Data Science with R
Decision Trees with Rattle

Graham.Williams@togaware.com

9th June 2014

Visit http://onepager.togaware.com/ for more OnePageR's.

In this module we use the **weather** dataset to explore the building of decision tree models in rattle (Williams, 2014).

The required packages for this module include:

```
library(rpart)        # Popular recursive partitioning decision tree algorithm
library(rattle)       # Graphical user interface for building decision trees
```

As we work through this chapter, new R commands will be introduced. Be sure to review the command's documentation and understand what the command does. You can ask for help using the `?` command as in:

```
?read.csv
```

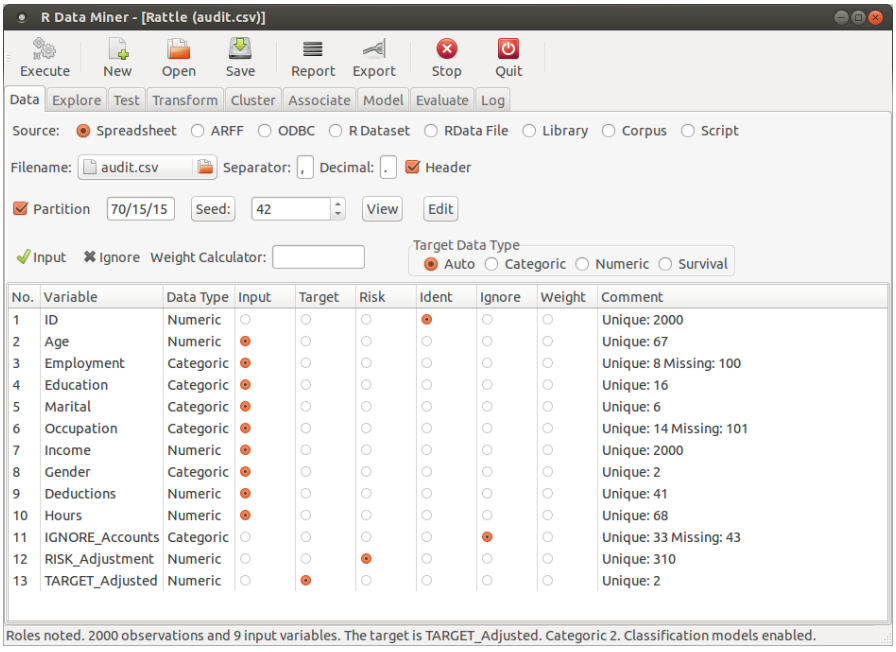
We can obtain documentation on a particular package using the *help=* option of `library()`:

```
library(help=rattle)
```

This chapter is intended to be hands on. To learn effectively, you are encouraged to have R running (e.g., RStudio) and to run all the commands as they appear here. Check that you get the same output, and you understand the output. Try some variations. Explore.

