Rattle: R for Data Mining

Experiences in Government and Industry

Graham Williams

Senior Director and Principal Data Miner
Australian Taxation Office

Adjunct Professor, University of Canberra and ANU
Fellow, Institute of Analytics Professionals of Australia

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

Overview

Setting the Context

Background

Australian Taxation Office

Tooling up for Data Mining

Technologies

Commodity and Open Source

Delivering Outcomes

Overview

Setting the Context

Background

Australian Taxation Office

Tooling up for Data Mining

Technologies

Commodity and Open Source

Delivering Outcomes

Data is Fundamental

Sherlock Holmes:

> *"It is a capital mistake to theorize before one has data. Insensibly, one begins to twist facts to suit theories, instead of theories to suit facts."*

A Scandal in Bohemia (1891)
Arthur Conan Doyle

Data Mining is fundamentally about delivering novel and actionable knowledge from mountains of data.

Data is Fundamental

Sherlock Holmes:

> *"It is a capital mistake to theorize before one has data. Insensibly, one begins to twist facts to suit theories, instead of theories to suit facts."*

A Scandal in Bohemia (1891)
Arthur Conan Doyle

Data Mining is fundamentally about delivering novel and actionable knowledge from mountains of data.

An Australian Journey

- Data Mining Research - CSIRO 1995
- Data Mining Practise - Health Insurance Commission 1995
- A Taste of Data Mining:
  - Esanda Finance
  - NRMA
  - Mount Stromlo
  - Health Insurance Commission
  - Commonwealth Bank
  - Department of Health
  - Australian Taxation Office
  - Australian Customs Service
  - Department of Veteran Affairs
  - . . .

An Australian Journey

- Data Mining Research - CSIRO 1995
- Data Mining Practise - Health Insurance Commission 1995
- A Taste of Data Mining:
  - Esanda Finance
  - NRMA
  - Mount Stromlo
  - Health Insurance Commission
  - Commonwealth Bank
  - Department of Health
  - Australian Taxation Office
  - Australian Customs Service
  - Department of Veteran Affairs
  - . . .

An Australian Journey

- Data Mining Research - CSIRO 1995
- Data Mining Practise - Health Insurance Commission 1995
- A Taste of Data Mining:
  - Esanda Finance
  - NRMA
  - Mount Stromlo
  - Health Insurance Commission
  - Commonwealth Bank
  - Department of Health
  - Australian Taxation Office
  - Australian Customs Service
  - Department of Veteran Affairs
  - . . .

An Australian Journey

- Data Mining Research - CSIRO 1995
- Data Mining Practise - Health Insurance Commission 1995
- A Taste of Data Mining:
    - Esanda Finance
    - NRMA
    - Mount Stromlo
    - Health Insurance Commission
    - Commonwealth Bank
    - Department of Health
    - Australian Taxation Office
    - Australian Customs Service
    - Department of Veteran Affairs
    - ...

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data, for the betterment of society and in the service of society — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data, for the betterment of society and in the service of society — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data, for the betterment of society and in the service of society — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data, for the betterment of society and in the service of society — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data, for the betterment of society and in the service of society — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data, for the betterment of society and in the service of society — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data, for the betterment of society and in the service of society — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data, for the betterment of society and in the service of society — privacy protocols.

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
- PAYG $100B, GST $40B, Excise $20B

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

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
- PAYG $100B, GST $40B, Excise $20B

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

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
- PAYG $100B, GST $40B, Excise $20B

- 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 a national capability in 2003

Team has been built up to 16 data mining specialists

Support 120 analysts throughout the organisation

Spread new technology throughout the whole organisation through a central R&D capability

Provide an over-arching framework for Risk Management

How: Analytics Community of Practise and roll out of Training Course

ATO Analytics - Deploying Data Mining

Established as a national capability in 2003

Team has been built up to 16 data mining specialists

Support 120 analysts throughout the organisation

Spread new technology throughout the whole organisation through a central R&D capability

Provide an over-arching framework for Risk Management

How: Analytics Community of Practise and roll out of Training Course

Overview

Setting the Context

Background

Australian Taxation Office

Tooling up for Data Mining

Technologies

Commodity and Open Source

Delivering Outcomes

Technologies

- Originally tooled up with commercial, expensive, 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 vendors have difficulty keeping up.

Technologies

- Originally tooled up with commercial, expensive, 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 vendors have difficulty keeping up.

Technologies

- Originally tooled up with commercial, expensive, 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 vendors have difficulty keeping up.

New Approaches Ensembles

Commercial software is lagging behind advances in Data Mining

- Current best off the shelf technology includes random forests, boosting and support vector machines - SAS/EM?
- Open source solutions allow investment in people, not software.

