Data Mining in Government

From R to Rattle and Back Again

Graham Williams

Chief Data Miner and Director of Analytics
Office of the Chief Knowledge Officer
Australian Taxation Office

Adjunct Professor, University of Canberra
Adjunct Professor, Australian National University
Fellow, Institute of Analytics Professionals of Australia

Graham.Williams@ato.gov.au
http://datamining.togaware.com

Download instructions: http://rattle.togaware.com

Overview

Analytics Case Study
- Data Mining
- Australian Taxation Office

Challenge — Deliver Data Mining Capability
- Competing Technologies
- Commodity and Open Source

Rattle in Action
- Rattle Screenshots
- Moving Forward with Rattle

Data Mining — In a Nutshell

- The non-trivial extraction of novel, implicit, and actionable knowledge from large databases and in a *timely* manner.
- Build models (regression, neural networks, decision trees, random forests, boosted trees, support vector machines, association and correlation analysis, cluster analysis, ...) from data that represent observations of the world, and often whole populations.
- *Knowledge from data any way we can.*
- Technology to enable data processing, data understanding and visualisation, and data analysis of very large databases for insight and hypothesis generation.

Australian Taxation Office — Case Study

- Employs 22,000 staff Australia wide
- Revenue Collection and Refund Management
- Compliance and Risk Modelling

- 12M Individuals, $450B Income, $100B Tax
- 2M Companies..., $1800B Income, $40B Tax
- Deductions $B

- Tax payer's charter:
  Fair but firm; *Protect privacy; Assume honest*
- Service standards — turn around refunds
- Whilst protecting integrity of revenue collection

ATO Analytics — Deploying Data Mining

- Established as an organisational capability in 2003
  Add a scientific basis to managing Risk
- Team of 16 data mining specialists
  Statistics and Machine Learning
- Support 120 data analysts
- Spread data mining throughout the organisation through a central centre of excellence
- Provide a unified framework for Risk Management across the organisation

Analytics Case Study — High Risk Refunds

High Risk Refunds (HRR) identified prior to payment of refunds.

- Current business rules identify too many "high risk" refunds.
- We might identify 100,000 cases each year.
- Sometimes as few as 5% require adjustment.
- But revenue at risk is significant (from $10m to $1b).

Current practise of simple business rules based on experience:

- Total claimed investment deductions $>$ $N
- Ratio of self education deductions to total income $>$ N
- Total international transfers $>$ N times taxable income
- Luxury vehicle purchase $M $>$ N times taxable income

Analytics Case Study — High Risk Refunds

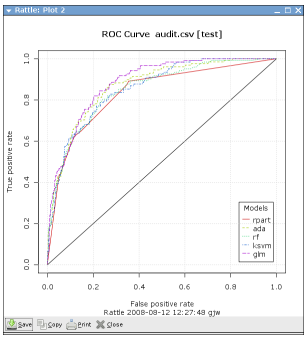
Almost, any kind of modelling will help ... data mining for HRR

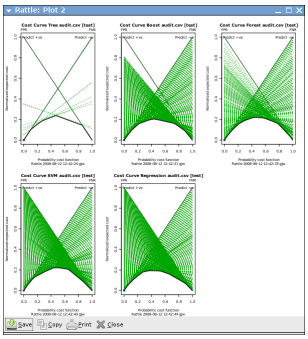
- Major task is all about the data:
  - data understanding/preparation, feature generation/selection
  - 100,000 cases by 1,000 variables
- Stock and trade: glm, rpart, ada, randomForest, kernlab
- Simple binary classification and $ regression

- Identify new characteristics to target high risk (5%);
- Focus resources on productive cases - $ and tax payer benefit;
- Decision trees and ensembles (random forests) are often effective.

Analytics Case Study — High Risk Refunds

"Out of the box" model performance:





Other Areas of Modelling

- High Risk Refunds
- Required to Lodge ($110M)
- Assessing Levels of Debt
  - Propensity to Pay
  - Capacity to Pay
- Determining Optimal Treatment Strategies
- Identity Theft — eTax and International
- Project Wickenby Text Mining
- Tax Havens — AUSTRAC

Overview

Analytics Case Study
Data Mining
Australian Taxation Office

Challenge — Deliver Data Mining Capability
Competing Technologies
Commodity and Open Source

Rattle in Action
Rattle Screenshots
Moving Forward with Rattle

Packaging Analytics

- Simple enough for the R jockey
- But how to roll this capability out to the point-n-click generation
- SAS/EM and SPSS/Clementine and KXEN and Salford Systems ...

- But, why not invest in people, not vendors?
- R provides quite a capable data mining environment, and has much more to offer than the competitors.

- Challenge: How to roll R out to a large community of data analysts?
- Answer: "Package" (or template) analyses tuned for the kinds of analyses required, reviewed and vetted by our Statisticians and Data Miners.

Background — Technologies

- Originally purchased data mining tools (SAS/EM, Teradata Warehouse Miner) and hardware (Big Iron MS/Windows 32 bit).

- But data mining needs skilled people, not off the shelf solutions (yet).