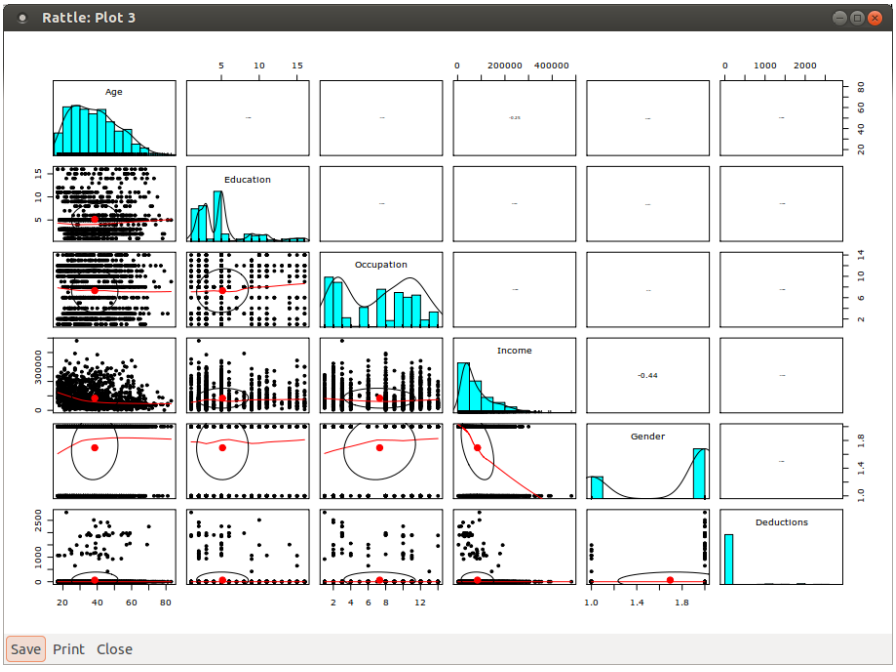
1 Loading the Data

1. On the Data tab, click the Execute button to load the default `weather` dataset (which is loaded after clicking `Yes`).
2. Note that we can click on the `Filename` chooser box to find some other datasets. Assuming we have just loaded the default `weather` dataset, we should be taken to the folder containing the actual CSV data files provided with Rattle.
3. Load the `audit` dataset.
4. Note that the variable TARGET_Adjusted is selected as the Target variable, and that the variable ID is identified as having a role of Ident(ifier).
5. Note also the variable RISK_Adjustment is set to have a role of Risk (this is based on its name). **For now, choose to give it the role of an `Input` variable.**
6. Choose to Partition the data. In fact, leave the 70/15/15 percentage split in the Partition text box as it is. Also, ensure the Partition checkbox is checked. This results in a random 70% of the dataset being used for training or building our models. A 15% sample of the dataset is used for validation, and is used whilst we are building and comparing different decision trees through the use of different parameters. The final 15% is for testing.
7. **Exercise: Research the issue of selecting a training/validation/testing dataset. Why is it important to partition the data in this way when data mining? Explain in a paragraph or two.**
8. **Exercise: Compare this approach to partitioning a dataset with the concept of cross fold validation. Report on this in one or two paragraphs.**
9. Be sure you have clicked the Execute button whilst on the Data tab. This will ensure that the sampling, for example, has taken place.



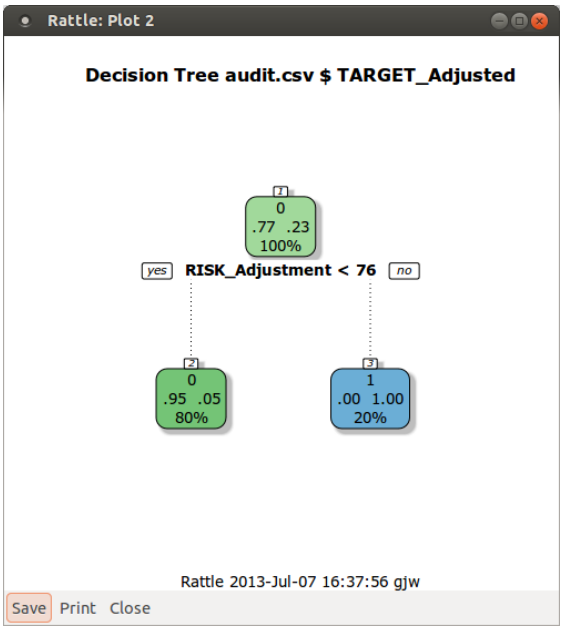
2 Exploring the Data

10. Click the View button to get a little familiar with the data.
11. **Exercise: What values of the variable TARGET_Adjusted correspond to an adjustment? How does TARGET_Adjusted relate to RISK_Adjustment?**
12. Explore the dataset further from the Explore tab.
13. Firstly, simply click Execute from the `Summary` option.
14. Explore some of the different summaries that we can generate.
15. Textual summaries are comprehensive, but sometimes take a little getting used to. Visual summaries can convey information more quickly. Select the `Distributions` option and then `Execute` (without any distributions being selected) to view a family of scatter plots.
16. **Exercise: What do the numbers in the plots represent?**
17. Choose one of each of the plot types for `Income`, then execute.
18. **Exercise: Research the use of Benford's law. How might it be useful in the context of the Benford's plot of Income? Discuss in one or two paragraphs.**
19. Rattle is migrating to a more sophisticated collection of graphics. To access this work in progress, from the `Settings` menu enable the `Advanced Graphics` option. Then select `Execute` again. It is experimental and if it fails, un-select the Option.
20. Now have a look at the distribution of Age against the target variable.
21. **Exercise: Are the different distributions significant? Explain each of the different elements of the plot.**



