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, wsrf)
- 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

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

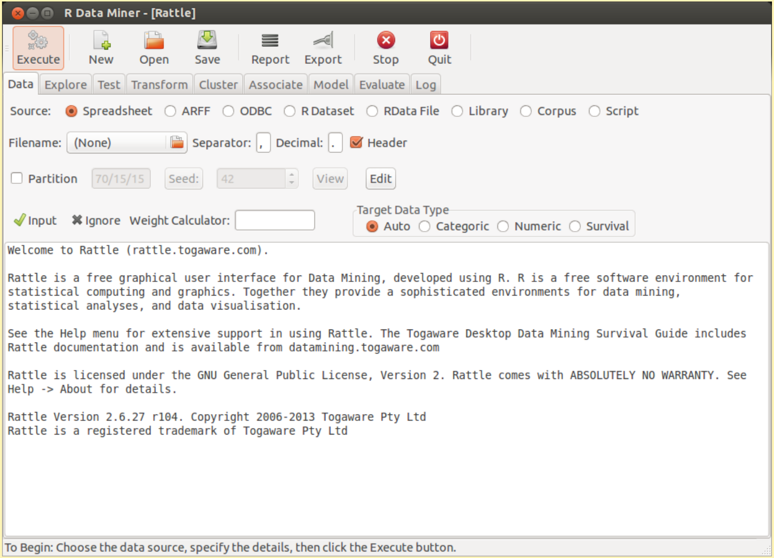
It is used by

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

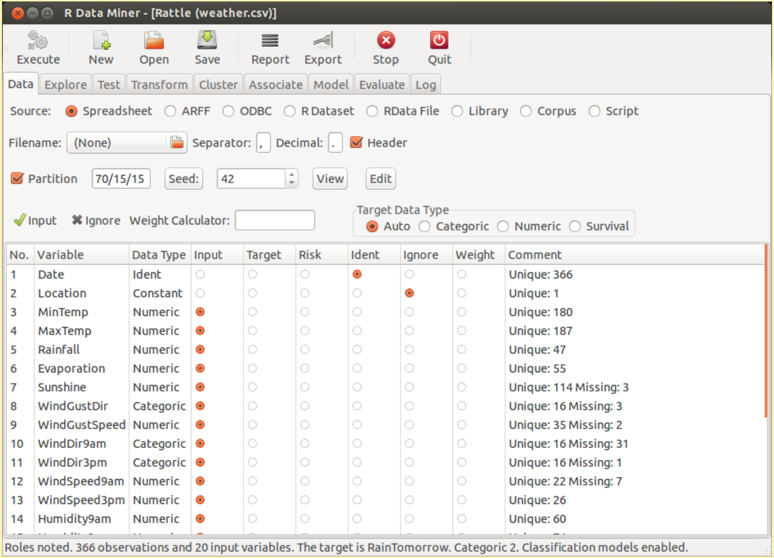
It is and will remain freely available.

