Module 3 CT Option 2

Shaun Stearns

8/2/2019

# Load Required Packages

library(rpart)  
library(rpart.plot)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(adabag)

## Loading required package: foreach

## Loading required package: doParallel

## Loading required package: iterators

## Loading required package: parallel

library(jtools)  
library(sandwich)

# Set Working Directory and read csv file

Toyota.ct <- read.csv("ToyotaCorolla.csv")

# Dimension of the frame

dim(Toyota.ct)

## [1] 1436 39

# First six rows

head(Toyota.ct)

## Id Model Price Age\_08\_04  
## 1 1 TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13500 23  
## 2 2 TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13750 23  
## 3 3 \xa0TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13950 24  
## 4 4 TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 14950 26  
## 5 5 TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors 13750 30  
## 6 6 TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors 12950 32  
## Mfg\_Month Mfg\_Year KM Fuel\_Type HP Met\_Color Color Automatic CC  
## 1 10 2002 46986 Diesel 90 1 Blue 0 2000  
## 2 10 2002 72937 Diesel 90 1 Silver 0 2000  
## 3 9 2002 41711 Diesel 90 1 Blue 0 2000  
## 4 7 2002 48000 Diesel 90 0 Black 0 2000  
## 5 3 2002 38500 Diesel 90 0 Black 0 2000  
## 6 1 2002 61000 Diesel 90 0 White 0 2000  
## Doors Cylinders Gears Quarterly\_Tax Weight Mfr\_Guarantee BOVAG\_Guarantee  
## 1 3 4 5 210 1165 0 1  
## 2 3 4 5 210 1165 0 1  
## 3 3 4 5 210 1165 1 1  
## 4 3 4 5 210 1165 1 1  
## 5 3 4 5 210 1170 1 1  
## 6 3 4 5 210 1170 0 1  
## Guarantee\_Period ABS Airbag\_1 Airbag\_2 Airco Automatic\_airco  
## 1 3 1 1 1 0 0  
## 2 3 1 1 1 1 0  
## 3 3 1 1 1 0 0  
## 4 3 1 1 1 0 0  
## 5 3 1 1 1 1 0  
## 6 3 1 1 1 1 0  
## Boardcomputer CD\_Player Central\_Lock Powered\_Windows Power\_Steering  
## 1 1 0 1 1 1  
## 2 1 1 1 0 1  
## 3 1 0 0 0 1  
## 4 1 0 0 0 1  
## 5 1 0 1 1 1  
## 6 1 0 1 1 1  
## Radio Mistlamps Sport\_Model Backseat\_Divider Metallic\_Rim Radio\_cassette  
## 1 0 0 0 1 0 0  
## 2 0 0 0 1 0 0  
## 3 0 0 0 1 0 0  
## 4 0 0 0 1 0 0  
## 5 0 1 0 1 0 0  
## 6 0 1 0 1 0 0  
## Parking\_Assistant Tow\_Bar  
## 1 0 0  
## 2 0 0  
## 3 0 0  
## 4 0 0  
## 5 0 0  
## 6 0 0

# Print a list of variables as a refernce guide

t(t(names(Toyota.ct)))

## [,1]   
## [1,] "Id"   
## [2,] "Model"   
## [3,] "Price"   
## [4,] "Age\_08\_04"   
## [5,] "Mfg\_Month"   
## [6,] "Mfg\_Year"   
## [7,] "KM"   
## [8,] "Fuel\_Type"   
## [9,] "HP"   
## [10,] "Met\_Color"   
## [11,] "Color"   
## [12,] "Automatic"   
## [13,] "CC"   
## [14,] "Doors"   
## [15,] "Cylinders"   
## [16,] "Gears"   
## [17,] "Quarterly\_Tax"   
## [18,] "Weight"   
## [19,] "Mfr\_Guarantee"   
## [20,] "BOVAG\_Guarantee"   
## [21,] "Guarantee\_Period"   
## [22,] "ABS"   
## [23,] "Airbag\_1"   
## [24,] "Airbag\_2"   
## [25,] "Airco"   
## [26,] "Automatic\_airco"   
## [27,] "Boardcomputer"   
## [28,] "CD\_Player"   
## [29,] "Central\_Lock"   
## [30,] "Powered\_Windows"   
## [31,] "Power\_Steering"   
## [32,] "Radio"   
## [33,] "Mistlamps"   
## [34,] "Sport\_Model"   
## [35,] "Backseat\_Divider"   
## [36,] "Metallic\_Rim"   
## [37,] "Radio\_cassette"   
## [38,] "Parking\_Assistant"  
## [39,] "Tow\_Bar"