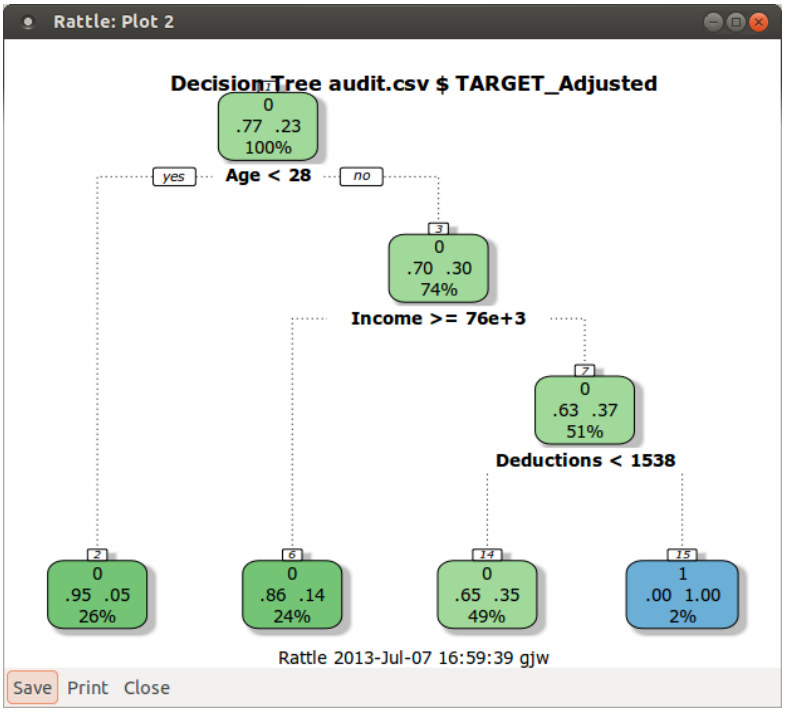
3 Naïvely Building Our First Decision Tree

22. Move to the Model tab, click the Execute button.
23. A decision tree is produced. Take a moment to understand what the description of the decision tree means. Click on the Draw button to see a visual presentation of the tree.
24. Go to the Evaluate tab and click the Execute button.
25. **Exercise: How accurate is this model? How many true positives are there? How many false positives are there? Which are better—false positives or false negatives? What are the consequences of false positives and of true positives in the audit scenario?**
26. **What is the fundamental flaw with the model we have just built?**
27. Go back to the Explore tab and have a look at the distribution of RISK_Adjustment against the target variable. **Does this explain the model performance?**
28. Go back to the Data tab and change RISK_Adjustment to be a Risk variable. Set to Ignore any variables that you feel are not suitable for decision tree classification. After having built further decision trees (or any models in fact) you might want to come back to the Data tab and change your variable selections. Be sure to click the Execute button each time.
29. **Exercise: What do you notice about the variables chosen in the new model? Why are categoric variables favoured over numeric variables? Research this issue and discuss in one or two paragraphs.**



