ETC 2420/5242 Lab 8 2017

SOLUTION

Week 8

Purpose

For this lab we are going to build models based on partitioning, and combine models built on bootstrap samples, using regression trees and forests.

Reading

Read the code in the lecture notes on regression trees and forests from weeks 7 and 8. We will work with data scraped from property auction reports, collected over the last couple of years. Dr Julia Polak collected the reports, and together we used the `pdftools` package in R to extract information about each property. We will compare the results from trees and forests with the multiple regression model.

Warmup

This is a description of the variables:

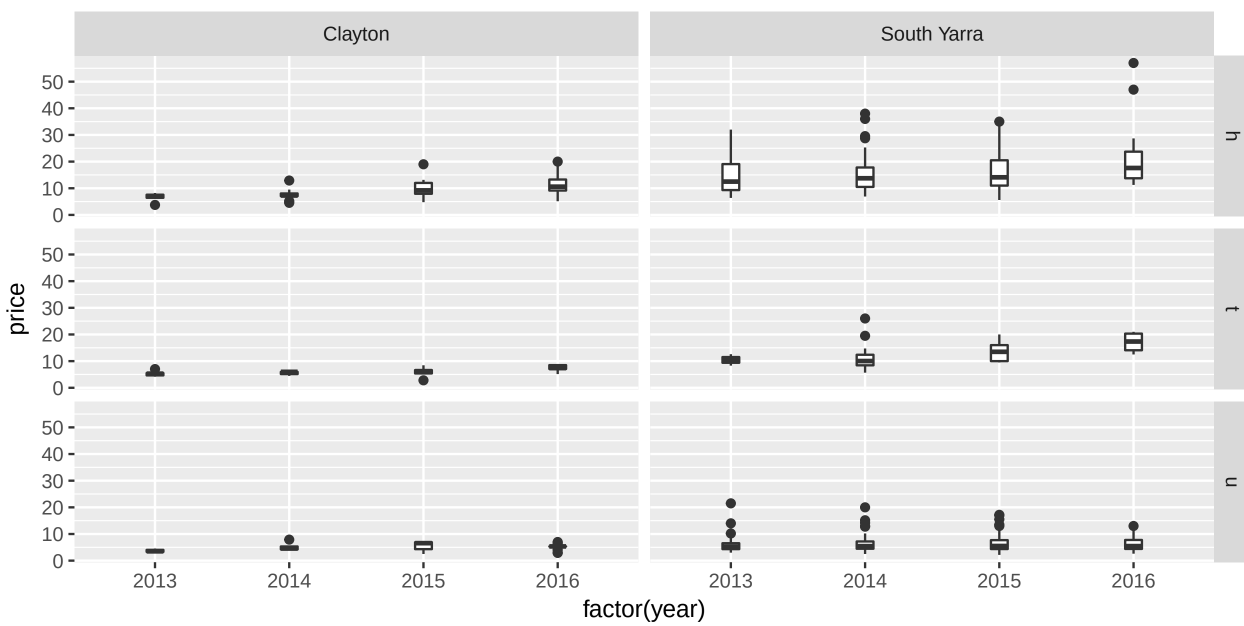
Variable	Description
id	unique id for property
suburb	suburb location of property
price	Price house sold for in AUD dollars, divided by 100,000
result	S indicates property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; o res - other residential; w - withdrawn prior to auction
agent	realtor in charge of sale
nbeds	Number of bedrooms
property type	h =house, t =townhouse, u =unit/apartment
day	day of the month of auction
month	month of auction
year	year of auction
nvisits	How many people came to open houses
ncars	Number of parking places
nbaths	Number of bathrooms
land size	Size of the lot, in sq m, units will be 0
house size	Internal size of property in sq m

We have subsetted the data to only use two suburbs, Clayton and South Yarra.