CRAN and `http://rattle.togaware.com`

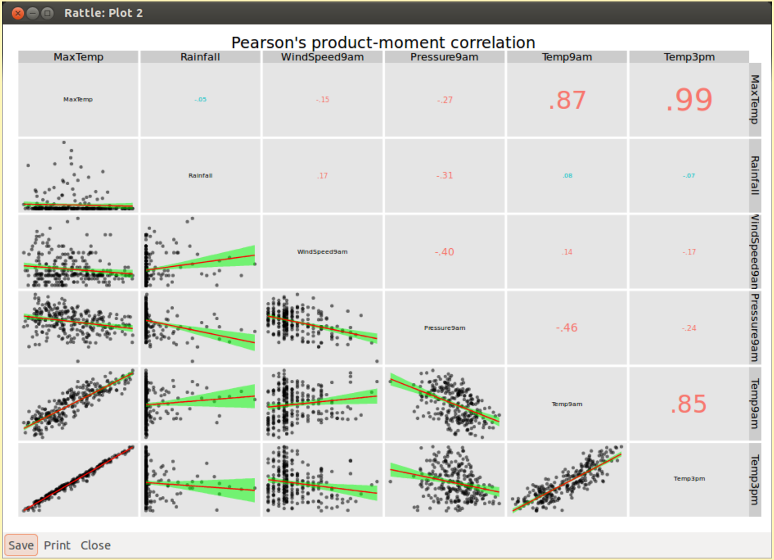
A Tour Thru Rattle: Startup



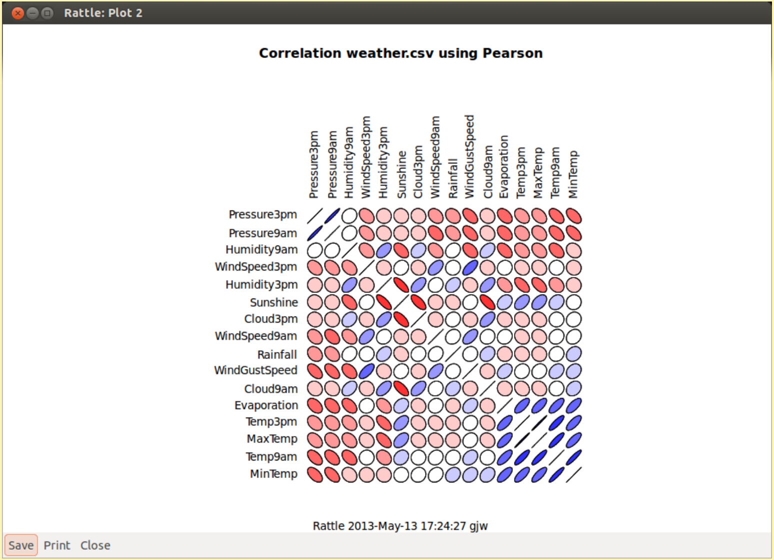
A Tour Thru Rattle: Loading Data



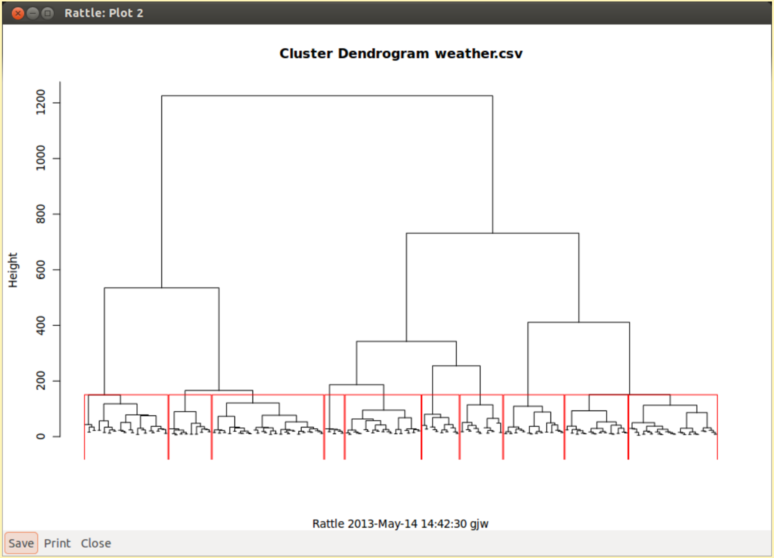
A Tour Thru Rattle: Explore Distribution



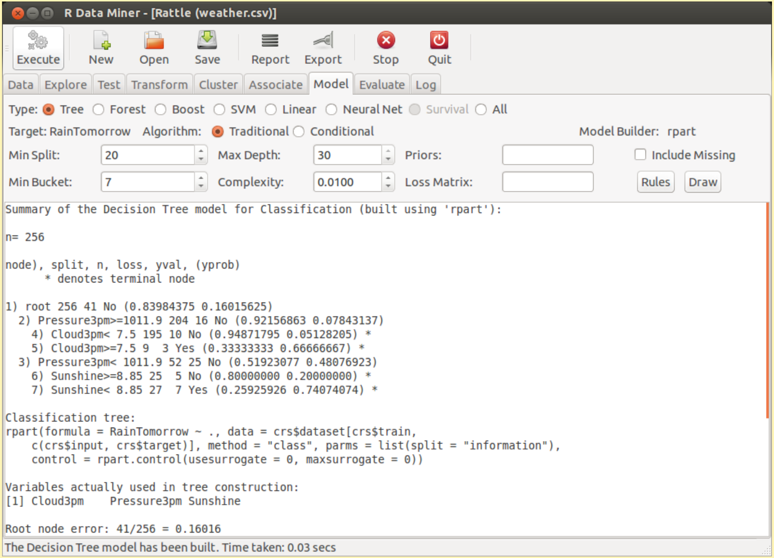
A Tour Thru Rattle: Explore Correlations



