Decision Tree Example: Statlog Heart

We will use statlog heart dataset again for decision tree model practice. Statlog (Heart) Data Set is downloaded from UCI machine learning repository. It has 13 different attributes and 1 class variable. The detailed feature attributes are presented in the below. You could also find it here: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

The attributes include: 1. Age; 2. Sex (female =0, male =1); 3. chestpain = chest pain type (4 values); 4. restbloodpr = resting blood pressure; 5. serumchol = serum cholestoral in mg/dl; 6. blsugar.120 = fasting blood sugar > 120 mg/dl (have =1, don’t have =0); 7. restelectro = resting electrocardiographic results (values 0,1,2); 8. maxheartrate = maximum heart rate achieved; 9. angina = exercise induced angina (don’t have =0, have =1); 10. oldpeak = ST depression induced by exercise relative to rest; 11. st.segment = the slope of the peak exercise ST segment; 12. majorvessels = number of major vessels (0-3) colored by flourosopy; 13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect;

The class variable is statlogheart: 1=Absence of heart disease; 2= presence of heart disease;

Among all attributes, we have real attributes: 1,4,5,8,10,12; ordered attributes: 11; binary attributes: 2,6,9 and nominal attributes: 3, 7, 13;

**Assignment 4:**

1. Import Boston Housing dataset from UCI machine learning repository: assignment4.csv

Code:

house <- read.csv(file.choose(), stringsAsFactors = F)

1. Use str() to inspect the type and data frame of variables. How many observations and how many variables in this dataset? Learn the meaning of each variable.

Ans:

#506 obs. of 14 variables

Code:

str(house)

1. Reform the CHAS to nominal (factor) variable. How many levels does CHAS have? How many instances belong to level 0?

Ans:

# There are 2 levels

# 471

Code:

house$CHAS <- factor(house$CHAS)

summary(house$CHAS )

1. What is the mean value of age variable? Since age is the proportion of homes built before 1940, how do you interpret this mean value? What is the maximum value of MEDV? Build a histogram for MEDV, which value range has the highest frequency?

Ans:

# 68.5749 , the mean value decreases

#50

# 20 - 25

Code:

Mea n(house$AGE)

max(house$MEDV)

hist(house$MEDV, main="Plot of House MEDV", ylab="house MEDV")

1. Check the correlation between all the variables except for CHAS. Which pairs have correlations larger than 0.6? What variables have correlations with MEDV larger than 0.6?

Ans:

#CRIM-CRIM; CRIM-RAD; ZN-ZN; ZN-DIS; INDUS-INDUS; INDUS-NOX; INDUS-AGE; INDUS-TAX; INDUS-LSTAT;NOX-INDUS; NOX-NOX; NOX-AGE; NOX-RAD; NOX-TAX; RM-RM; RM-MEDV; AGE-INDUS; AGE-NOX; AGE-AGE;AGE-LSTAT; DIS-ZN; DIS-DIS; RAD-CRIM; RAD-NOX; RAD-RAD; RAD-TAX;TAX-INDUS; TAX-NOX; TAX-RAD; TAX-TAX;PTRATIO-PTRATIO; B-B; LSTAT-INDUS; LSTAT-AGE; LSTAT-LSTAT;

# MEDV larger than 0.6

MEDV-RM; MEDV-MEDV;

Code:

house1<-cor(house[-4])

house1

1. What are your assumptions of the relations between crime and home value, between DIS and home value? Does higher pupil-teacher ratio lead to lower home value? Evaluate your assumptions with the results from correlation.

Ans:

# -0.3883046 # 0.2499287

#yes, higher pupil-teacher ratio lead to lower home value

Code:

cor(house$CRIM, house$MEDV)

cor(house$DIS, house$MEDV)

cor(house$PTRATIO, house$MEDV)

1. Use pairs.panels() to visualize the relations between “RM”, “LSTAT” and “MEDV”. What are the histograms depicting on the diagonal? Does more room lead to higher home values? Does higher percentage of lower status of the population cause higher home values?

Ans:

#the histogram on diagonal represents the plots for MEDV , LSTAT ,RM

#yes,more room lead to higher home values

#No,higher percentage of lower status of the population does not cause higher home values

Code:

install.packages("psych")

library(psych)

pairs.panels(house[c("MEDV","LSTAT","RM")])

1. **Randomize** the dataset and use the first 80% of observations for training. Set seed as 1234.