Take a quick glimpse of the data, by making some numerical and visual summaries. What is the average sale price for Clayton and South Yarra, over this period? Is there an increase in price over the four years?

```
#        id           suburb              price            result         
#  Min.   :  115   Length:614         Min.   : 2.165   Length:614        
#  1st Qu.:41830   Class :character   1st Qu.: 4.900   Class :character  
#  Median :54787   Mode  :character   Median : 6.810   Mode  :character  
#  Mean   :52464                      Mean   : 8.695                     
#  3rd Qu.:64814                      3rd Qu.:10.182                     
```

```
#  Max.   :75347                      Max.   :57.000
#      nbeds      property_type           day          month
#  Min.   :1.00   Length:614         Min.   : 1.00   Length:614
#  1st Qu.:2.00   Class :character   1st Qu.: 9.00   Class :character
#  Median :2.00   Mode  :character   Median :16.00   Mode  :character
#  Mean   :2.27                      Mean   :16.13
#  3rd Qu.:3.00                      3rd Qu.:23.00
#  Max.   :8.00                      Max.   :31.00
#       year         nvisits          rating           ncars
#  Min.   :2013   Min.   :  7.00   Min.   : 0.000   Min.   :0.000
#  1st Qu.:2014   1st Qu.: 52.00   1st Qu.: 3.000   1st Qu.:0.000
#  Median :2015   Median : 92.00   Median : 5.000   Median :0.000
#  Mean   :2015   Mean   : 93.65   Mean   : 4.971   Mean   :0.614
#  3rd Qu.:2015   3rd Qu.:137.00   3rd Qu.: 7.000   3rd Qu.:2.000
#  Max.   :2016   Max.   :181.00   Max.   :10.000   Max.   :3.000
#      nbaths        land_size        house_size
#  Min.   :1.000   Min.   :   0.0   Min.   : 70.46
#  1st Qu.:1.000   1st Qu.:   0.0   1st Qu.: 74.70
#  Median :1.500   Median :   0.0   Median :152.15
#  Mean   :1.619   Mean   : 238.4   Mean   :173.72
#  3rd Qu.:2.000   3rd Qu.: 479.5   3rd Qu.:241.93
#  Max.   :3.000   Max.   :1115.0   Max.   :496.00
```



Model building will be done using:

- Response: `price`
- Explanatory variables: `suburb, result, nbeds and property type`.

Subset the data to contain just these variables.

Now to correctly evaluate a tree model, you should fit the model to half of the data, and calculate the error on the predictions of the other half. We are going to make the split equally for the two suburbs, so that both are reqpresented

To compare models we will compute the mean square error (MSE):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

Write a function to compute the MSE.

Question 1

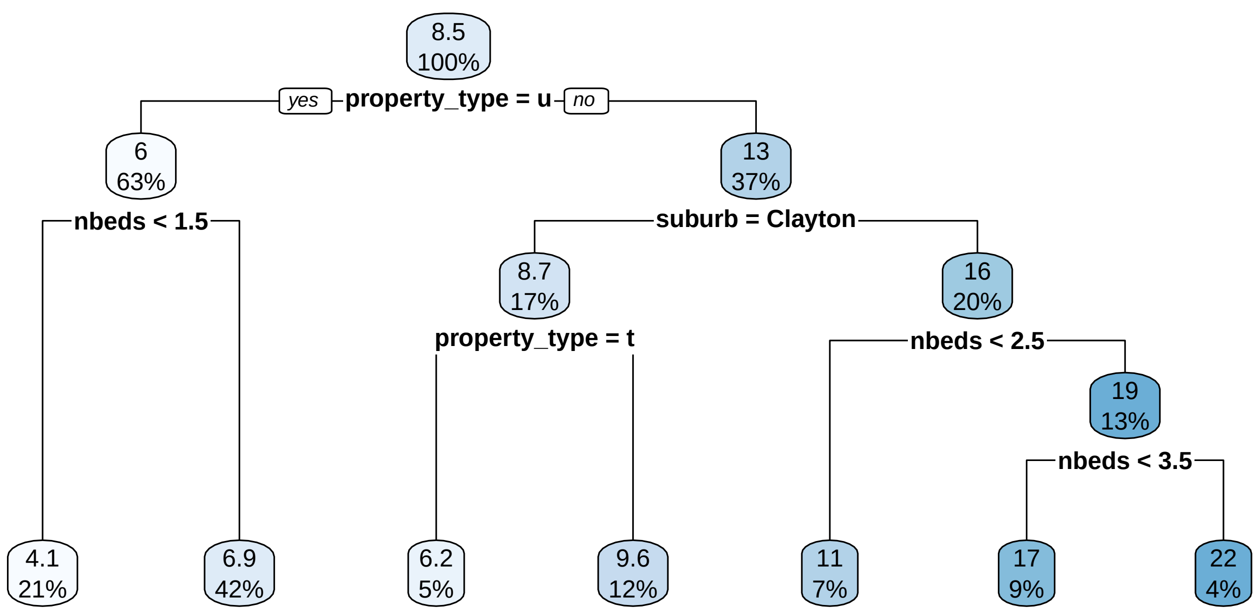
a. (2pts) Fit a regression tree, to the training data, with the default parameters to the data.

```
# n= 307
#
# node), split, n, deviance, yval
#       * denotes terminal node
#
#  1) root 307 10383.58000  8.490146
#    2) property_type=u 193  1290.10000  5.983482
#      4) nbeds< 1.5 65    89.34170  4.128723 *
#      5) nbeds>=1.5 128   863.59890  6.925352 *
#    3) property_type=h,t 114  5827.72500 12.733880
#      6) suburb=Clayton 52   544.70960  8.692594
#       12) property_type=t 14    20.97634  6.218071 *
#       13) property_type=h 38   406.42450  9.604260 *
#      7) suburb=South Yarra 62  3721.46200 16.123350
#       14) nbeds< 2.5 22   323.22850 11.187180 *
#       15) nbeds>=2.5 40  2567.36000 18.838250
#         30) nbeds< 3.5 27  1062.48200 17.172960 *
#         31) nbeds>=3.5 13  1274.49000 22.296920 *
```