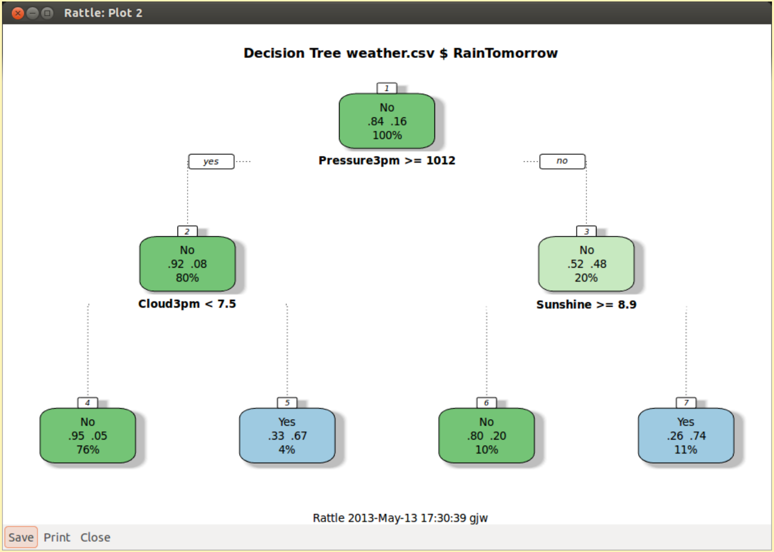
A Tour Thru Rattle: Hierarchical Cluster