# Subset of predictors requested by the shareholder

Toyota.ct <- Toyota.ct[c(3,4,7,8,9,12,14,17,19,21,25,26,28,30,34,39)]

# Summary Statistics for each varaible

summary(Toyota.ct)

## Price Age\_08\_04 KM Fuel\_Type   
## Min. : 4350 Min. : 1.00 Min. : 1 CNG : 17   
## 1st Qu.: 8450 1st Qu.:44.00 1st Qu.: 43000 Diesel: 155   
## Median : 9900 Median :61.00 Median : 63390 Petrol:1264   
## Mean :10731 Mean :55.95 Mean : 68533   
## 3rd Qu.:11950 3rd Qu.:70.00 3rd Qu.: 87021   
## Max. :32500 Max. :80.00 Max. :243000   
## HP Automatic Doors Quarterly\_Tax   
## Min. : 69.0 Min. :0.00000 Min. :2.000 Min. : 19.00   
## 1st Qu.: 90.0 1st Qu.:0.00000 1st Qu.:3.000 1st Qu.: 69.00   
## Median :110.0 Median :0.00000 Median :4.000 Median : 85.00   
## Mean :101.5 Mean :0.05571 Mean :4.033 Mean : 87.12   
## 3rd Qu.:110.0 3rd Qu.:0.00000 3rd Qu.:5.000 3rd Qu.: 85.00   
## Max. :192.0 Max. :1.00000 Max. :5.000 Max. :283.00   
## Mfr\_Guarantee Guarantee\_Period Airco Automatic\_airco   
## Min. :0.0000 Min. : 3.000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.: 3.000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :0.0000 Median : 3.000 Median :1.0000 Median :0.00000   
## Mean :0.4095 Mean : 3.815 Mean :0.5084 Mean :0.05641   
## 3rd Qu.:1.0000 3rd Qu.: 3.000 3rd Qu.:1.0000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :36.000 Max. :1.0000 Max. :1.00000   
## CD\_Player Powered\_Windows Sport\_Model Tow\_Bar   
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :1.000 Median :0.0000 Median :0.0000   
## Mean :0.2187 Mean :0.562 Mean :0.3001 Mean :0.2779   
## 3rd Qu.:0.0000 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.000 Max. :1.0000 Max. :1.0000

# Check for Missing values

### No missing values were found

data.frame(miss.val=sapply(Toyota.ct, function(x)   
 sum(length(which(is.na(x))))))

## miss.val  
## Price 0  
## Age\_08\_04 0  
## KM 0  
## Fuel\_Type 0  
## HP 0  
## Automatic 0  
## Doors 0  
## Quarterly\_Tax 0  
## Mfr\_Guarantee 0  
## Guarantee\_Period 0  
## Airco 0  
## Automatic\_airco 0  
## CD\_Player 0  
## Powered\_Windows 0  
## Sport\_Model 0  
## Tow\_Bar 0

# Multiple Regression

### Run to see if any variables can be eliminated, partial and part coefficients added to understand how much variability is accounted for by each variable. Jtools was used to provide a more comrpehensive regression analysis.

