   
## Detection Prevalence : 0.7915   
## Balanced Accuracy : 0.6906   
##   
## 'Positive' Class : No   
##

d<-table(p,test$Churn)  
acc<-sum(diag(d))/sum(d)  
acc

## [1] 0.7853081

#################################################################################  
##using information gain split  
  
fit1= rpart(Churn ~ gender+SeniorCitizen+Partner+Dependents+  
 PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup+  
 DeviceProtection+TechSupport+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+ PaymentMethod+MonthlyCharges+TotalCharges,  
 method = "class",control=rpart.control(minsplit = 10),parms = list(split="information"),train)

#rpart.plot(fit1)  
  
fit1$cptable

## CP nsplit rel error xerror xstd  
## 1 0.05772171 0 1.0000000 1.0000000 0.02369296  
## 2 0.01146789 3 0.7805810 0.7951070 0.02189606  
## 3 0.01000000 5 0.7576453 0.7714067 0.02165319

fit1$variable.importance

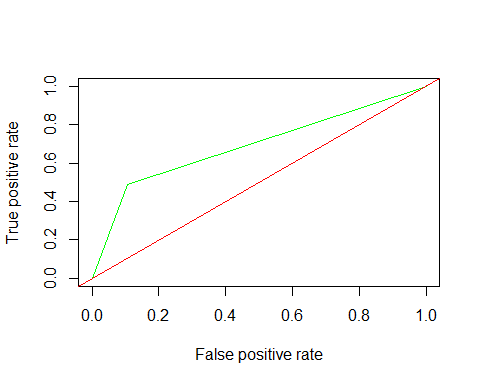
## Contract TotalCharges TechSupport OnlineSecurity   
## 457.654939 211.404185 197.013891 183.297470   
## DeviceProtection PaymentMethod MonthlyCharges InternetService   
## 140.096262 126.056075 122.234855 103.431491   
## MultipleLines OnlineBackup StreamingTV Partner   
## 53.808895 47.129486 14.130373 14.065418   
## StreamingMovies   
## 0.696265

p<-predict(fit1,test,type="class")

d<-table(p,test$Churn)  
acc<-sum(diag(d))/sum(d)  
acc

## [1] 0.7853081

library(ROCR)  
pr<-prediction(as.numeric(p),as.numeric(test$Churn))  
  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf,col="green")  
abline(0,1,col="red")



############################################################################## using loss matrix  
  
  
  
lossmatrix=matrix(c(0,0.2,0.6,0),nrow=2,byrow=TRUE)  
fit2= rpart(Churn ~ gender+SeniorCitizen+Partner+Dependents+  
 PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup+  
 DeviceProtection+TechSupport+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+ PaymentMethod+MonthlyCharges+TotalCharges,  
 method = "class",control=rpart.control(minsplit = 10),parms = list(loss=lossmatrix,split="information"),train)

#rpart.plot(fit2)  
fit2$cptable

## CP nsplit rel error xerror xstd  
## 1 0.44881018 0 1.0000000 3.0000000 0.02572529  
## 2 0.02434975 1 0.5511898 1.3281682 0.02686284  
## 3 0.01162147 3 0.5024903 0.9277809 0.02353892  
## 4 0.01000000 4 0.4908688 0.8904261 0.02316057

fit2$variable.importance

## Contract TechSupport OnlineSecurity TotalCharges   
## 616.51552 256.61230 248.28754 234.63904   
## DeviceProtection PaymentMethod InternetService MonthlyCharges   
## 197.07124 158.45286 121.06263 115.82127   
## OnlineBackup MultipleLines PhoneService   
## 43.07982 37.08717 14.48072

cp1<-fit2$cptable[which.min(fit$cptable[,"xerror"])]  
prune(fit2,cp=cp1)

## n= 4922   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 4922 722.8 Yes (0.73425437 0.26574563)   
## 2) Contract=One year,Two year 2210 88.2 No (0.93348416 0.06651584) \*  
## 3) Contract=Month-to-month 2712 310.2 Yes (0.57190265 0.42809735)   
## 6) InternetService=DSL,No 1223 176.4 Yes (0.72117743 0.27882257)   
## 12) TotalCharges>=0.02362393 688 76.8 No (0.81395349 0.18604651) \*  
## 13) TotalCharges< 0.02362393 535 64.4 Yes (0.60186916 0.39813084) \*  
## 7) InternetService=Fiber optic 1489 133.8 Yes (0.44929483 0.55070517) \*

p<-predict(fit2,test,type="class")

