Course 2 Section 4.13 - YOUR TURN

Jiaying Wu

17/10/2020

What you need to do

In a previous exercise, regression trees were used to predict housing auction prices in Melbourne, which is a continuous variable. In this exercise, your task is to determine the location of the house.

Since the response variable is a categorical variable, classification trees are the best model for you to use.

About the data

A subset of the data, that includes two suburbs (Carlton and Brighton) and five variables is used for this exercise. The popular split criteria for classification trees are Gini and Entropy.

find out how the model responds to the two criteria.

```
# load library
library(tidyverse)
library(rpart)
library(rpart.plot)

# load data
houses_raw <- read_csv("https://raw.githubusercontent.com/datascienceprogram/ids_course_data/master/Mel

# subset data
houses_suburb2 <- houses_raw %>%
  select(Suburb, Price, Landsize, Rooms, Type) %>%
  filter(Suburb %in% c("Carlton", "Brighton"))

houses_suburb2
```

```
## # A tibble: 554 x 5
##    Suburb     Price Landsize Rooms Type
##    <chr>      <dbl>    <dbl> <dbl> <chr>
##  1 Brighton 1550000      663     3 h
##  2 Brighton      NA      683     4 h
##  3 Brighton 1635000      366     3 h
##  4 Brighton      NA      688     3 h
##  5 Brighton      NA      318     3 h
##  6 Brighton 1830000      436     4 h
##  7 Brighton 1300000       NA     3 t
##  8 Brighton 3695000      836     4 h
##  9 Brighton      NA      845     5 h
## 10 Brighton      NA        0     4 h
## # ... with 544 more rows
```

use Gini to split the tree.

```r
rp_fit_gini <- rpart(Suburb ~ ., data = houses_suburb2, parms = list(split = "gini"))
rp_fit_gini
```

```
## n= 554
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 554 98 Brighton (0.82310469 0.17689531)
##    2) Landsize>=160.5 370 21 Brighton (0.94324324 0.05675676)
##      4) Landsize< 1319.5 357 14 Brighton (0.96078431 0.03921569) *
##      5) Landsize>=1319.5 13  6 Carlton (0.46153846 0.53846154) *
##    3) Landsize< 160.5 184 77 Brighton (0.58152174 0.41847826)
##      6) Type=t,u 137 40 Brighton (0.70802920 0.29197080)
##       12) Price>=614750 106 24 Brighton (0.77358491 0.22641509) *
##       13) Price< 614750 31 15 Carlton (0.48387097 0.51612903)
##         26) Price>=492500 19  9 Brighton (0.52631579 0.47368421) *
##         27) Price< 492500 12  5 Carlton (0.41666667 0.58333333) *
##      7) Type=h 47 10 Carlton (0.21276596 0.78723404)
##       14) Landsize< 50.5 13  4 Brighton (0.69230769 0.30769231) *
##       15) Landsize>=50.5 34  1 Carlton (0.02941176 0.97058824) *
```

Identify the number of terminal nodes

Q1.Based on the print output, how many terminal nodes does the model produce? What's the first split for the tree?

```r
printcp(rp_fit_gini)
```

```
##
## Classification tree:
## rpart(formula = Suburb ~ ., data = houses_suburb2, parms = list(split = "gini"))
##
## Variables actually used in tree construction:
## [1] Landsize Price    Type
##
## Root node error: 98/554 = 0.1769
##
## n= 554
##
##         CP nsplit rel error  xerror     xstd
## 1 0.137755      0   1.00000 1.00000 0.091646
## 2 0.051020      2   0.72449 0.72449 0.080283
## 3 0.010204      3   0.67347 0.77551 0.082630
## 4 0.010000      6   0.64286 0.82653 0.084858
```

There are seven terminal nodes, the first split is if the landsize larger than or equal to 160.5.

Identify the number of splits

The previous print output displays the 'CP' table for the model fit, which contains information about the model's goodness of fit.

Q2.How many splits are performed during the model fitting?

There are six splits (nsplit in bottom row).

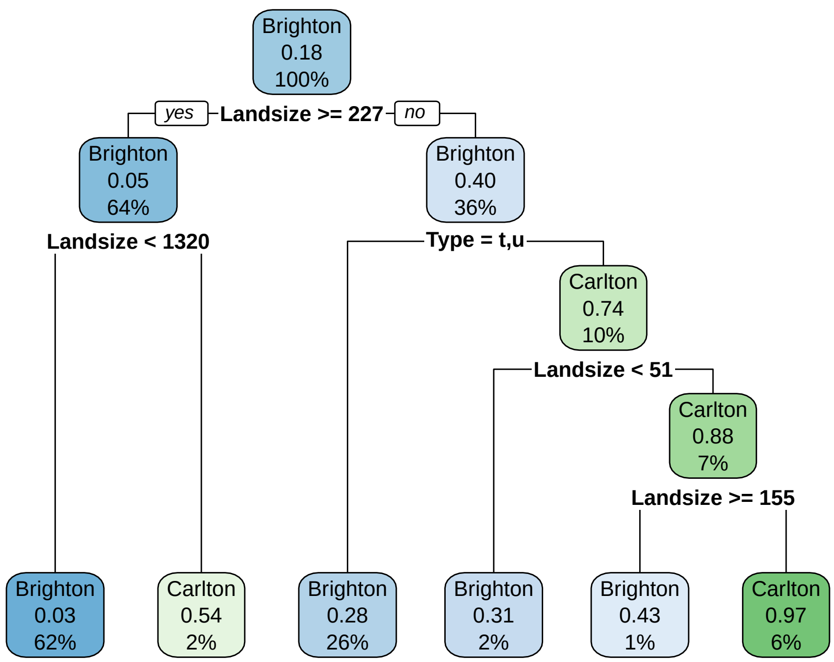
Q3.After two splits, what is R^2? Interpret that R^2.

1 - 0.72449 = 0.27551

The model explained 27.551% of the variance after two splits.

Entropy criteria

```
rp_fit_entropy <- rpart(Suburb ~ ., data = houses_suburb2, parms = list(split = "information"))
rpart.plot(rp_fit_entropy)
```



Compute the confusion table

```
pred_gini <- predict(rp_fit_gini, houses_suburb2, type = "class")
pred_entropy <- predict(rp_fit_entropy, houses_suburb2, type = "class")
table(houses_suburb2$Suburb, pred_gini)
```

```
##           pred_gini
##            Brighton Carlton
##   Brighton      444      12
##   Carlton        51      47
```

```r
table(houses_suburb2$Suburb, pred_entropy)
```

```
##           pred_entropy
##            Brighton Carlton
##   Brighton      449       7
##   Carlton        59      39
```

Q3.How many cases have been correctly classified for both models?

Gini criteria:

444+47 = 491

Entropy criteria:

449+39 = 488

Q4.Which criteria gives a better model?

Gini criteria might gives a better model.