Data Science with R

Introducing Data Mining with Rattle and R

Graham.Williams@togaware.com

Senior Director and Chief Data Miner, Analytics
Australian Taxation Office

Visiting Professor, SIAT, Chinese Academy of Sciences
Adjunct Professor, Australian National University
Adjunct Professor, University of Canberra
Fellow, Institute of Analytics Professionals of Australia

Graham.Williams@togaware.com
http://datamining.togaware.com
Overview

1. An Introduction to Data Mining
2. The Rattle Package for Data Mining
3. Moving Into R
4. Getting Started with Rattle
Overview

1. An Introduction to Data Mining
2. The Rattle Package for Data Mining
3. Moving Into R
4. Getting Started with Rattle
Data Mining and Big Data

- Application of
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - Intuition

- To Big Data — Volume, Velocity, Variety

- ... to discover new knowledge
- ... to improve business outcomes
- ... to deliver better tailored services
The Business of Data Mining

- Australian Taxation Office
  - Lodgment ($110M)
  - Tax Havens ($150M)
  - Tax Fraud ($250M)

- IBM Buys SPSS for $1.2B in 2009
- SAS has annual revenue approaching $3B
- Analytics is >$100B business
- Amazon, eBay/PayPal, Google . . .
Basic Tools: Data Mining Algorithms

- Linear Discriminant Analysis (lda)
- Logistic Regression (glm)
- Decision Trees (rpart, wsrpart)
- Random Forests (randomForest, wsr)
- Boosted Stumps (ada)
- Neural Networks (nnet)
- Support Vector Machines (kernlab)
- ...

That’s a lot of tools to learn in R!
Many with different interfaces and options.
Overview

1. An Introduction to Data Mining
2. The Rattle Package for Data Mining
3. Moving Into R
4. Getting Started with Rattle
Why a GUI?

- Statistics can be complex and traps await
- **So many** tools in R to deliver insights
- Effective analyses should be scripted
- Scripting also required for repeatability
- R is a language for **programming** with data

How to remember how to do all of this in R?
How to skill up 150 data analysts with Data Mining?
Users of Rattle

Today, Rattle is used world wide in many industries

- Health analytics
- Customer segmentation and marketing
- Fraud detection
- Government

It is used by

- Consultants and Analytics Teams across business
- Universities to teach Data Mining

It is and will remain freely available.

CRAN and http://rattle.togaware.com
A Tour Thru Rattle: Startup

Welcome to Rattle (rattle.togaware.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.togaware.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 2.6.27 r104. Copyright 2006-2013 Togaware Pty Ltd
Rattle is a registered trademark of Togaware Pty Ltd

To Begin: Choose the data source, specify the details, then click the Execute button.
A Tour Thru Rattle: Loading Data

### Data Source Selection
- **Spreadsheet**
- **ARFF**
- **ODBC**
- **R Dataset**
- **R Data File**
- **Library**
- **Corpus**
- **Script**