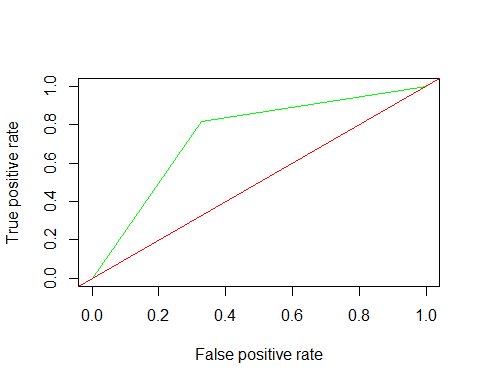
d<-table(p,test$Churn)  
acc<-sum(diag(d))/sum(d)  
acc

## [1] 0.7118483

confusionMatrix(p,test$Churn,positive="Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1044 103  
## Yes 505 458  
##   
## Accuracy : 0.7118   
## 95% CI : (0.692, 0.7311)  
## No Information Rate : 0.7341   
## P-Value [Acc > NIR] : 0.9899   
##   
## Kappa : 0.3992   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.8164   
## Specificity : 0.6740   
## Pos Pred Value : 0.4756   
## Neg Pred Value : 0.9102   
## Prevalence : 0.2659   
## Detection Rate : 0.2171   
## Detection Prevalence : 0.4564   
## Balanced Accuracy : 0.7452   
##   
## 'Positive' Class : Yes   
##

library(ROCR)  
pr<-prediction(as.numeric(p),as.numeric(test$Churn))  
  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf,col="green")  
abline(0,1,col="red")



#############################################################################  
# Using surrogate parameter  
  
fit3= rpart(Churn ~ gender+SeniorCitizen+Partner+Dependents+  
 PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup+  
 DeviceProtection+TechSupport+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+ PaymentMethod+MonthlyCharges+TotalCharges,  
 method = "class",control=rpart.control(minsplit = 10,maxsurrogate = 10),parms = list(split="information"),train)

#rpart.plot(fit3)  
fit3$cptable

## CP nsplit rel error xerror xstd  
## 1 0.05772171 0 1.0000000 1.0000000 0.02369296  
## 2 0.01146789 3 0.7805810 0.7966361 0.02191145  
## 3 0.01000000 5 0.7576453 0.7958716 0.02190376

fit3$variable.importance

## Contract TotalCharges TechSupport OnlineSecurity   
## 457.654939 241.850086 206.418579 191.953112   
## DeviceProtection PaymentMethod OnlineBackup StreamingTV   
## 183.279978 166.223438 156.262587 130.769718   
## StreamingMovies MonthlyCharges Partner InternetService   
## 128.587897 122.234855 106.042642 103.431491   
## Dependents MultipleLines PaperlessBilling   
## 77.889252 53.808895 0.243561

#to get the cp value having minimum xerror against it  
  
fit3$cptable[which.min(fit$cptable[,"xerror"])]

## [1] 0.01

prune(fit3,cp=0.01)

## n= 4922   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 4922 1308 No (0.73425437 0.26574563)   
## 2) Contract=One year,Two year 2210 147 No (0.93348416 0.06651584) \*  
## 3) Contract=Month-to-month 2712 1161 No (0.57190265 0.42809735)   
## 6) InternetService=DSL,No 1223 341 No (0.72117743 0.27882257) \*  
## 7) InternetService=Fiber optic 1489 669 Yes (0.44929483 0.55070517)   
## 14) TotalCharges>=0.1749798 728 296 No (0.59340659 0.40659341)   
## 28) PaymentMethod=Bank transfer (automatic),Credit card (automatic),Mailed check 327 106 No (0.67584098 0.32415902) \*  
## 29) PaymentMethod=Electronic check 401 190 No (0.52618454 0.47381546)   
## 58) TotalCharges>=0.353658 219 84 No (0.61643836 0.38356164) \*  
## 59) TotalCharges< 0.353658 182 76 Yes (0.41758242 0.58241758) \*  
## 15) TotalCharges< 0.1749798 761 237 Yes (0.31143233 0.68856767) \*

p<-predict(fit3,test,type="class")  
head(p)

## 3 19 23 56 96 98   
## No No No Yes Yes No   
## Levels: No Yes

#View(p)  
nrow(test)

## [1] 2110

length(p)

## [1] 2110

d<-table(p,test$Churn)  
acc<-sum(diag(d))/sum(d)  
acc

## [1] 0.7853081

library(ROCR)  
pr<-prediction(as.numeric(p),as.numeric(test$Churn))  
  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf,col="green")  
abline(0,1,col="red")

