HW5

Sanduni Talagala

March 5, 2020

# 1)

get the file

crime<-read.csv("crimedata.csv")

## 2) build regression tree

vriables: ExpenditureYear, StateSize, BelowWage, Education, YouthUnemployment,Youth Note: Im not making a training dataset since I have another dataset to test on.

library(rpart)  
  
#summary of dataset  
summary(crime)

## CrimeRate Youth Southern Education   
## Min. : 45.5 Min. :119.0 Min. :0.0000 Min. :10.00   
## 1st Qu.: 82.7 1st Qu.:130.0 1st Qu.:0.0000 1st Qu.:11.55   
## Median :103.0 Median :136.0 Median :0.0000 Median :12.40   
## Mean :102.8 Mean :138.6 Mean :0.3404 Mean :12.39   
## 3rd Qu.:120.7 3rd Qu.:146.0 3rd Qu.:1.0000 3rd Qu.:13.20   
## Max. :161.8 Max. :177.0 Max. :1.0000 Max. :15.10   
## ExpenditureYear LabourForce Males MoreMales   
## Min. : 45.0 Min. :480.0 Min. : 934.0 Min. :0.0000   
## 1st Qu.: 62.5 1st Qu.:530.5 1st Qu.: 964.5 1st Qu.:0.0000   
## Median : 78.0 Median :560.0 Median : 977.0 Median :0.0000   
## Mean : 85.0 Mean :561.2 Mean : 983.0 Mean :0.1915   
## 3rd Qu.:104.5 3rd Qu.:593.0 3rd Qu.: 992.0 3rd Qu.:0.0000   
## Max. :166.0 Max. :641.0 Max. :1071.0 Max. :1.0000   
## StateSize YouthUnemployment MatureUnemployment HighYouthUnemploy  
## Min. : 3.00 Min. : 70.00 Min. :20.00 Min. :0.0000   
## 1st Qu.: 10.00 1st Qu.: 80.50 1st Qu.:27.50 1st Qu.:0.0000   
## Median : 25.00 Median : 92.00 Median :34.00 Median :0.0000   
## Mean : 36.62 Mean : 95.47 Mean :33.98 Mean :0.3191   
## 3rd Qu.: 41.50 3rd Qu.:104.00 3rd Qu.:38.50 3rd Qu.:1.0000   
## Max. :168.00 Max. :142.00 Max. :58.00 Max. :1.0000   
## Wage BelowWage   
## Min. :288.0 Min. :126.0   
## 1st Qu.:459.5 1st Qu.:165.5   
## Median :537.0 Median :176.0   
## Mean :525.4 Mean :194.0   
## 3rd Qu.:591.5 3rd Qu.:227.5   
## Max. :689.0 Max. :276.0

#Regression tree with whole dataset  
wholecrime<-rpart(CrimeRate ~ ExpenditureYear+StateSize+BelowWage+Education+YouthUnemployment+Youth, data = crime)

## 3) Print summary of tree

most important variables: ExpenditureYear was the most important variable, followed closely by Youth, then BelowWage, then StateSize, then YouthUnemployment and finally Education as the least important.

#summary of tree using all the data  
summary(wholecrime)