New Approaches Ensembles

Commercial software is lagging behind advances in Data Mining

- Current best off the shelf technology includes random forests, boosting and support vector machines - SAS/EM?
- Open source solutions allow investment in people, not software.

New Approaches Ensembles

Commercial software is lagging behind advances in Data Mining

- Current best off the shelf technology includes random forests, boosting and support vector machines - SAS/EM?
- Open source solutions allow investment in people, not software.

Hardware Platform - 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 (Optimal)
- 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

Data Warehouse (Netezza/SQLite) as the workhorse data server

Hardware Platform - 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 (Optimal)
- 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

Data Warehouse (Netezza/SQLite) as the workhorse data server

Hardware Platform - 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 (Optimal)
- 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

Data Warehouse (Netezza/SQLite) as the workhorse data server

Hardware Platform - 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 (Optimal)
- 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

Data Warehouse (Netezza/SQLite) as the workhorse data server

Hardware Platform - 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 (Optimal)
- 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

Data Warehouse (Netezza/SQLite) as the workhorse data server

Overview

Setting the Context

- Background
- Australian Taxation Office

Tooling up for Data Mining

- Technologies
- Commodity and Open Source

Delivering Outcomes

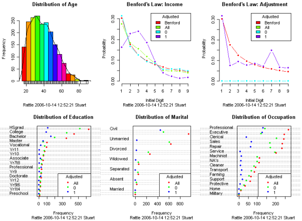
Rattle

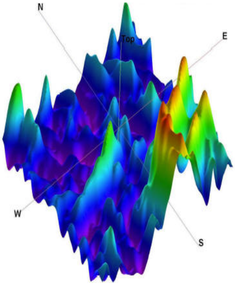
Invest in expertise — tools follow.

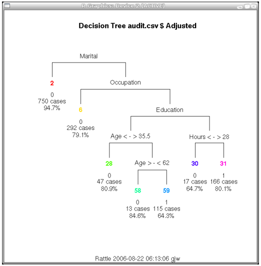
Free software for data mining based on R
+ Weka, AlphaMiner, KNIME, RapidMiner, . . .

Exploratory Data Analysis + Mining: R is second to none

Importance of effectively communicating results.

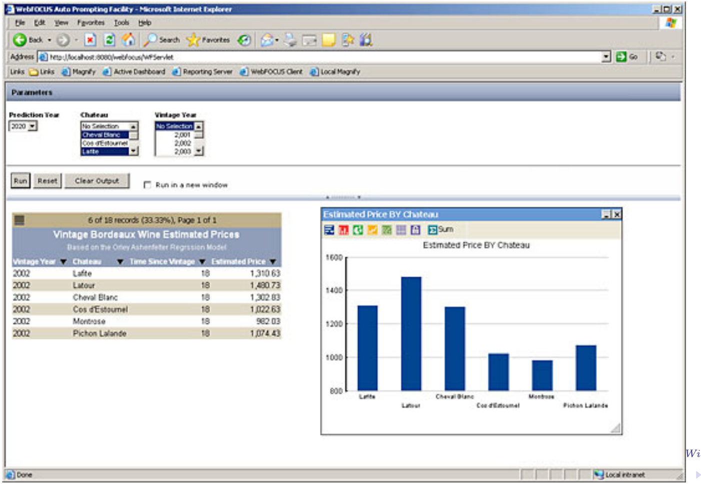






Business Intelligence and Data Mining

- Press Release 2 Jun 2008 from Information Builders (BI Tool — WebFOCUS)
- Announced partnership to incorporate open source Rattle (as RStat) into WebFOCUS.


WebFOCUS Auto Prompting Facility - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Go
Links Links Magnify Active Dashboard Reporting Server WebFOCUS Client Local Magnify
Run Reset Clear Output
Estimated Price BY Chateau
Sum
Estimated Price BY Chateau
1600
1400
1200
1000
800
Lafite
Latour
Cheval Blanc
Cos d'Estournel
Montrose
Pichon Lalande
Done
Local intranet

Address http://localhost:8080/webfocus/WFServlet

Parameters

Prediction Year: 2020

Chateau: No Selection, Cheval Blanc, Cos d'Estournel, Lafite

Vintage Year: No Selection, 2,001, 2,002, 2,003

☐ Run in a new window

6 of 18 records (33.33%), Page 1 of 1

Vintage Bordeaux Wine Estimated Prices

Based on the Orley Ashenfelter Regression Model

Vintage Year	Chateau	Time Since Vintage	Estimated Price
2002	Lafite	18	1,310.63
2002	Latour	18	1,480.73
2002	Cheval Blanc	18	1,302.83
2002	Cos d'Estournel	18	1,022.63
2002	Montrose	18	982.03
2002	Pichon Lalande	18	1,074.43

Analytics in Action

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

- Current rules identify too many “high risk” refunds.
- Some tests might identify 100,000 cases each year.
- Sometimes as few as 5% are found to require adjustment.
- Revenue at risk can be very significant (from \$10m to \$1b).