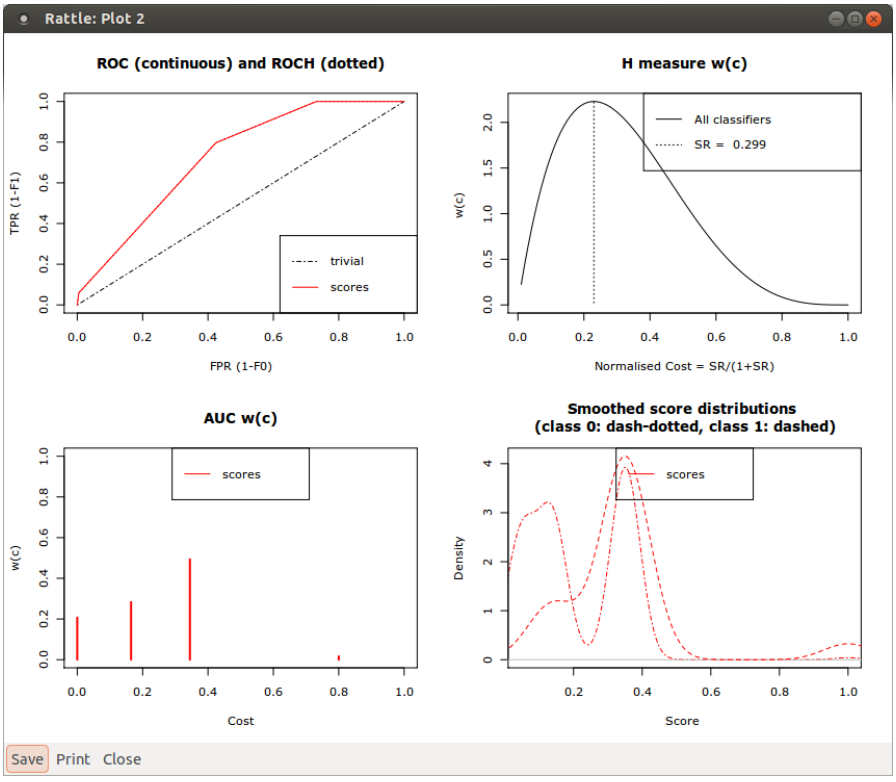
4 Building a Useful Decision Tree

30. Go to the `Model` tab and make sure the Tree radio button is selected.
31. Note the various parameters that can be set and modified. Read the Rattle documentation on Decision Trees for more information. You can also get additional help for these parameters from R by typing into the R console of RStudio: help(rpart.control).
32. **Which control options noted in the documentation of the rpart() command correspond to which Rattle options?**
33. Generate a new decision tree by clicking on `Execute` and inspect what is printed into the Rattle textview.
34. Click on `Draw` and a window with a decision tree will be shown.
35. **Which leaf node is most predictive of a tax return requiring an adjustment? Which is most predictive of not requiring an adjustment?**
36. Compare the decision tree drawing with the Summary of the rpart model in the main Rattle textview. Each leaf node in the drawing has a coloured number (which corresponds to the leaf node number), a 0 or 1 (which is the class label from the audit data set according to the target variable TARGET_Adjusted), and a percentage number (which corresponds to the accuracy of the classified training records in this leaf node).