Code:

nrows <- nrow(house)

nrows

train.size<- floor(0.8 \* nrows)

train.size

set.seed(1234)

train.index <- sample(1:nrows,train.size,replace = F)

house.train<- house[train.index,] # 404 training observations

house.test<- house[-train.index,] # 102 testing observations

str(house.train)

str(house.test)

1. Build a linear regression model with MEDV as dependent variable and the rest variables as independent variables. Which variables have insignificant impact on MEDV?

Ans: # INDUS; AGE

Code:

model <- lm(MEDV ~., data=house.train)

summary(model)

1. How to interpret the estimate for variable RM? What is the p-value? Is this variable statistically significant? Why?

Ans:

#increase in 1 unit of RM will increase the MEDV value by 3.810 units.

#p-value: 2e-16

#yes it is significant as 2e-16 is less than 0.001

1. What are the R-squared value and adjusted R-squared value?

Ans:

#R-squared value is 0.7589

#Adjusted R-squared value is 0.7509

1. Try Crim^2 as one of the dependent variable and build a liner regression model include this variable (all other variables included). Does this Crim^2 have significant impact on MEDV?

Ans:

#Crim^2 does not have significant impact on MEDV.

Code:

house.train$crim2 <- house.train$CRIM^2

model.crim2 <- lm(MEDV ~.,data=house.train)

summary(model.crim2)

1. Also create a new tax2 variable with tax2 = 1 if TAX >=350 and tax2 = 0 if TAX < 350. Build a linear regression model include this tax2 (all other variables included). What is the adjusted R-squared value?

Ans:

#Adjusted R-squared: 0.7502

Code:

house.train$tax2 <- ifelse(house.train$TAX >=350, 1, 0)

model.tax2 <- lm(MEDV ~ ., house.train)

summary(model.tax2) Create a new simple regression model and estimate MEDV using only CHAS and CRIM.

1. How to interpret the estimate for variable CHAS1? What is the p-value? Is this variable statistically significant? Why?

Ans:

#the value of MEDV for CHAS1 is 5.57772 more than CHAS0

# CHAS1 p-value is 2.2e-16

# Yes, it is statistically significant as p-value 2.2e-16

1. Calculate MAE using the function as discussed in class using training dataset. Also, calculate the correlation between predicted and the actual using training dataset

Ans:

# 3.163262

# 4.531315

Code:

model <- lm(MEDV ~., data=house.train)

summary(model)

predicted <- predict(model, house.train)

MAE <- function (v1, v2) mean(abs(v1-v2))

MAE(house.train$MEDV, predicted)

RMSE <- function (v1, v2) (mean((v1 -v2)^2))^0.5

RMSE(house.train$MEDV, predicted)

1. Calculate MAE and correlation using testing dataset.

Ans:

#3.44646

#0.8626647

Code:

model.test <- lm(MEDV ~., house.test)

summary(model.test)

predicted.test <- predict(model.test, house.test)

MAE <- function (v1, v2) mean(abs(v1-v2))

MAE(house.test$MEDV, predicted.test)

cor(predicted.test, house.test$MEDV)

1. Estimate MEDV via regression tree using CHAS and CRIM.
   1. Plot the tree using rpart.plot

Code:

library("rpart")

install.packages("rpart.plot")

library("rpart.plot")

mytree <- rpart(house$MEDV ~ house$CHAS + house$CRIM, data = house.train)

mytree

rpart.plot(mytree)

* 1. What is the most important feature? What is the rule for the first leaf node?

Ans:

# CRIM

# decision taken after computing all attributes , the first leaf node represent the class label.

* 1. How many examples/records/data points are there in the first leaf node? What is the predicted value for these data points?

Ans:

# 5% records are in the first leaf node

# 10 is the predicted value for this data point

* 1. Calculate the correlation & MAE between predicted and the actual using training dataset

Ans:

#6.074679

# 0.446937

Code:

model.train <- lm(MEDV ~CHAS+CRIM, house.train)

summary(model.train)

predicted.train <- predict(model.train, house.train)

MAE <- function (v1, v2) mean(abs(v1-v2))

MAE(house.train$MEDV, predicted.train)

cor(predicted.train, house.train$MEDV)

* 1. Calculate the correlation & MAE between predicted and the actual using testing dataset

Ans:

#6.135301

# 0.2675173

Code:

model.test <- lm(MEDV ~CHAS+CRIM, house.test)

summary(model.test)

predicted.test <- predict(model.test, house.test)

MAE <- function (v1, v2) mean(abs(v1-v2))

MAE(house.test$MEDV, predicted.test)

cor(predicted.test, house.test$MEDV)