

# Preface

Knowledge leads to wisdom and better understanding. Data mining builds knowledge from information, adding value to the ever-increasing stores of electronic data that abound today. Emerging from the database community in the late 1980s' data mining grew quickly to encompass researchers and technologies from machine learning, high-performance computing, visualisation, and statistics, recognising the growing opportunity to add value to data. Today, this multidisciplinary and transdisciplinary effort continues to deliver new techniques and tools for the analysis of very large collections of data. Working on databases that are now measured in the terabytes and petabytes, data mining delivers discoveries that can improve the way an organisation does business. Data mining enables companies to remain competitive in this modern, data-rich, information-poor, knowledge-hungry, and wisdom-scarce world. Data mining delivers knowledge to drive the getting of wisdom.

A wide range of techniques and algorithms are used in data mining. In performing data mining, many decisions need to be made regarding the choice of methodology, data, tools, and algorithms.

Throughout this book, we will be introduced to the basic concepts and algorithms of data mining. We use the free and open source software Rattle (Williams, 2009), built on top of the R statistical software package (R Development Core Team, 2011). As free software the source code of Rattle and R is available to everyone, without limitation. Everyone is permitted, and indeed encouraged, to read the source code to learn, understand verify, and extend it. R is supported by a worldwide network of some of the world's leading statisticians and implements all of the key algorithms for data mining.

This book will guide the reader through the various options that Rattle provides and serves to guide the new data miner through the use of Rattle. Many excursions into using R itself are presented, with the aim

of encouraging readers to use R directly as a scripting language. Through scripting comes the necessary integrity and repeatability required for professional data mining.

## Features

A key feature of this book, which differentiates it from many other very good textbooks on data mining, is the focus on the hands-on end-to-end process for data mining. We cover data understanding, data preparation, model building, model evaluation, data refinement, and practical deployment. Most data mining textbooks have their primary focus on just the model building—that is, the algorithms for data mining. This book, on the other hand, shares the focus with data and with model evaluation and deployment.

In addition to presenting descriptions of approaches and techniques for data mining using modern tools, we provide a very practical resource with actual examples using Rattle. Rattle is easy to use and is built on top of R. As mentioned above, we also provide excursions into the command line, giving numerous examples of direct interaction with R. The reader will learn to rapidly deliver a data mining project using software obtained for free from the Internet. Rattle and R deliver a very sophisticated data mining environment.

This book encourages the concept of programming with data, and this theme relies on some familiarity with the programming of computers. However, students without that background will still benefit from the material by staying with the Rattle application. All readers are encouraged, though, to consider becoming familiar with some level of writing commands to process and analyse data.

The book is accessible to many readers and not necessarily just those with strong backgrounds in computer science or statistics. At times, we do introduce more sophisticated statistical, mathematical, and computer science notation, but generally aim to keep it simple. Sometimes this means oversimplifying concepts, but only where it does not lose the intent of the concept and only where it retains its fundamental accuracy.

At other times, the presentation will leave the more statistically sophisticated wanting. As important as the material is, it is not always easily covered within the confines of a short book. Other resources cover such material in more detail. The reader is directed to the extensive

mathematical treatment by Hastie et al. (2009). For a more introductory treatment using R for statistics, see Dalgaard (2008). For a broader perspective on using R, including a brief introduction to the tools in R for data mining, Adler (2010) is recommended. For an introduction to data mining with a case study orientation, see Torgo (2010).

## Organisation

Chapter 1 sets the context for our data mining. It presents an overview of data mining, the process of data mining, and issues associated with data mining. It also canvasses open source software for data mining.

Chapter 2 then introduces *Rattle* as a graphical user interface (GUI) developed to simplify data mining projects. This covers the basics of interacting with R and *Rattle*, providing a quick-start guide to data mining.

Chapters 3 to 7 deal with data—we discuss the data, exploratory, and transformational steps of the data mining process. We introduce data and how to select variables and the partitioning of our data in Chapter 3. Chapter 4 covers the loading of data into *Rattle* and R. Chapters 5 and 6 then review various approaches to exploring the data in order for us to gain our initial insights about the data. We also learn about the distribution of the data and how to assess the appropriateness of any analysis. Often, our exploration of the data will lead us to identify various issues with the data. We thus begin cleaning the data, dealing with missing data, transforming the data, and reducing the data, as we describe in Chapter 7.