ToyReg <- lm(Price~ ., data = Toyota.ct)  
summ(ToyReg, scale = TRUE, part.corr = TRUE, digits = 3)

## MODEL INFO:  
## Observations: 1436  
## Dependent Variable: Price  
## Type: OLS linear regression   
##   
## MODEL FIT:  
## F(16,1419) = 712.506, p = 0.000  
## R² = 0.889  
## Adj. R² = 0.888   
##   
## Standard errors: OLS  
## --------------------------------------------------------------------------  
## Est. S.E. t val. p partial.r  
## ---------------------- ----------- --------- --------- ------- -----------  
## (Intercept) 8159.668 332.922 24.509 0.000   
## Age\_08\_04 -2046.912 51.220 -39.964 0.000 -0.728  
## KM -674.782 45.694 -14.767 0.000 -0.365  
## Fuel\_TypeDiesel 1890.588 325.476 5.809 0.000 0.152  
## Fuel\_TypePetrol 1863.418 345.302 5.396 0.000 0.142  
## HP 511.209 42.827 11.937 0.000 0.302  
## Automatic 551.370 141.424 3.899 0.000 0.103  
## Doors 155.542 33.623 4.626 0.000 0.122  
## Quarterly\_Tax 680.754 65.113 10.455 0.000 0.267  
## Mfr\_Guarantee 198.909 68.698 2.895 0.004 0.077  
## Guarantee\_Period 245.414 34.648 7.083 0.000 0.185  
## Airco 154.206 82.335 1.873 0.061 0.050  
## Automatic\_airco 3093.551 164.400 18.817 0.000 0.447  
## CD\_Player 301.242 91.358 3.297 0.001 0.087  
## Powered\_Windows 452.320 78.609 5.754 0.000 0.151  
## Sport\_Model 365.990 74.988 4.881 0.000 0.128  
## Tow\_Bar -245.090 74.369 -3.296 0.001 -0.087  
## --------------------------------------------------------------------------  
##   
## -------------------------------  
## part.r  
## ---------------------- --------  
## (Intercept)   
## Age\_08\_04 -0.353  
## KM -0.130  
## Fuel\_TypeDiesel 0.051  
## Fuel\_TypePetrol 0.048  
## HP 0.105  
## Automatic 0.034  
## Doors 0.041  
## Quarterly\_Tax 0.092  
## Mfr\_Guarantee 0.026  
## Guarantee\_Period 0.063  
## Airco 0.017  
## Automatic\_airco 0.166  
## CD\_Player 0.029  
## Powered\_Windows 0.051  
## Sport\_Model 0.043  
## Tow\_Bar -0.029  
## -------------------------------  
##   
## Continuous predictors are mean-centered and scaled by 1 s.d.

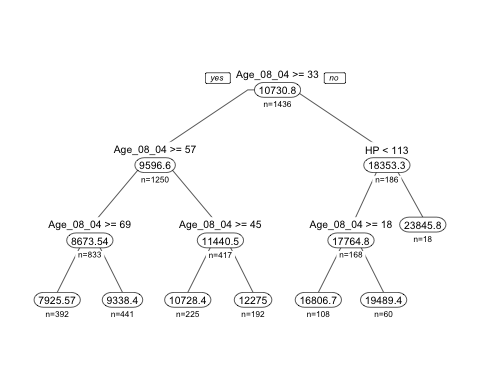
# Default Regression Tree

### Provides a preview of the most relevant variables predicting Price.

ToyReg.tree <- rpart(Price ~ ., data = Toyota.ct, method = "anova")  
length(ToyReg.tree$frame$var[ToyReg.tree$frame$var == "<leaf>"])

## [1] 7

prp(ToyReg.tree, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10, digits=-6)



# Deep Regression Tree

### At 121 leafs, I chose not to produce this tree due to technical limitations.

ToyReg.tree <- rpart(Price ~ ., data = Toyota.ct, method = "anova", cp = 0, maxdepth = 30)  
length(ToyReg.tree$frame$var[ToyReg.tree$frame$var == "<leaf>"])