Data Mining modelling for HRR.

- Has identified numerous characteristics to better target risk (5%)
- More effectively deploy resources on productive cases.
- Uses decision trees and ensembles (random forests).

Analytics in Action

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

- Current rules identify too many "high risk" refunds.
- Some tests might identify 100,000 cases each year.
- Sometimes as few as 5% are found to require adjustment.
- Revenue at risk can be very significant (from $10m to $1b).

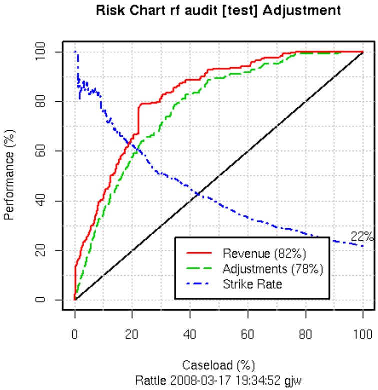
Data Mining modelling for HRR.

- Has identified numerous characteristics to better target risk (5%)
- More effectively deploy resources on productive cases.
- Uses decision trees and ensembles (random forests).

Communicating Outcomes

Complex black box models or explainable insights for intelligence
ROC Versus Risk Charts

- Sort cases by the risk score
- Review from the top of the list
- Trade off caseload against performance
- 40% reduction in effort with little impact.



Other Areas of Modelling

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

Deploying Data Mining

Placing Data Mining Models into Production — Difficulties

- Much data mining is **not** deployed!
- Mostly ad-hoc model runs for case selection using original platform.
- How best to deploy into production?
  - As SQL — 2 million lines (20x200x500)
  - As PMML — interoperability (new engines)
  - As C — DWH (Netezza) 15M entities in 90 seconds

Deploying Data Mining

Placing Data Mining Models into Production — Difficulties

- Much data mining is **not** deployed!
- Mostly ad-hoc model runs for case selection using original platform.
- How best to deploy into production?
  - As SQL — 2 million lines (20x200x500)
  - As PMML — interoperability (new engines)
  - As C — DWH (Netezza) 15M entities in 90 seconds

Deploying Data Mining

Placing Data Mining Models into Production — Difficulties

- Much data mining is **not** deployed!
- Mostly ad-hoc model runs for case selection using original platform.
- How best to deploy into production?
  - As SQL — 2 million lines (20x200x500)
  - As PMML — interoperability (new engines)
  - As C — DWH (Netezza) 15M entities in 90 seconds

Deploying Data Mining

Placing Data Mining Models into Production — Difficulties

- Much data mining is **not** deployed!
- Mostly ad-hoc model runs for case selection using original platform.
- How best to deploy into production?
  - As SQL — 2 million lines (20×200×500)
  - As PMML — interoperability (new engines)
  - As C — DWH (Netezza) 15M entities in 90 seconds

Deploying Data Mining

Placing Data Mining Models into Production — Difficulties

- Much data mining is **not** deployed!
- Mostly ad-hoc model runs for case selection using original platform.
- How best to deploy into production?
  - As SQL — 2 million lines (20×200×500)
  - As PMML — interoperability (new engines)
  - As C — DWH (Netezza) 15M entities in 90 seconds

Deploying Data Mining

Placing Data Mining Models into Production — Difficulties

- Much data mining is **not** deployed!
- Mostly ad-hoc model runs for case selection using original platform.
- How best to deploy into production?
  - As SQL — 2 million lines (20×200×500)
  - As PMML — interoperability (new engines)
  - As C — DWH (Netezza) 15M entities in 90 seconds

Deploying Data Mining

Placing Data Mining Models into Production — Difficulties

- Much data mining is **not** deployed!
- Mostly ad-hoc model runs for case selection using original platform.
- How best to deploy into production?
  - As SQL — 2 million lines (20x200x500)
  - As PMML — interoperability (new engines)
  - As C — DWH (Netezza) 15M entities in 90 seconds

Demonstrating Rattle

A stepping stone into R
or
A self contained tool for data mining

1. Start Rattle
2. Explore the interface
3. Load sample audit dataset
4. Explore the data: Summary, Plots, GGobi, Correlations
5. Transform the data: Rescale, Impute, Remap
6. Cluster, Associate
7. Predictive Model
8. Evaluate and Score
9. Log

Resources

- Togaware
  `http://datamining.togaware.com`

- Tools:
  - rattle.togaware.com
  - www.cs.waikato.ac.nz/ml/weka/
  - www.knime.org
  - rapid-i.com