Chapters 8 to 14 then cover the building of models. This is the next step in data mining, where we begin to represent the knowledge discovered. The concepts of modelling are introduced in Chapter 8, introducing descriptive and predictive data mining. Specific descriptive data mining approaches are then covered in Chapters 9 (clusters) and 10 (association rules). Predictive data mining approaches are covered in Chapters 11 (decision trees), 12 (random forests), 13 (boosting), and 14 (support vector machines). Not all predictive data mining approaches are included, leaving some of the well-covered topics (including linear regression and neural networks) to other books.

Having built a model, we need to consider how to evaluate its performance. This is the topic for Chapter 15. We then consider the task of deploying our models in Chapter 16.

Appendix A can be consulted for installing R and Rattle. Both R and Rattle are open source software and both are freely available on multiple platforms. Appendix B describes in detail how the datasets used throughout the book were obtained from their sources and how they were transformed into the datasets made available through **rattle**.

## Production and Typographical Conventions

This book has been typeset by the author using L<sup>A</sup>T<sub>E</sub>X and R's `Sweave()`. All R code segments included in the book are run at the time of typesetting the book, and the results displayed are directly and automatically obtained from R itself. The Rattle screen shots are also automatically generated as the book is typeset.

Because all R code and screen shots are automatically generated, the output we see in the book should be reproducible by the reader. All code is run on a 64 bit deployment of R on a Ubuntu GNU/Linux system. Running the same code on other systems (particularly on 32 bit systems) may result in slight variations in the results of the numeric calculations performed by R.

Other minor differences will occur with regard to the widths of lines and rounding of numbers. The following options are set when typesetting the book. We can see that `width=` is set to 58 to limit the line width for publication. The two options `scipen=` and `digits=` affect how numbers are presented:

```
> options(width=58, scipen=5, digits=4, continue="  ")
```

Sample code used to illustrate the interactive sessions using R will include the R prompt, which by default is “>”. However, we generally do not include the usual continuation prompt, which by default consists of “+”. The continuation prompt is used by R when a single command extends over multiple lines to indicate that R is still waiting for input from the user. For our purposes, including the continuation prompt makes it more difficult to cut-and-paste from the examples in the electronic version of the book. The `options()` example above includes this change to the continuation prompt.

R code examples will appear as code blocks like the following example (though the continuation prompt, which is shown in the following example, will not be included in the code blocks in the book).

```
> library(rattle)

Rattle: A free graphical interface for data mining with R.
Version 2.6.7 Copyright (c) 2006-2011 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.

> rattle()

Rattle timestamp: 2011-06-13 09:57:52

> cat("Welcome to Rattle",
+     "and the world of Data Mining.\n")

Welcome to Rattle and the world of Data Mining.
```

In providing example output from commands, at times we will truncate the listing and indicate missing components with [...]. While most examples will illustrate the output exactly as it appears in R, there will be times where the format will be modified slightly to fit publication limitations. This might involve silently removing or adding blank lines.

In describing the functionality of **Rattle**, we will use a sans serif font to identify a **Rattle** widget (a graphical user interface component that we interact with, such as a button or menu). The kinds of widgets that are used in **Rattle** include the check box for turning options on and off, the radio button for selecting an option from a list of alternatives, file selectors for identifying files to load data from or to save data to, combo boxes for making selections, buttons to click for further plots or information, spin buttons for setting numeric options, and the text view, where the output from R commands will be displayed.

R provides very many *packages* that together deliver an extensive toolkit for data mining. **rattle** is itself an R package—we use a bold font to refer to R packages. When we discuss the functions or commands that we can type at the R prompt, we will include parentheses with the function name so that it is clearly a reference to an R function. The command **rattle()**, for example, will start the user interface for **Rattle**. Many functions and commands can also take arguments, which we indicate by trailing the argument with an equals sign. The **rattle()** command, for example, can accept the command argument **csvfile=**.

