“Data Mining with Rattle and R”

DIT Analytics Club
07 March, 2013
“Peadar Kearney’s”, Dame Street

Colman McMahon
colman@thedata.co

www.linkedin.com/in/colmanmcmahon
Agenda

► Intro
► Data mining
► Rattle installation
► Rattle workflow
► Appendix
Agenda

- Intro
- Data mining
- Rattle installation
- Rattle workflow
- Appendix
Bio

► Professional
  » Background: Film, VFX, Digital Media
  » Currently: PhD Fellow, UCD Dynamics Lab

► PhD:
  » Policy network analysis
  » Creative Industries

► Technologies:
  » Social Network Analysis
  » Agent Based Modelling
  » Data/statistical analysis
Data Mining with Rattle and R

The Art of Excavating Data for Knowledge Discovery

Graham Williams

Springer, 2011

Rattle - “the R Analytical Tool To Learn Easily”

- Presents statistical & visual summaries of data
- Transforms data into forms that can be readily modelled
- Builds models (unsupervised and supervised) from the data
- Graphically presents the performance of models
- Scores new datasets
Rattle

► Built on the **statistical language R**
   
   » an understanding of R is not required in order to use it

► **Simple** to use, quick to deploy, and allows us to **rapidly work through** the data processing, modelling, and evaluation **phases** of a data mining project

► Can **migrate from Rattle to R** when we need to fine-tune and further develop our data mining projects
Rattle GUI
Agenda

- Intro
- Data mining
- Rattle stack
- Rattle workflow
- Appendix
CRISP-DM

Cross Industry Process for Data Mining (CRISP-DM, 1996)

» framework for delivering data mining projects.

1) Problem Understanding
2) Data Understanding
3) Data Preparation
4) Modelling
5) Evaluation
6) Deployment
Rattle GUI
Agenda

► Intro
► Data mining
► Rattle installation
► Rattle workflow
► Appendix
Rattle stack

Rattle

R
R - a sophisticated statistical software package

- easily installed, instructional, state-of-the-art, and it is free and open source

- The basic *modus operandi* - **write scripts** using the R language

- Steeper learning curve than using GUI based systems, but once over the hurdle, becomes relatively easy
R Project (www.r-project.org)

R : Copyright 2003, The R Development Core Team
Version 1.7.0 (2003-04-16)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type `license()` or `licence()` for distribution details.

R is a collaborative project with many contributors.
Type `contributors()` for more information.

Type `demo()` for some demos, `help()` for on-line help, or
`help.start()` for a HTML browser interface to help.
Type `q()` to quit R.

[Previously saved workspace restored]

>
Rattle (rattle.togware.com)

Welcome to Rattle (rattle.togware.com).

Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for statistical computing and graphics. Together they provide a sophisticated environments for data mining, statistical analyses, and data visualisation.

See the Help menu for extensive support in using Rattle. The Togaware Desktop Data Mining Survival Guide includes Rattle documentation and is available from datamining.togware.com

Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See Help -> About for details.

Rattle Version 2.6.7. Copyright 2006-2011 Togaware Pty Ltd
Rattle is a registered trademark of Togaware Pty Ltd
Installation

Install R

www.r-project.org

Start R

Install Rattle

> install.packages("rattle")

Load rattle into the R library

> library(rattle)
> rattle()
Agenda

- Intro
- Data mining
- Rattle installation
- Rattle workflow
- Appendix
Rattle workflow

1) Load a **dataset**

2) Select **variables and entities** for exploring and mining

3) **Explore the data** to understand how it is distributed or spread

4) **Transform the data** to suit our data mining purposes

5) Build our **models**

6) **Evaluate the models** on other datasets

7) **Export the models** for deployment
Workflow - tabs
Roles noted. 366 observations and 20 input variables. The target is RainTomorrow. Categoric 2. Classification models enabled.
Explore