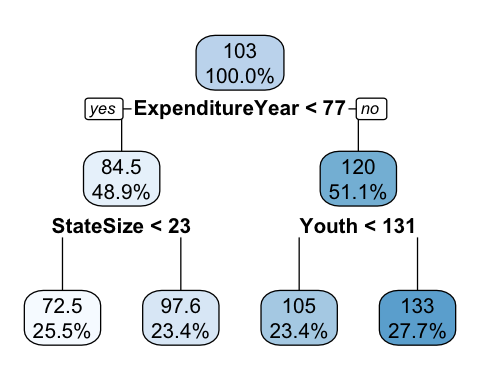
## Call:  
## rpart(formula = CrimeRate ~ ExpenditureYear + StateSize + BelowWage +   
## Education + YouthUnemployment + Youth, data = crime)  
## n= 47   
##   
## CP nsplit rel error xerror xstd  
## 1 0.39334682 0 1.0000000 1.0899008 0.1872031  
## 2 0.11890083 1 0.6066532 0.7648871 0.1331704  
## 3 0.09454168 2 0.4877524 0.9948696 0.1929716  
## 4 0.01000000 3 0.3932107 0.8841833 0.1923362  
##   
## Variable importance  
## ExpenditureYear Youth BelowWage StateSize   
## 27 21 17 16   
## YouthUnemployment Education   
## 10 9   
##   
## Node number 1: 47 observations, complexity param=0.3933468  
## mean=102.8085, MSE=817.0587   
## left son=2 (23 obs) right son=3 (24 obs)  
## Primary splits:  
## ExpenditureYear < 76.5 to the left, improve=0.39334680, (0 missing)  
## StateSize < 48.5 to the left, improve=0.11935310, (0 missing)  
## BelowWage < 248 to the right, improve=0.06924067, (0 missing)  
## Education < 13.65 to the left, improve=0.04244413, (0 missing)  
## YouthUnemployment < 104 to the right, improve=0.03701363, (0 missing)  
## Surrogate splits:  
## BelowWage < 171.5 to the right, agree=0.745, adj=0.478, (0 split)  
## Youth < 132.5 to the right, agree=0.745, adj=0.478, (0 split)  
## StateSize < 48.5 to the left, agree=0.681, adj=0.348, (0 split)  
## Education < 11.55 to the left, agree=0.638, adj=0.261, (0 split)  
## YouthUnemployment < 77.5 to the left, agree=0.638, adj=0.261, (0 split)  
##   
## Node number 2: 23 observations, complexity param=0.09454168  
## mean=84.49565, MSE=375.2891   
## left son=4 (12 obs) right son=5 (11 obs)  
## Primary splits:  
## StateSize < 22.5 to the left, improve=0.4206110, (0 missing)  
## Youth < 145 to the left, improve=0.3381542, (0 missing)  
## ExpenditureYear < 57.5 to the left, improve=0.2012354, (0 missing)  
## YouthUnemployment < 93 to the right, improve=0.1755920, (0 missing)  
## BelowWage < 215 to the left, improve=0.1001957, (0 missing)  
## Surrogate splits:  
## Youth < 145 to the left, agree=0.783, adj=0.545, (0 split)  
## ExpenditureYear < 56.5 to the left, agree=0.739, adj=0.455, (0 split)  
## BelowWage < 203 to the left, agree=0.739, adj=0.455, (0 split)  
## YouthUnemployment < 93 to the right, agree=0.739, adj=0.455, (0 split)  
## Education < 11.15 to the right, agree=0.696, adj=0.364, (0 split)  
##   
## Node number 3: 24 observations, complexity param=0.1189008  
## mean=120.3583, MSE=611.0374   
## left son=6 (11 obs) right son=7 (13 obs)  
## Primary splits:  
## Youth < 130.5 to the left, improve=0.31135580, (0 missing)  
## ExpenditureYear < 107.5 to the left, improve=0.15843150, (0 missing)  
## BelowWage < 165.5 to the left, improve=0.09890595, (0 missing)  
## YouthUnemployment < 104 to the right, improve=0.08606562, (0 missing)  
## StateSize < 21.5 to the right, improve=0.08245995, (0 missing)  
## Surrogate splits:  
## BelowWage < 172.5 to the left, agree=0.792, adj=0.545, (0 split)  
## StateSize < 50.5 to the right, agree=0.708, adj=0.364, (0 split)  
## YouthUnemployment < 78.5 to the left, agree=0.667, adj=0.273, (0 split)  
## ExpenditureYear < 88.5 to the right, agree=0.625, adj=0.182, (0 split)  
## Education < 11.8 to the right, agree=0.625, adj=0.182, (0 split)  
##   
## Node number 4: 12 observations  
## mean=72.46667, MSE=265.8972   
##   
## Node number 5: 11 observations  
## mean=97.61818, MSE=164.5742   
##   
## Node number 6: 11 observations  
## mean=105.3636, MSE=587.826   
##   
## Node number 7: 13 observations  
## mean=133.0462, MSE=279.4471

## 4)Plot my tree and describe

Description tree: The main and most important split is for yearly police expenditure and if it is greater than or less than 77. This shows that 51.1% of the data has early police expenditure above 77 and the mean crime rate in that group is 120. Then, of the ones with police expenditure above 77, if there is above 131 youths then the crime rate mean is 133 which inludes 27.7% of the data. Alternatively if there’s less than 131 youths then the mean crime rate is 105 which includes 23.4% of the data.

When the police expenditure is less than 77 the most important next factor is the state size (not youth like when expenditure is above 77). The mean crime rate when expenditure is below 77 is 84.5 which includes 48.9% of the data. In that group, if the state size is above 23 hundred thousand, the mean crime rate is 97.6 which includes 23.4% of the data. Alternatively, if the state size is below 23 hundred thousand, the mean crime rate is 72.5 which includes 25.5% of the data.