## Implementing Rattle

**Rattle** has been developed using the Gnome (1997) toolkit with the Glade (1998) graphical user interface (GUI) builder. Gnome is independent of any programming language, and the GUI side of **Rattle** started out using the Python (1989) programming language. I soon moved to **R** directly, once **RGtk2** (Lawrence and Temple Lang, 2010) became available, providing access to Gnome from **R**. Moving to **R** allowed us to avoid the idiosyncrasies of interfacing multiple languages.

The Glade graphical interface builder is used to generate an XML file that describes the interface independent of the programming language. That file can be loaded into any supported programming language to display the GUI. The actual functionality underlying the application is then written in any supported language, which includes Java, C, C++, Ada, Python, Ruby, and **R**! Through the use of Glade, we have the freedom to quickly change languages if the need arises.

**R** itself is written in the procedural programming language C. Where computation requirements are significant, **R** code is often translated into C code, which will generally execute faster. The details are not important for us here, but this allows **R** to be surprisingly fast when it needs to be, without the users of **R** actually needing to be aware of how the function they are using is implemented.

## Currency

New versions of **R** are released twice a year, in April and October. **R** is free, so a sensible approach is to upgrade whenever we can. This will ensure that we keep up with bug fixes and new developments, and we won't annoy the developers with questions about problems that have already been fixed.

The examples included in this book are from version 2.13.0 of **R** and version 2.6.7 of **Rattle**. **Rattle** is an ever-evolving package and, over time, whilst the concepts remain, the details will change. For example, the advent of **ggplot2** (Wickham, 2009) provides an opportunity to significantly develop its graphics capabilities. Similarly, **caret** (Kuhn et al., 2011) offers a newer approach to interfacing various data mining algorithms, and we may see **Rattle** take advantage of this. New data mining algorithms continue to emerge and may be incorporated over time.

Similarly, the screen shots included in this book are current only for the version of **Rattle** available at the time the book was typeset. Expect some minor changes in various windows and text views, and the occasional major change with the addition of new functionality.

Appendix A includes links to guides for installing **Rattle**. We also list there the versions of the primary packages used by **Rattle**, at least as of the date of typesetting this book.

## Acknowledgements

This book has grown from a desire to share experiences in using and deploying data mining tools and techniques. A considerable proportion of the material draws on over 20 years of teaching data mining to undergraduate and graduate students and running industry-based courses. The aim is to provide recipe-type material that can be easily understood and deployed, as well as reference material covering the concepts and terminology a data miner is likely to come across.

Many thanks are due to students from the Australian National University, the University of Canberra, and elsewhere who over the years have been the reason for me to collect my thoughts and experiences with data mining and to bring them together into this book. I have benefited from their insights into how they learn best. They have also contributed in a number of ways with suggestions and example applications. I am also in debt to my colleagues over the years, particularly Peter Milne, Joshua Huang, Warwick Graco, John Maindonald, and Stuart Hamilton, for their support and contributions to the development of data mining in Australia.

Colleagues in various organisations deploying or developing skills in data mining have also provided significant feedback, as well as the motivation, for this book. Anthony Nolan deserves special mention for his enthusiasm and ongoing contribution of ideas that have helped fine-tune the material in the book.

Many others have also provided insights and comments. Illustrative examples of using **R** have also come from the **R** mailing lists, and I have used many of these to guide the kinds of examples that are included in the book. The many contributors to those lists need to be thanked.

Thanks also go to the reviewers, who have added greatly to the readability and usability of the book. These include Robert Muenchen, Pe-

ter Christen, Peter Helmsted, Bruce McCullough, and Balázs Bárány. Thanks also to John Garden for his encouragement and insights in choosing a title for the volume.

*My very special thanks to my wife, Catharina, and children, Sean and Anita, who have endured my indulgence in bringing this book together.*

Canberra

*Graham J. Williams*



Data Mining with Rattle and R

The Art of Excavating Data for Knowledge Discovery

Williams, G.

2011, XX, 374 p. 95 illus., 80 illus. in color., Softcover

ISBN: 978-1-4419-9889-7