- Also data mining technology is rapidly developing, and commercial tools not always up to date.

New Approaches Ensembles

Commercial software is lagging behind advances in Data Mining (PhD 1989 combining multiple decision trees.... not new technology)

- Current best off the shelf technology includes random forests, boosting and support vector machines - SAS/EM, SPSS?
- Open source solutions allow organisational investment in people.
- Decide in 2005 to move to open source platform — 3 years to deploy!!

AnalyticsNet

Build a network of DataMining Nodes:

- 1 CPU (2 Cores), AMD64, 16GB RAM, 300GB Disk
- 4 CPU (8 Cores), AMD64, 32GB RAM, 1TB Disk (Good Price Point)
- 8 CPU (16 Cores), AMD64, 128GB RAM, 10TB Disk (Near Term)

- Best of class open source operating system (Debian GNU/Linux)
- Open Source data mining tools R, Rattle, Weka, AlphaMiner
- Open Source does deliver quality software
  A difficult sell (commercial interests).

Data source: Data Warehouse (Teradata or Netezza or SQLite) as the workhorse data server

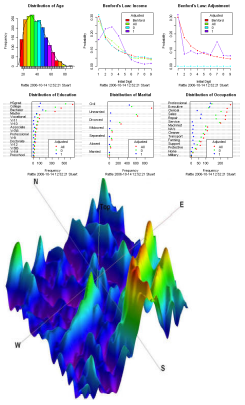
The Challenge Remains...

But how to roll out R across an organisation?

- Shortage of skilled data miners
- Skilled data analysts

- What do you mean "a CLI?"
- Development of training - using a GUI

- Integrity of statistical analyses
- Community of Practise for peer review



Overview

Analytics Case Study
Data Mining
Australian Taxation Office

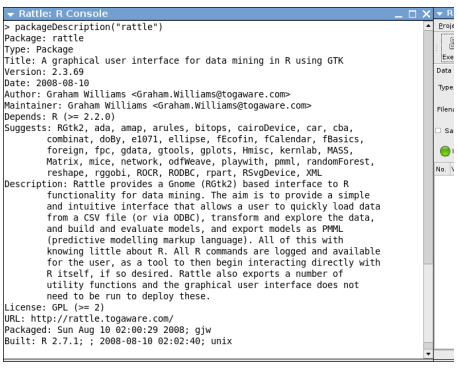
Challenge — Deliver Data Mining Capability
Competing Technologies
Commodity and Open Source

Rattle in Action
Rattle Screenshots
Moving Forward with Rattle

Introducing Rattle

A stepping stone into the full power of R (quick ref) — original plan
or
A self contained tool for data mining — sufficient but limiting

1. Load Data
   CSV, ARFF, ODBC
2. Explore
   Plots, GGobi
3. Transform
   Rescale, Remap
4. Cluster, Associate
5. Predictive Models
   Tree, Forest, NNet
6. Evaluate and Score
   ROC, Risk, Cost
7. Log



Overview

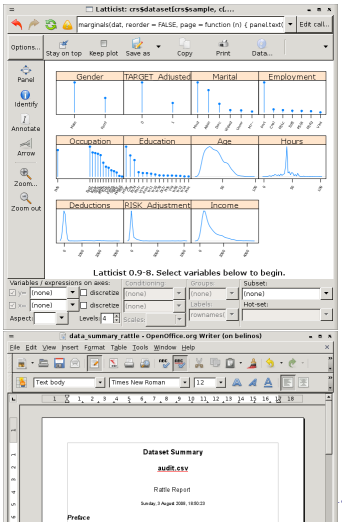
Analytics Case Study
Data Mining
Australian Taxation Office

Challenge — Deliver Data Mining Capability
Competing Technologies
Commodity and Open Source

Rattle in Action
Rattle Screenshots
Moving Forward with Rattle

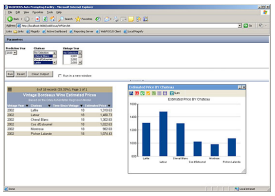
Future Directions

- latticist and playwith
- odfWeave report generator
- tm for text mining
- pmml consumer
- dealing with large data: wide and deep



Business Intelligence and Data Mining

- Information Builders
  BI Tool — WebFOCUS
  (Announced 2 Jun)
- Incorporate Open Source Rattle
  Badged as RStat
  Open Source Contributions



- SPSS can now create an R data frame object and (somehow — but not in the same memory space?) shares the object between R and SPSS.
- Allows customers to drive the SPSS technology directions!

Interoperabililty using PMML

- Share models: import and export
- Commercial (... extensions are bad) and Open Source vendor support
- Uptake of the technology?
- Rattle's PMML now exports PMML for:
  - lm; glm; rpart; kmeans; randomForest; randomSurvivalForest; arules, nnet
- Import into Zementis' ADAPA (Amazon Clouds)
- Integrated into WebFOCUS

Resources

- Data Mining
  - `datamining.togaware.com`
- Tools:
  - rattle.togaware.com (Rattle and PMML beta versions)
  - www.cs.waikato.ac.nz/ml/weka/
  - www.knime.org
  - rapid-i.com

Thank You