b. (4pts) Plot the tree. How many terminal nodes? What variables are used?

```
7 terminal nodes
```

```
property type, nbeds, suburb
```

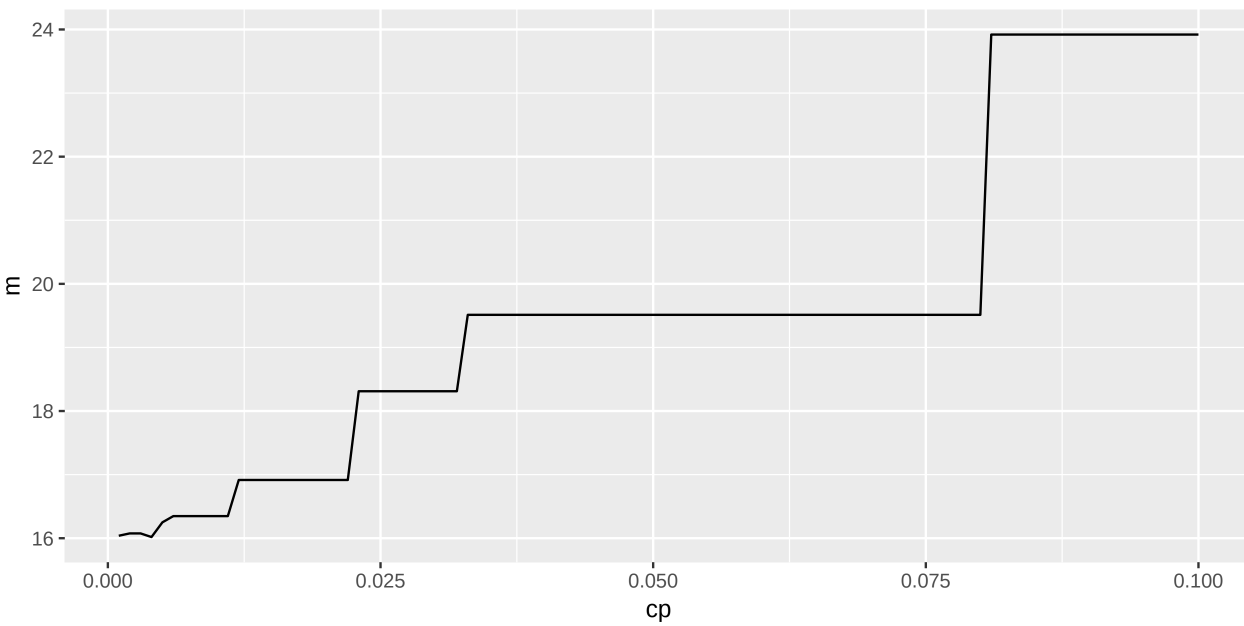


c. (2pts) Compute the MSE of the test data.

```
# [1] 16.34752
```

d. (3pts) Change the `cp` input parameter, try several different values. What cp value gives the best model, as measured by the smallest test MSE?

```
Several smallest cp values (0.001, 0.002, 0.003) give lowest MSE.
```