Provides a textual overview of the data

```r
> summary(weather[7:9])

Sunshine      WindGustDir      WindGustSpeed
Min.    :0.00      NW     :73      Min.    :13.0
1st Qu.:5.95      NNW    :44      1st Qu.:31.0
Median :8.60        E      :37      Median :39.0
Mean   :7.91       WNW    :35      Mean   :39.8
3rd Qu.:10.50      ENE    :30      3rd Qu.:46.0
Max.   :13.60      (Other):144  Max.   :98.0
NA's   :3.00       NA's   :3     NA's   :2.0
```
```
> library(fBasics)
> basicStats(weather$Sunshine)

          X..weather.Sunshine
nobs       366.0000
NAs         3.0000
Minimum     0.0000
Maximum    13.6000
1. Quartile 5.9500
3. Quartile 10.5000
Mean        7.9094
Median      8.6000
Sum        2871.1000
SE Mean     0.1827
LCL Mean    7.5500
UCL Mean    8.2687
Variance    12.1210
Stdev       3.4815
Skewness    -0.7235
Kurtosis    -0.2706
```
Explore (graphically)

<table>
<thead>
<tr>
<th>No.</th>
<th>Variable</th>
<th>Box Plot</th>
<th>Histogram</th>
<th>Cumulative</th>
<th>Benford</th>
<th>Min; Median/Mean; Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>MinTemp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-5.30; 7.45/7.27; 20.90</td>
</tr>
<tr>
<td>4</td>
<td>MaxTemp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.60; 19.65/20.55; 35.80</td>
</tr>
<tr>
<td>5</td>
<td>Rainfall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00; 0.00/1.43; 39.80</td>
</tr>
<tr>
<td>6</td>
<td>Evaporation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.20; 4.20/4.52; 13.80</td>
</tr>
<tr>
<td>7</td>
<td>Sunshine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00; 8.60/7.91; 13.60</td>
</tr>
<tr>
<td>9</td>
<td>WindGustSpeed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.00; 39.00/39.84; 98.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Variable</th>
<th>Bar Plot</th>
<th>Dot Plot</th>
<th>Mosaic</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>WindGustDir</td>
<td></td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>WindDir9am</td>
<td></td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>WindDir3pm</td>
<td></td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>RainToday</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>24</td>
<td>RainTomorrow</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
Explore (graphically)
Figure 6.3: The parallel coordinates plot from latticist
Figure 6.9: Colourful brushing of multiple scatterplots
Figure 6.10: GGobi’s scatter plot matrix.
Figure 6.11: GGobi’s parallel coordinates plot.
various statistical tests, e.g. the T-test and F-test
- Normalising
- Filling in missing values
- Turning numeric variables into categoric variables (and vice versa)
- Dealing with outliers
- Removing variables or observations with missing values
Cluster

- Allows data miners to break data into more meaningful groups and then contrast the different clusters against each other.
Associate

Summary of the Apriori Association Rules:
Number of Rules: 127

Summary of the Measures of Interestingness:

<table>
<thead>
<tr>
<th>support</th>
<th>confidence</th>
<th>lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. : 0.100</td>
<td>Min. : 0.100</td>
<td>Min. : 0.714</td>
</tr>
<tr>
<td>1st Qu. : 0.100</td>
<td>1st Qu. : 0.500</td>
<td>1st Qu. : 1.429</td>
</tr>
<tr>
<td>Median : 0.100</td>
<td>Median : 1.000</td>
<td>Median : 2.500</td>
</tr>
<tr>
<td>Mean : 0.145</td>
<td>Mean : 0.759</td>
<td>Mean : 3.015</td>
</tr>
<tr>
<td>3rd Qu. : 0.100</td>
<td>3rd Qu. : 1.000</td>
<td>3rd Qu. : 5.000</td>
</tr>
<tr>
<td>Max. : 0.700</td>
<td>Max. : 1.000</td>
<td>Max. : 10.000</td>
</tr>
</tbody>
</table>

