Module 3 CT Option 2

Dan Amrine

August 5, 2019

# Loading the Data

# Read the .csv into a data frame

Cars<-as.data.frame(read.csv(“ToyotaCorolla.csv”))

# A Quick Exploration of the data structure

colnames(Cars) str(Cars)

## 

## More Exploration using Max, Min, and Mean

# Comparing the Max, min and mean of price and the age of the cars

Price\_mean<-mean(Cars$

Price)

Price\_Min<-min(Cars$Price)

Age\_mean<-mean(Cars

Age\_08\_04)

Age\_Min<-min(Cars$Age\_08\_04)

### 

### The Regression Tree

# Loading the necessary package for Regressions trees

library(rpart)

# A simple Regression tree

# with a single node limited to 30 splits,

# and cp=0.001

fit<-rpart(Price ~ Age\_08\_04+KM+Fuel\_Type+HP+Automatic+Doors+Quarterly\_Tax+Mfr\_Guarantee+ Guarantee\_Period+Airco+Automatic\_airco+CD\_Player+Powered\_Windows+Sport\_Model +Tow\_Bar,data = Cars, method = “anova”, control = rpart.control(minsplit=30L, cp=0.001))

# plotting the tree and adding text

plotcp(fit)

plot(fit, uniform=TRUE, main=“Regression Tree for Price”)

text(fit, use.n=TRUE, all=TRUE, cex=.8)

### What I learned

We were tasked to load the *ToyotaCarolla.csv* into R. We then had to explore the data and then generate a regression tree to predict price or at least understand which predictor influenced the price the most.

The initial part was like the original critical thinking assignments. I chose to look at the column names and structure because these two functions break the down data and helps us understand just how large it is. Structure breaks it down by type and you can use this to target which columns to use in your analysis based on if it deals with factor or integer types.

Looking at the Price and Age of the cars helped me to understand the scope of the focus. I was surprised by how old the cars were give then mean of 55.

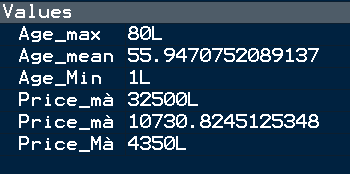


Figure 1: The Min, Max, and Mean of the Age and price of the cars.

Based on the Regression Tree generated using rpart the most important variables are the Age, Automatic\_airco, KM, HP, and Quarterly\_Tax.



Figure 2: The “Variable.Importance” Table.

The most difficult part of this was getting the formula to work and then making sure there were no capital letters in the arguments, other than capital letters in the variables. I also found it difficult to understand the results. I did a basic plot of the tree to see how the data was divided, but the real results are located in the tables generated by rpart. The table in (Fig. 2) outlines it well and ranks the variables but I’m curious as to why the numbers are so large.