Question 2

a. (2pts) Fit a generalised linear model to the same set of variables.

```
#
# Call:  glm(formula = price ~ suburb + result + nbeds + property_type,
#     data = train)
#
# Coefficients:
#      (Intercept)  suburbSouth Yarra            resultS
#           1.3055             6.6904             0.4160
#         resultSP           resultVB              nbeds
#          -0.1047             0.4913             2.5862
#    property_typet     property_typeu
#          -2.8109            -6.3424
#
# Degrees of Freedom: 306 Total (i.e. Null);  299 Residual
# Null Deviance:        10380
# Residual Deviance: 4714   AIC: 1728
```

b. (3pts) Summarise the variable importance.

```
Based on the p-value suburb, nbeds and property type are very important. Result is not
important.
```

```
#                      Estimate Std. Error    t value     Pr(>|t|)
# (Intercept)         1.3055377  1.4578887  0.8954988 3.712408e-01
```

```
# suburbSouth Yarra  6.6903879  0.6596074 10.1429849 5.700794e-21
# resultS            0.4160357  0.7511930  0.5538333 5.801069e-01
# resultSP          -0.1047329  0.9156470 -0.1143813 9.090123e-01
# resultVB           0.4913121  0.9007158  0.5454684 5.858381e-01
# nbeds              2.5861912  0.3244152  7.9718550 3.331030e-14
# property_typet    -2.8108828  0.9245741 -3.0401919 2.573461e-03
# property_typeu    -6.3424178  0.6592840 -9.6201600 2.910570e-19
```

c. (2pts) Compute the MSE of the test data.

```
# [1] 18.49644
```

d. (3pts) Try including some interaction terms to improve the model, by reducing the test MSE.

```
This is the most complicated interaction model. It reduces the MSE a little, to beat the
default tree model.
```

```
#
# Call:  glm(formula = price ~ suburb * nbeds * property_type, data = train)
#
# Coefficients:
#                          (Intercept)
#                               8.0514
#                    suburbSouth Yarra
#                              -6.0044
#                                nbeds
#                               0.4574
#                       property_typet
#                              -4.9243
#                       property_typeu
#                              -7.2597
#              suburbSouth Yarra:nbeds
#                               4.5078
#     suburbSouth Yarra:property_typet
#                               2.8268
#     suburbSouth Yarra:property_typeu
#                               6.7558
#                 nbeds:property_typet
#                               0.4257
#                 nbeds:property_typeu
#                               1.3751
# suburbSouth Yarra:nbeds:property_typet
#                               0.1248
# suburbSouth Yarra:nbeds:property_typeu
#                              -3.7011
#
# Degrees of Freedom: 306 Total (i.e. Null);  295 Residual
# Null Deviance:        10380
# Residual Deviance: 3925   AIC: 1680
#                                      Estimate Std. Error     t value
# (Intercept)                         8.0513867  2.0672729  3.89468972
# suburbSouth Yarra                  -6.0044190  2.8023699 -2.14262185
# nbeds                               0.4574356  0.5834860  0.78397008
# property_typet                     -4.9242819  5.1918100 -0.94847112
# property_typeu                     -7.2597200  5.1427665 -1.41163710
# suburbSouth Yarra:nbeds             4.5078023  0.8601761  5.24055759
```

```
# suburbSouth Yarra:property_typet        2.8268224  6.9112353  0.40901840
# suburbSouth Yarra:property_typeu        6.7557655  5.5408703  1.21926072
# nbeds:property_typet                    0.4256978  1.4541012  0.29275664
# nbeds:property_typeu                    1.3750644  2.0090282  0.68444259
# suburbSouth Yarra:nbeds:property_typet  0.1248021  2.1667830  0.05759785
# suburbSouth Yarra:nbeds:property_typeu -3.7011225  2.1542821 -1.71803055
#                                            Pr(>|t|)
# (Intercept)                            1.216742e-04
# suburbSouth Yarra                      3.296168e-02
# nbeds                                  4.336867e-01
# property_typet                         3.436660e-01
# property_typeu                         1.591107e-01
# suburbSouth Yarra:nbeds                3.054226e-07
# suburbSouth Yarra:property_typet       6.828229e-01
# suburbSouth Yarra:property_typeu       2.237192e-01
# nbeds:property_typet                   7.699140e-01
# nbeds:property_typeu                   4.942331e-01
# suburbSouth Yarra:nbeds:property_typet 9.541080e-01
# suburbSouth Yarra:nbeds:property_typeu 8.684034e-02
# [1] 15.24866
```