Summary of the Execution of the Apriori Command:

parameter specification:
  confidence minval smax arem aval originalSupport
  0.1 0.1 1 none FALSE TRUE
  support minlen maxlen target ext
  0.1 1 10 rules FALSE

algorithmic control:
  filter tree heap memopt load sort verbose
  0.1 TRUE TRUE FALSE TRUE 2 TRUE

apriori - find association rules with the apriori algorithm

The Association Rules model has been built. Time taken: 0.02 secs
Model

Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 256

node), split n, loss, yval, (yprob)  * denotes terminal node

1) root 256 41 No (0.83984 0.16016)
   2) Pressure3pm>=1012 294 16 No (0.92157 0.07843)
      4) Cloud3pm< 7.5 195 10 No (0.94072 0.05128) *
      5) Cloud3pm>=7.5 9 3 Yes (0.33333 0.66667) *
   3) Pressure3pm< 1012 52 25 No (0.51923 0.48077)
      6) Sunshine>=8.85 25 5 No (0.80000 0.20000) *
      7) Sunshine< 8.85 27 7 Yes (0.25926 0.74074) *
Decision Trees

Decision Tree weather.csv $ RainTomorrow

Pressure3pm $\geq 1011.9$

Cloud3pm $\leq 7.5$

4
- No
  - 195 obs
  - 94.9%

5
- Yes
  - 9 obs
  - 66.7%

Sunshine $\leq 8.85$

6
- No
  - 25 obs
  - 80%

7
- Yes
  - 27 obs
  - 74.1%
Boosting

Summary of the Ada Boost model:

Call:
ada(RainTomorrow ~ ., data = crs$dataset[crs$train, c(crs$input,
crs$target)], control = rpart.control(maxdepth = 30, cp = 0.01,
minsplits = 20, xval = 10), iter = 50)

Loss: exponential Method: discrete Iteration: 50

Final Confusion Matrix for Data:

<table>
<thead>
<tr>
<th>True value</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>213</td>
<td>2</td>
</tr>
<tr>
<td>Yes</td>
<td>15</td>
<td>26</td>
</tr>
</tbody>
</table>

Train Error: 0.066

Out-Of-Bag Error: 0.094 iteration= 41

Additional Estimates of number of iterations:

train.err1 train.kap1

The Ada Boost model has been built. Time taken: 1.62 secs
Boosting

Figure 13.3: The variable importance plot for a boosted model.
Evaluate

A variety of options for evaluating the performance of our models.

- Training
- Validation
- Testing
- Full
<table>
<thead>
<tr>
<th>Count</th>
<th>Predict</th>
<th>Percentage</th>
<th>Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>No 205 10</td>
<td>Actual</td>
<td>No 80 4</td>
</tr>
<tr>
<td></td>
<td>Yes 15 26</td>
<td></td>
<td>Yes 6 10</td>
</tr>
</tbody>
</table>

**Evaluation Using the Validation Dataset:**

<table>
<thead>
<tr>
<th>Count</th>
<th>Predict</th>
<th>Percentage</th>
<th>Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>No 39 5</td>
<td>Actual</td>
<td>No 72 9</td>
</tr>
<tr>
<td></td>
<td>Yes 5 5</td>
<td></td>
<td>Yes 9 9</td>
</tr>
</tbody>
</table>

**Evaluation Using the Testing Dataset:**

<table>
<thead>
<tr>
<th>Count</th>
<th>Predict</th>
<th>Percentage</th>
<th>Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>No 35 6</td>
<td>Actual</td>
<td>No 62 11</td>
</tr>
<tr>
<td></td>
<td>Yes 5 10</td>
<td></td>
<td>Yes 9 18</td>
</tr>
</tbody>
</table>