## [1] 121

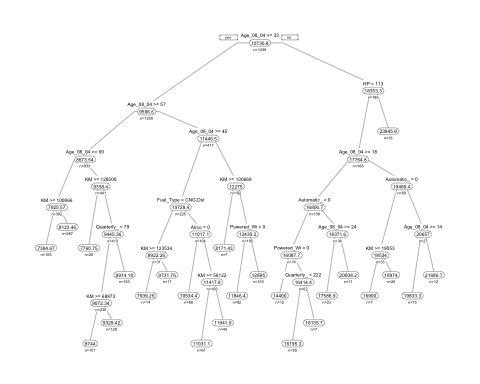
# Specified Regression Tree

### This tree specifically meets the parameters requested by the share holder.

ToyReg.tree <- rpart(Price ~ ., data = Toyota.ct, method = "anova", cp = 0.001, maxdepth = 30)  
length(ToyReg.tree$frame$var[ToyReg.tree$frame$var == "<leaf>"])

## [1] 24

prp(ToyReg.tree, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10, digits=-6)



# Pruned Regression Tree: Best pruned regression tree.

### It’s interesting that there are more leaves produced than the code requested by the shareholder.

Toy.ct <- rpart(Price ~ ., data = Toyota.ct, method = "anova",  
 cp = 0.001, minsplit = 5, xval = 5)  
length(ToyReg.tree$frame$var[ToyReg.tree$frame$var == "<leaf>"])

## [1] 24

printcp(Toy.ct)

##   
## Regression tree:  
## rpart(formula = Price ~ ., data = Toyota.ct, method = "anova",   
## cp = 0.001, minsplit = 5, xval = 5)  
##   
## Variables actually used in tree construction:  
## [1] Age\_08\_04 Airco Automatic\_airco Fuel\_Type   
## [5] HP KM Powered\_Windows Quarterly\_Tax   
##   
## Root node error: 1.8877e+10/1436 = 13145711  
##   
## n= 1436   
##   
## CP nsplit rel error xerror xstd  
## 1 0.6576727 0 1.000000 1.00035 0.0631131  
## 2 0.1127049 1 0.342327 0.35436 0.0218508  
## 3 0.0318482 2 0.229622 0.23977 0.0200888  
## 4 0.0219443 3 0.197774 0.22722 0.0166121  
## 5 0.0147050 4 0.175830 0.18889 0.0127083  
## 6 0.0131259 5 0.161125 0.18411 0.0126605  
## 7 0.0114420 6 0.147999 0.17771 0.0122842  
## 8 0.0064805 7 0.136557 0.15291 0.0101421  
## 9 0.0064372 8 0.130077 0.14893 0.0096332  
## 10 0.0062134 9 0.123639 0.14893 0.0096332  
## 11 0.0039421 10 0.117426 0.14260 0.0092003  
## 12 0.0038260 11 0.113484 0.13829 0.0090444  
## 13 0.0035456 12 0.109658 0.13720 0.0089757  
## 14 0.0026595 13 0.106112 0.13207 0.0087742  
## 15 0.0024398 14 0.103453 0.13069 0.0085844  
## 16 0.0023072 15 0.101013 0.12947 0.0085073  
## 17 0.0022228 16 0.098706 0.12937 0.0085043  
## 18 0.0021611 17 0.096483 0.13022 0.0086093  
## 19 0.0019877 18 0.094322 0.12988 0.0086022  
## 20 0.0019245 19 0.092334 0.12988 0.0086022  
## 21 0.0016246 20 0.090410 0.12664 0.0086822  
## 22 0.0012936 21 0.088785 0.12392 0.0086463  
## 23 0.0012567 22 0.087491 0.12345 0.0086584  
## 24 0.0012130 23 0.086235 0.12368 0.0086950  
## 25 0.0011379 24 0.085022 0.12165 0.0085252  
## 26 0.0010992 25 0.083884 0.12148 0.0085327  
## 27 0.0010513 26 0.082785 0.12144 0.0085364  
## 28 0.0010284 28 0.080682 0.11975 0.0083705  
## 29 0.0010000 29 0.079654 0.11970 0.0083724

pruned.ct <- prune(Toy.ct,  
 cp = Toy.ct$cptable[which.min(Toy.ct$cptable[,"xerror"]), "CP"])  
length(pruned.ct$frame$var[pruned.ct$frame$var == "<leaf>"])

## [1] 30

prp(pruned.ct, type = 1, extra = 1, split.font = 1, varlen = -10, digits=-6)