### Filename Input
- **(None)**

#### Separation and Decimal Settings
- **Separator:**
- **Decimal:**

### Input Settings
- **Partition:** 70/15/15
- **Seed:** 42

### Target Data Type
- **Auto**
- **Categoric**
- **Numeric**
- **Survival**

### Variable Table
<table>
<thead>
<tr>
<th>No.</th>
<th>Variable</th>
<th>Data Type</th>
<th>Input</th>
<th>Target</th>
<th>Risk</th>
<th>Ident</th>
<th>Ignore</th>
<th>Weight</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Date</td>
<td>Ident</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 366</td>
</tr>
<tr>
<td>2</td>
<td>Location</td>
<td>Constant</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 1</td>
</tr>
<tr>
<td>3</td>
<td>MinTemp</td>
<td>Numeric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 180</td>
</tr>
<tr>
<td>4</td>
<td>MaxTemp</td>
<td>Numeric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 187</td>
</tr>
<tr>
<td>5</td>
<td>Rainfall</td>
<td>Numeric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 47</td>
</tr>
<tr>
<td>6</td>
<td>Evaporation</td>
<td>Numeric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 55</td>
</tr>
<tr>
<td>7</td>
<td>Sunshine</td>
<td>Numeric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 114 Missing: 3</td>
</tr>
<tr>
<td>8</td>
<td>WindGustDir</td>
<td>Categoric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 16 Missing: 3</td>
</tr>
<tr>
<td>9</td>
<td>WindGustSpeed</td>
<td>Numeric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 35 Missing: 2</td>
</tr>
<tr>
<td>10</td>
<td>WindDir9am</td>
<td>Categoric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 16 Missing: 31</td>
</tr>
<tr>
<td>11</td>
<td>WindDir3pm</td>
<td>Categoric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 16 Missing: 1</td>
</tr>
<tr>
<td>12</td>
<td>WindSpeed9am</td>
<td>Numeric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 22 Missing: 7</td>
</tr>
<tr>
<td>13</td>
<td>WindSpeed3pm</td>
<td>Numeric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 25</td>
</tr>
<tr>
<td>14</td>
<td>Humidity9am</td>
<td>Numeric</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>🗑️</td>
<td>Unique: 60</td>
</tr>
</tbody>
</table>

Roles noted. 366 observations and 20 input variables. The target is RainTomorrow. Categoric 2. Classification models enabled.
A Tour Thru Rattle: Explore Distribution
Correlation between weather.csv using Pearson

- Pressure3pm vs Pressure9am
- Pressure3pm vs Humidity9am
- Pressure3pm vs WindSpeed3pm
- Pressure3pm vs Humidity3pm
- Pressure3pm vs Sunshine
- Pressure3pm vs Cloud3pm
- Pressure3pm vs WindSpeed9am
- Pressure3pm vs Rainfall
- Pressure3pm vs WindGustSpeed
- Pressure3pm vs Cloud9am
- Pressure3pm vs Evaporation
- Pressure3pm vs Temp3pm
- Pressure3pm vs MaxTemp
- Pressure3pm vs Temp9am
- Pressure3pm vs MinTemp

Rattle 2013-May-13 17:24:27 gjw
A Tour Thru Rattle: Hierarchical Cluster
A Tour Thru Rattle: Decision Tree

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 (0.83984375 0.16015625)
   2) Pressure3pm=1011.9 204 16 No (0.92156863 0.07843137)
      4) Cloud3pm= 7.5 195 10 No (0.94871795 0.05128205)
      5) Cloud3pm> 7.5 9 3 Yes (0.33333333 0.66666667)
   3) Pressure3pm< 1011.9 52 25 No (0.51923077 0.48076923)
      6) Sunshine>= 0.85 25 5 No (0.80000000 0.20000000)
      7) Sunshine< 8.85 27 7 Yes (0.25925926 0.74074074)

Classification tree:
rpart(formula = RainTomorrow ~ ., data = crs$dataset[crs$train,
c(r$input, crs$target)], method = "class", parms = list(split = "information"),
control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] Cloud3pm Pressure3pm Sunshine

Root node error: 41/256 = 0.16016

The Decision Tree model has been built. Time taken: 0.03 secs
A Tour Thru Rattle: Decision Tree Plot

Decision Tree weather.csv $ RainTomorrow

Pressure3pm $= 1012

Cloud3pm $< 7.5

Sunshine $\geq 8.9

Cloud3pm $< 7.5

No

.92 .08

80%

Yes

.33 .67

4%

No

.80 .20

10%

Yes

.26 .74

11%

Rattle 2013-May-13 17:30:39 gjw

Save  Print  Close
A Tour Thru Rattle: Random Forest

---

**Summary of the Random Forest Model**

Number of observations used to build the model: 256
Missing value imputation is active.