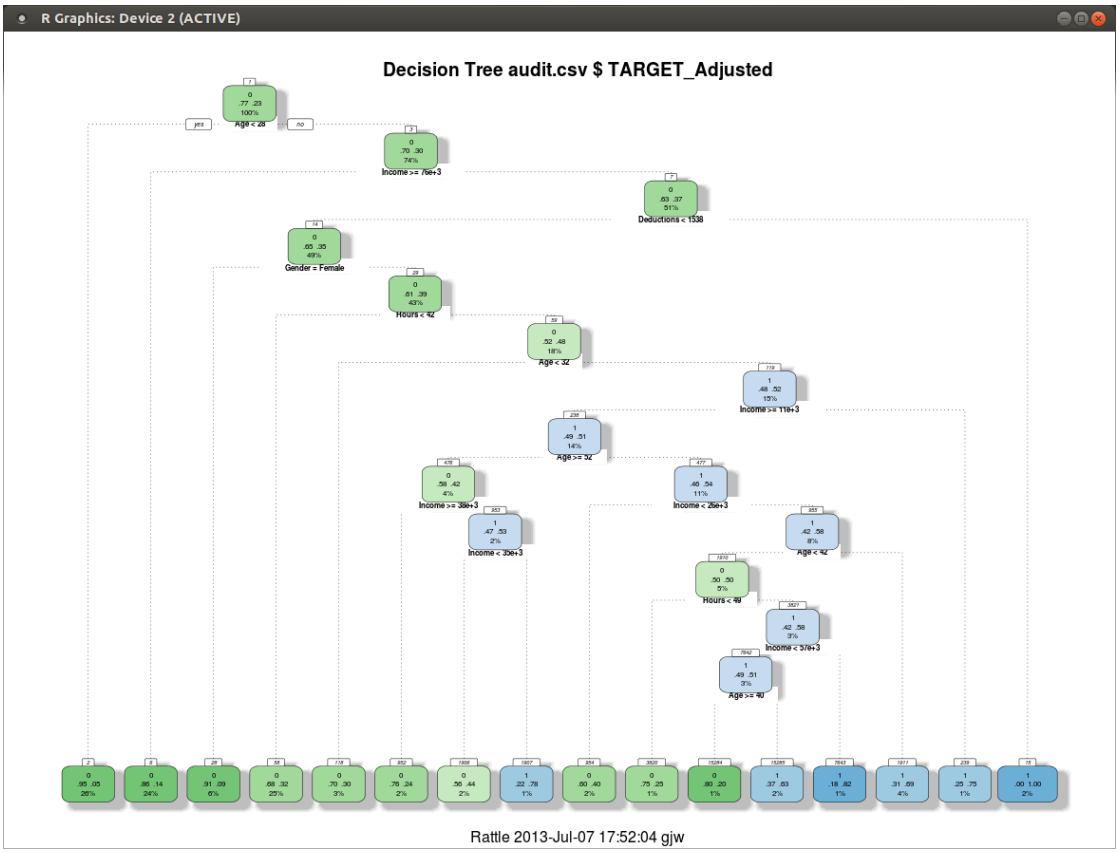
5 Evaluating the Model

37. On the `Evaluate` tab examine the different options to evaluate the accuracy of the decision tree we generated. Make sure the Validation and the Error Matrix radio buttons are both selected, and then click on Execute.
38. Check the error matrix that is printed (and write down the four numbers for each tree you generate).
39. **What is the error rate of the new decision tree? What is its accuracy?**
40. Click the `Training` radio button and again click on Execute.
41. **What is the error rate and what is the accuracy?**
42. **Why are the error rate and accuracy different between the validation and training settings?**
43. Select different Evaluate options then click on `Execute`.
44. You can read more on these evaluation measures in the Rattle book.
45. Investigate the ROC Curve graphical measure, as this is a widely used method to assess the performance of classifiers.



6 Fine Tuning the Decision Tree

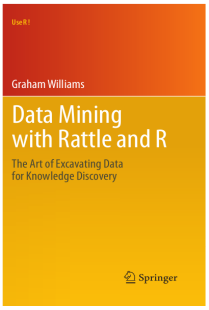
46. Go back to the `Model` tab and experiment with different values for the parameters Complexity, Min Bucket, Min Split and Max Depth.
47. **Which tree will give you the best accuracy, which one the worst? Which tree is the easiest to interpret? Which is hardest?** Remember that you need to click on `Execute` each time you change a parameter setting and want to generate a new tree.
48. Click on the `Rules` button to see the rules generated from a given tree.
49. **What is easier to understand, the tree diagram or the rules?**
50. **What is the highest accuracy you can achieve on the audit data set using the Rattle decision tree classifier? Which is the 'best' tree you can generate?**



7 Finishing Up

51. When finished we might like to save our project. Click on the `Save` icon.
52. We can give our project a name, like `audit_140609.rattle`.
53. Quit from Rattle and from R.
54. Now start up R and then Rattle again and load the project you saved.

8 Further Reading



The Rattle Book, published by Springer, provides a comprehensive introduction to data mining and analytics using Rattle and R. It is available from Amazon. Other documentation on a broader selection of R topics of relevance to the data scientist is freely available from `http://datamining.togaware.com`, including the

- Datamining Desktop Survival Guide.

This module is one of many OnePageR modules available from `http://onepager.togaware.com`. In particular follow the links on the website with a * which indicates the generally more developed OnePageR modules.

9 References

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Williams GJ (2009). "Rattle: A Data Mining GUI for R." *The R Journal*, **1**(2), 45–55. URL http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf.

Williams GJ (2011). *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Use R! Springer, New York. URL http://www.amazon.com/gp/product/1441998896/ref=as_li_qf_sp_asin_tl?ie=UTF8&tag=togaware-20&linkCode=as2&camp=217145&creative=399373&creativeASIN=1441998896.

Williams GJ (2014). *rattle: Graphical user interface for data mining in R*. R package version 3.0.4, URL http://rattle.togaware.com/.

This document, sourced from DTreesG.Rnw revision 419, was processed by KnitR version 1.6 of 2014-05-24 and took 1 seconds to process. It was generated by gjw on nyx running Ubuntu 14.04 LTS with Intel(R) Xeon(R) CPU W3520 @ 2.67GHz having 4 cores and 12.3GB of RAM. It completed the processing 2014-06-09 10:37:50.