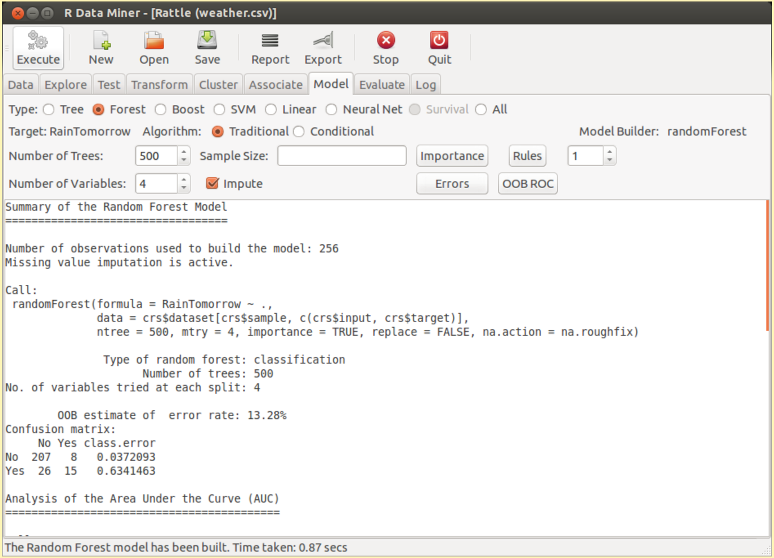
A Tour Thru Rattle: Decision Tree



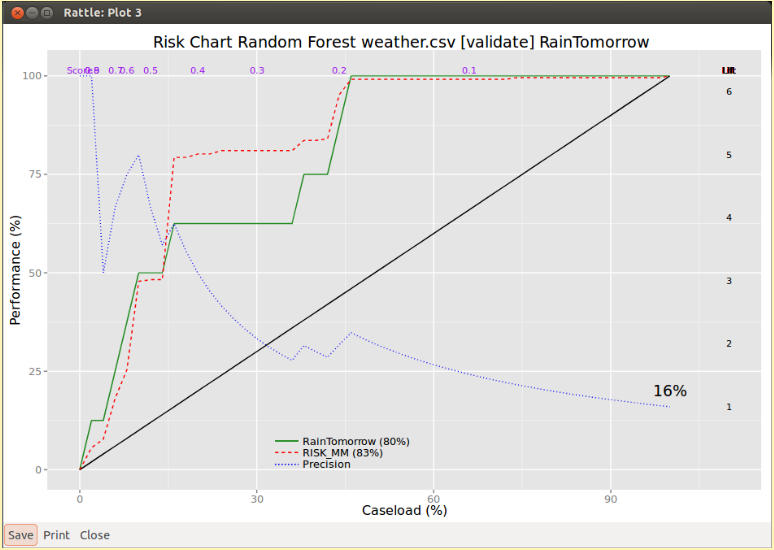
A Tour Thru Rattle: Decision Tree Plot



A Tour Thru Rattle: Random Forest



A Tour Thru Rattle: Risk Chart



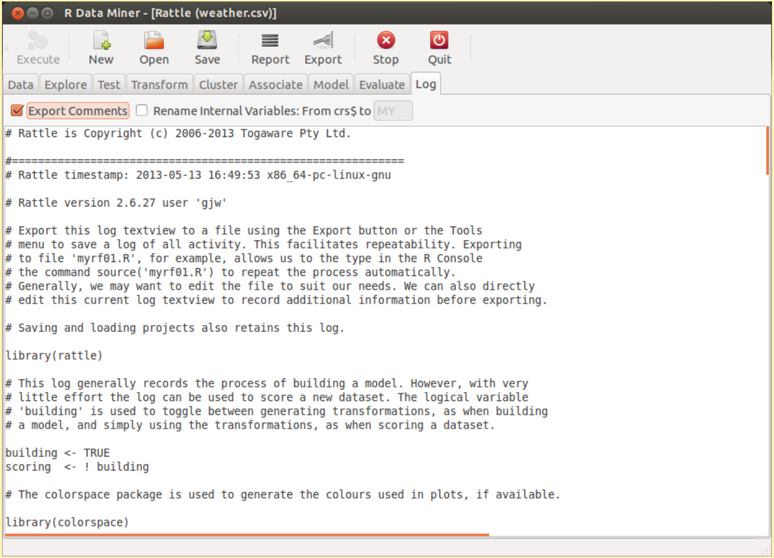
Overview

Data Miners are Programmers of Data

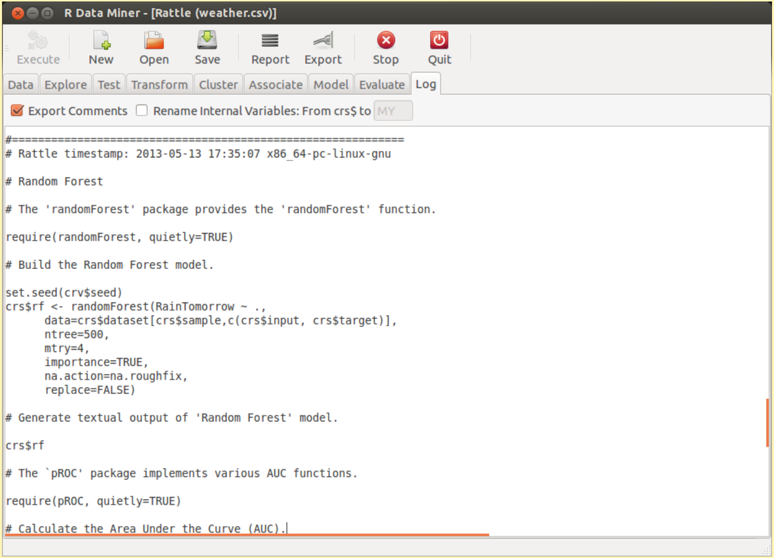
- Data miners are programmers of data
- A GUI can only do so much
- R is a powerful statistical language

- Professional data mining
  - Scripting
  - Transparency
  - Repeatability

From GUI to CLI — Rattle's Log Tab



From GUI to CLI — Rattle's Log Tab



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

```
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        ...
....

tail(ds)

##            Date Location MinTemp MaxTemp Rainfall Evapo...
## 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

```
str(ds)

## '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...
....

summary(ds)

##       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

```
(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

```
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
```

Step 4: Build the Model—Train/Test

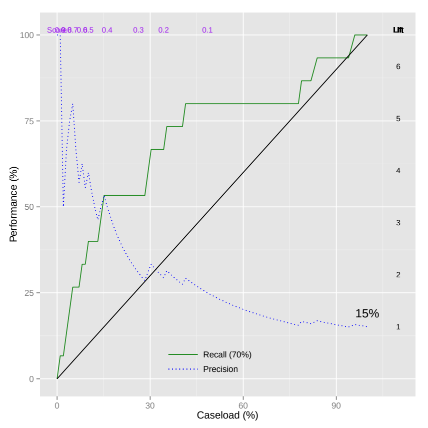
```
(n <- nrow(ds))

## [1] 328

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

## [1] 229

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

## [1] 99
```

Step 4: Build the Model—Random Forest

```
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

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



Overview

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:

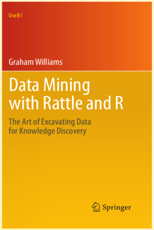
```
rattle()
```

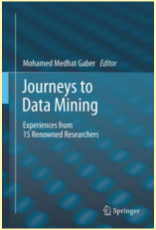
- Required packages are loaded as needed.

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)






Contributed Research Articles 45

Rattle: A Data Mining GUI for R

by Graham J Williams

Implementation

Introduction

The R Journal Vol. 1/2, December 2009 ISSN 2073-4859

Thank You