Question 3

a. (2pts) Build a random forest model, using the default parameters. what is the reported MSE? (This is the training set MSE.)

```
MSE=18.76
```

```
#
# Call:
#  randomForest(formula = price ~ suburb + result + nbeds + property_type,      data = train_sub, impor
#                Type of random forest: regression
#                      Number of trees: 500
# No. of variables tried at each split: 1
#
#           Mean of squared residuals: 16.71716
#                     % Var explained: 50.57
```

b. (3pts) Summarise the variable importance. Which variable is the most important?

```
property type, nbeds and suburb are all important, but result is not.
```

```
#                   %IncMSE IncNodePurity
# suburb         7.53756464      638.4552
# result         0.07739069      398.9497
# nbeds         11.27918161     1790.6590
# property_type 13.44467749     2057.6298
```

c. (2pts) Compute the MSE of the test data.

```
# [1] 19.33436
```

d. (3pts) Explore the effect of `mtry` and `ntree` parameters, on the MSE.

```
Its not easy to get a better fit than the single tree!
```

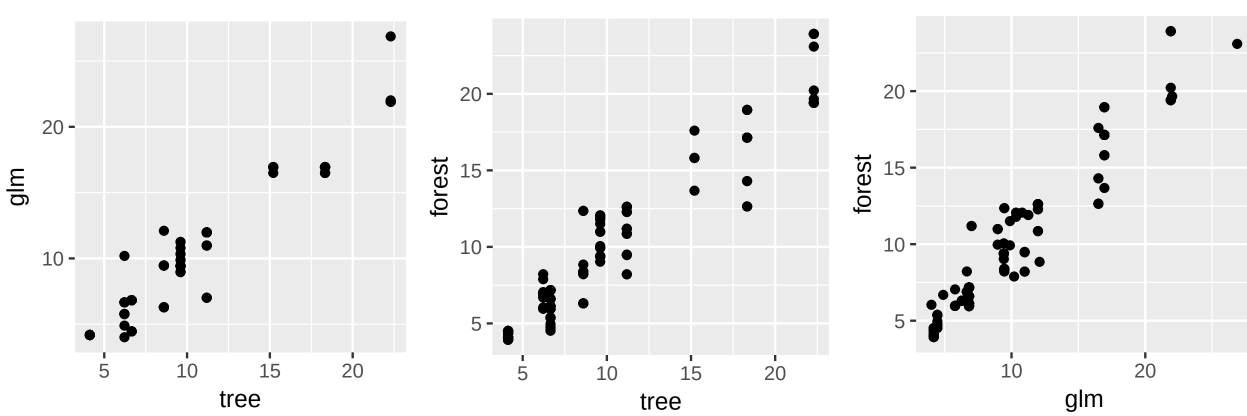
```
#
# Call:
```

```
#  randomForest(formula = price ~ suburb + result + nbeds + property_type,      data = train_sub, impor
#                Type of random forest: regression
#                      Number of trees: 500
# No. of variables tried at each split: 2
#
#           Mean of squared residuals: 15.89924
#                     % Var explained: 52.99
# [1] 16.79411
#
# Call:
#  randomForest(formula = price ~ suburb + result + nbeds + property_type,      data = train_sub, impor
#                Type of random forest: regression
#                      Number of trees: 10000
# No. of variables tried at each split: 2
#
#           Mean of squared residuals: 16
#                     % Var explained: 52.69
# [1] 16.80033
```

Question 4

(3pts) How do the predicted values compare for the different models? (Use the best model for each method.)

`There is positive linear association between the predictions from each method. The single tree model, and a fixed set of predicted values - see the stripes in the plots - but the forest produces more continuous predctions like the glm.`



TURN IN

- Your `.Rmd` file
- Your `html` file that results from knitting the Rmd.