**Call:**
```r
randomForest(formula = RainTomorrow ~ ., 
data = crss$dataset[crss$sample, c(crss$input, crss$target)], 
ntree = 500, mtry = 4, importance = TRUE, replace = FALSE, na.action = na.roughfix)
```

Type of random forest: classification
No. of variables tried at each split: 4

0.00B estimate of error rate: 13.28%

Confusion matrix:
```
No | Yes
---|---
26 | 8 0.0372093
26 | 15 0.6341463
```

Analysis of the Area Under the Curve (AUC)

The Random Forest model has been built. Time taken: 0.87 secs
A Tour Thru Rattle: Risk Chart
Overview

1. An Introduction to Data Mining
2. The Rattle Package for Data Mining
3. Moving Into R
4. Getting Started with Rattle
Data Miners are Programmers of Data

- Data miners are programmers of data
- A GUI can only do so much
- R is a powerful statistical language

Professional data mining
- Scripting
- Transparency
- Repeatability
# Rattle is Copyright (c) 2006-2013 Togaware Pty Ltd.

# Rattle timestamp: 2013-05-13 16:49:53 x86_64-pc-linux-gnu

# Rattle version 2.6.27 user 'gjw'

# Export this log textview to a file using the Export button or the Tools
# menu to save a log of all activity. This facilitates repeatability. Exporting
# to file 'myRf01.R', for example, allows us to the type in the R Console
# the command source('myRf01.R') to repeat the process automatically.
# Generally, we may want to edit the file to suit our needs. We can also directly
# edit this current log textview to record additional information before exporting.

# Saving and loading projects also retains this log.
library(rattle)

# This log generally records the process of building a model. However, with very
# little effort the log can be used to score a new dataset. The logical variable
# 'building' is used to toggle between generating transformations, as when building
# a model, and simply using the transformations, as when scoring a dataset.

building <- TRUE
scoring <- ! building

# The colorspace package is used to generate the colours used in plots, if available.
library(colorspace)
# Rattle timestamp: 2013-05-13 17:35:07 x86_64-pc-linux-gnu

# Random Forest

# The 'randomForest' package provides the 'randomForest' function.
require(randomForest, quietly=TRUE)

# Build the Random Forest model.
set.seed(crv$seed)

# The 'pROC' package implements various AUC functions.
require(pROC, quietly=TRUE)

# Calculate the Area Under the Curve (AUC)
Step 1: Identify the Data

dsname <- "weather"
target <- "RainTomorrow"
risk <- "RISK_MM"
form <- formula(paste(target, "~ ."))
ds <- get(dsname)
dim(ds)
## [1] 366 24

names(ds)
## [1] "Date" "Location" "MinTemp" "..."
## [5] "Rainfall" "Evaporation" "Sunshine" "..."
## [9] "WindGustSpeed" "WindDir9am" "WindDir3pm" "..."
## [13] "WindSpeed3pm" "Humidity9am" "Humidity3pm" "..."
....
### Step 2: Observe the Data

```r
def(head(ds)
## Date Location MinTemp MaxTemp Rainfall Evapora...
## 1 2007-11-01 Canberra 8.0 24.3 0.0 ...  
## 2 2007-11-02 Canberra 14.0 26.9 3.6 ...  
## 3 2007-11-03 Canberra 13.7 23.4 3.6 ...  
## 4 2007-11-04 Canberra 13.3 15.5 39.8 ...  
## 5 2007-11-05 Canberra 7.6 16.1 2.8 ... 
....

def(tail(ds)
## Date Location MinTemp MaxTemp Rainfall Evapora...
## 361 2008-10-26 Canberra 7.9 26.1 0 ...  
## 362 2008-10-27 Canberra 9.0 30.7 0 ...  
## 363 2008-10-28 Canberra 7.1 28.4 0 ...  
## 364 2008-10-29 Canberra 12.5 19.9 0 ...  
## 365 2008-10-30 Canberra 12.5 26.9 0 ... 
....
```
**Step 2: Observe the Data**

```r
str(ds)
```

```r
# 'data.frame': 366 obs. of 24 variables:
# $ Date       : Date, format: "2007-11-01" "2007-11-... 
# $ Location   : Factor w/ 46 levels "Adelaide","Alba... 
# $ MinTemp    : num  8 14 13.7 13.3 7.6 6.2 6.1 8.3 ... 
# $ MaxTemp    : num  24.3 26.9 23.4 15.5 16.1 16.9 1... 
# $ Rainfall   : num  0 3.6 3.6 39.8 2.8 0 0.2 0 0 16... 
```

```r
summary(ds)
```

```r
# Date Location MinTemp ...
# Min. :2007-11-01 Canberra :366 Min. : -5.3...
# 1st Qu.:2008-01-31 Adelaide : 0 1st Qu.: 2.3...
# Median :2008-05-01 Albany : 0 Median : 7.4...
# Mean  :2008-05-01 Albury : 0 Mean : 7.2...
# 3rd Qu.:2008-07-31 AliceSprings : 0 3rd Qu.:12.5...
```