#install.packages("rpart.plot")  
library(rpart.plot)  
  
#Plot tree  
rpart.plot(wholecrime, digits = 3, fallen.leaves = TRUE,tweak=1.3)



## 5) Mean crime rate for each group

\*[Police expenditure above 77 - mean crime rate 120]

\*[Police expenditure below 77 - mean crime rate 84.5]

\*[State size above 23 (with police expenditure below 77) - mean crime rate is 97.6]

\*[State size below 23 (with police expenditure below 77) - mean crime rate is 72.5]

\*[Youth count above 131 (with police expenditure above 77) - mean crime rate is 133]

\*[Youth count below 131 (with police expenditure above 77) - mean crime rate is 105]

## 6) Were variables excluded? Why?

Yes, some predictor variables were excluded. This finds the first variable that splits the data into 2 (the best split that give the most homogeneous groups). Then once the data is seperated, this is again done to each seperate group (the 2 groups that were split). This repeats where each subgroup and split until the groups have minimul datapoints (depends on the dataset size) or until no more splits can be made (to make the tree any more homogeneous). They basically make splits until the cp(comlexity parameter) cannot get better any more (this is when cp is very low). In our case, the best split was done using the expenditure of the year. Then depending on the first split, the next split was done based in state size or youth. After this no any more splits would be too small of a cp(not much improvement), so we do not want to add other variables which would make my tree less parsimonious.

## 7) Predict 10 year crime with ALL the predictor variables available in crime10

crime10<-read.csv("crimedata10.csv")  
  
#model predict crime   
Myp.rpart <- predict(wholecrime, crime10)   
Myp.rpart

## 1 2 3 4 5 6 7 8   
## 72.46667 72.46667 97.61818 72.46667 105.36364 97.61818 72.46667 72.46667   
## 9 10 11 12 13 14 15 16   
## 72.46667 72.46667 72.46667 133.04615 97.61818 133.04615 97.61818 72.46667   
## 17 18 19 20 21 22 23 24   
## 97.61818 105.36364 133.04615 97.61818 97.61818 97.61818 97.61818 97.61818   
## 25 26 27 28 29 30 31 32   
## 97.61818 133.04615 97.61818 72.46667 105.36364 133.04615 133.04615 105.36364   
## 33 34 35 36 37 38 39 40   
## 72.46667 133.04615 105.36364 97.61818 97.61818 133.04615 72.46667 133.04615   
## 41 42 43 44 45 46 47   
## 133.04615 133.04615 133.04615 133.04615 133.04615 133.04615 133.04615

## 8) correlation

correlation coefficient = 0.6175003

cor(Myp.rpart, crime10[["CrimeRate"]],method="pearson")

## [1] 0.6175003

## 9) Calculate MAE, is your model good at predicting?

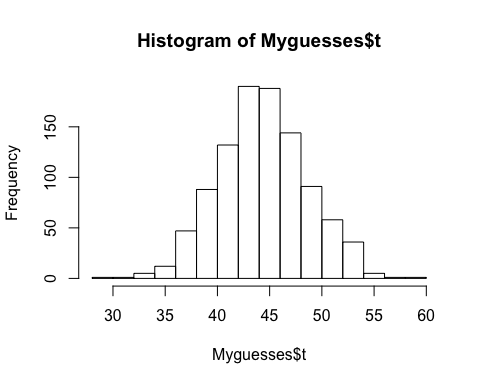
MAE = 24.32924 meaning: we are on average 24.32924 far off of predicting the actual thing. 24.32924 Does not seem that far off so I would say our model is good, but in the next question we will compare it with random guess and know exactly how good ours is.

MyMAE <- function(actual, predicted) {  
 mean(abs(actual - predicted))  
}  
  
MyMAE(predicted = Myp.rpart,actual = crime10[["CrimeRate"]])

## [1] 24.32924

## 10)Null distribution

Myactual = crime10[["CrimeRate"]]  
  
MyMAE2 <- function(data,indices) {  
 d<-data[indices]  
 return(mean(abs(Myactual - d)))  
}  
  
library(boot)  
Myguesses=boot(data=Myactual, statistic=MyMAE2, R=1000)  
  
hist(Myguesses$t)



## 11) random vs real.

Random MAE was 44.49198 (this changes every time) and the real MAE was 24.32924. My model did better than random chance it seems (as our model prediction is very far from the mean in the histogram). Next question will say if it was significantly better or not.

Mymean<- mean(Myguesses$t)  
Mymean

## [1] 44.3363

## 12) Is random and real significantly different?

They are significantly different, so my model was significantly better at predicting the crime10 data than at random.

Myp.value=length(which((Myguesses$t<24.32924)==T))/1000  
Myp.value

## [1] 0