**Evaluation Using the Full Dataset:**

<table>
<thead>
<tr>
<th>Count</th>
<th>Predict</th>
<th>Percentage</th>
<th>Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>No 279 21</td>
<td>Actual</td>
<td>No 76 6</td>
</tr>
<tr>
<td></td>
<td>Yes 25 41</td>
<td></td>
<td>Yes 7 11</td>
</tr>
</tbody>
</table>
Figure 15.5: Four risk charts displayed to compare performances of multiple model builders on the audit dataset.
First Model
(example)
First model

Once we have processed our data, we can build a model

1) Click on the Execute button
   • Rattle will notice that no dataset has been identified
2) The sample “weather” dataset will be offered
   • Click “Yes”
3) Click on the Model tab
   • This is where we tell Rattle what kind of model we want to build
4) Click on the Execute button.
Roles noted. 366 observations and 20 input variables. The target is RainTomorrow. Categoric 2. Classification models enabled.
Building a Model

Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 256

node), split, n, loss, yval, (yprob)
  * denotes terminal node
1) root 256 41 No (0.83984 0.16016)
   2) Pressure3pm>=1012 204 16 No (0.92157 0.07843)
      4) Cloud3pm< 7.5 195 10 No (0.94872 0.05128) *
      5) Cloud3pm>=7.5 9 3 Yes (0.33333 0.66667) *
   3) Pressure3pm< 1012 52 25 No (0.51923 0.48077)
      6) Sunshine>=8.85 25 5 No (0.80000 0.20000) *
      7) Sunshine< 8.85 27 7 Yes (0.25926 0.74074) *

Classification tree:

The Decision Tree model has been built. Time taken: 0.03 secs
Figure 2.5

Decision Tree weather.csv $ RainTomorrow

Pressure3pm $ 1011.9

Cloud3pm $ 7.5

4
No
195 obs
94.9%

5
Yes
9 obs
66.7%

Sunshine $ 8.85

6
No
25 obs
80%

7
Yes
27 obs
74.1%
Figure 2.6: The target variable, RainTomorrow, is skewed, with Yes being quite under-represented
Appendix
Installation

Install R

www.r-project.org

Start R

Install Rattle

> install.packages("rattle")

Load rattle into the R library

> library(rattle)
> rattle()
Articles on Rattle

- Rattle: A Data Mining GUI for R

- Getting started with data mining in R using Rattle
  - http://techpad.co.uk/content.php?sid=240
R Resources (sample)

- **Data Manipulation with R** (Spector, 2008) - covers basic data structures, reading and writing data, subscripting, manipulating, aggregating, and reshaping data.

- **Introductory Statistics with R** (Dalgaard, 2008) - good introduction to statistics using R.

- **Modern Applied Statistics with S** (Venables and Ripley, 2002) - an extensive introduction to statistics using R.

- **Data Analysis and Graphics Using R** (Maindonald and Braun, 2007) - excellent practical coverage of many aspects of exploring and modelling data using R.

- **The Elements of Statistical Learning** (Hastie et al., 2009) is a more mathematical treatise, covering all of the machine learning techniques discussed in this book in quite some mathematical depth.

- **R for SAS and SPSS Users** (Muenchen, 2008) is an excellent choice.

- **Lattice: Multivariate Data Visualization with R** (Sarkar, 2008) - covers the extensive capabilities of one of the graphics/plotting packages available for R.

- **ggplot2: Elegant Graphics for Data Analysis** (Wickham, 2009) - newer graphics framework is detailed.

- **Bivand et al.** (2008) cover applied spatial data analysis,

- **Kleiber and Zeileis** (2008) cover applied econometrics.

- **Cowpertwait and Metcalfe** (2009) cover time series.
Agenda

► Intro
► Data mining
► Rattle installation
► Rattle workflow
► Appendix
“Data Mining with Rattle and R”

DIT Analytics Club
07 March, 2013
“Peadar Kearney’s”, Dame Street

Colman McMahon
colman@thedata.co

www.linkedin.com/in/colmanmcmahon