Step 3: Clean the Data — Identify Variables

```r
(ignore <- c(names(ds)[c(1,2)], risk))

## [1] "Date" "Location" "RISK_MM"

(vars <- setdiff(names(ds), ignore))

## [1] "MinTemp" "MaxTemp" "Rainfall" ...
## [5] "Sunshine" "WindGustDir" "WindGustSpeed" ...
## [9] "WindDir3pm" "WindSpeed9am" "WindSpeed3pm" ...
## [13] "Humidity3pm" "Pressure9am" "Pressure3pm" ...

dim(ds[vars])

## [1] 366 21
```
Step 3: Clean the Data — Remove Missing

```r
dim(ds[vars])
## [1] 366 21

sum(is.na(ds[vars]))
## [1] 47

ds <- na.omit(ds[vars])
sum(is.na(ds))
## [1] 0

dim(ds)
## [1] 328 21
```
Step 3: Clean the Data—Target as Categoric

```
summary(ds[[target]])

## RainTomorrow
## Min.  :0.000
## 1st Qu.:0.000
## Median :0.000
## Mean  :0.183
## 3rd Qu.:0.000
## Max.  :1.000
##

ds[[target]] <- as.factor(ds[[target]])
levels(ds[[target]]) <- c("No", "Yes")
summary(ds[[target]])

## RainTomorrow
## 0:268
## 1: 60
```

http://togaware.com
Step 4: Build the Model—Train/Test

(n <- nrow(ds))

## [1] 328

train <- sample(1:n, 0.70*n)
length(train)

## [1] 229

test <- setdiff(1:n, train)
length(test)

## [1] 99
library(randomForest)

## randomForest 4.6-7
## Type rfNews() to see new features/changes/bug fixes.

m <- randomForest(form, ds[train,])
m

## Call:
## randomForest(formula=form, data=ds[train,])
## Type of random forest: classification
## Number of trees: 500
## No. of variables tried at each split: 4
##
## OOB estimate of error rate: 12.23%
## Confusion matrix:
## ....
Step 5: Evaluate the Model—Risk Chart

```r
pr <- predict(m, ds[test,], type="prob")[,2]
ev <- evaluateRisk(pr, ds[test, target], ds[test, risk])
riskchart(ev)
```
Overview

1. An Introduction to Data Mining
2. The Rattle Package for Data Mining
3. Moving Into R
4. Getting Started with Rattle
Installation

- Rattle is built using R
- Need to download and install R from cran.r-project.org
- Recommend also install RStudio from www.rstudio.org

Then start up RStudio and install Rattle:

    install.packages("rattle")

Then we can start up Rattle:

    rattle()

- Required packages are loaded as needed.
Resources and References

- Rattle: http://rattle.togaware.com
- OnePageR: http://onepager.togaware.com
- Guides: http://datamining.togaware.com
- Practise: http://analystfirst.com

- Book: Data Mining using Rattle/R
- Chapter: Rattle and Other Tales
- Paper: A Data Mining GUI for R — R Journal, Volume 1(2)
Thank You