Data based decision making in retail banking

Gordon Blunt

Gordon Blunt Analytics Ltd

LTCC
17th February 2010

Outline

Outline

My background

Education and professional

- BA Mathematics
- PhD Data mining
- Fellow of the Institute of Mathematics & its Applications

Work - ‘client side’

- Fast moving consumer goods
- Royal Mail
- Barclaycard

Work - consultancy

- CACI Ltd
- GfK NOP Ltd
- Gordon Blunt Analytics Ltd

Type of work

Client side

- almost always in, or around, marketeers, so rarely amongst other statisticians or mathematicians
- always about data analysis in one form or another, the majority either exploratory data analysis, or in helping the non-numerate understand the implications of what the data can tell us

Consultancy

- similar to work on the client side, but typically a broader range of projects, with less of an ‘end to end’ view
- often brought in to satisfy short term resource issues
- in recent years, mostly in retail financial services

Issues in retail financial services

Lending

- banks became less willing to lend than they had been
- to consumers, businesses, or each other

Cost of funds

- inter-bank lending became expensive in 2007
  - reverted to the norm in Q3 2009?

Risk

- more concern with bad debt than a few years ago

Margin

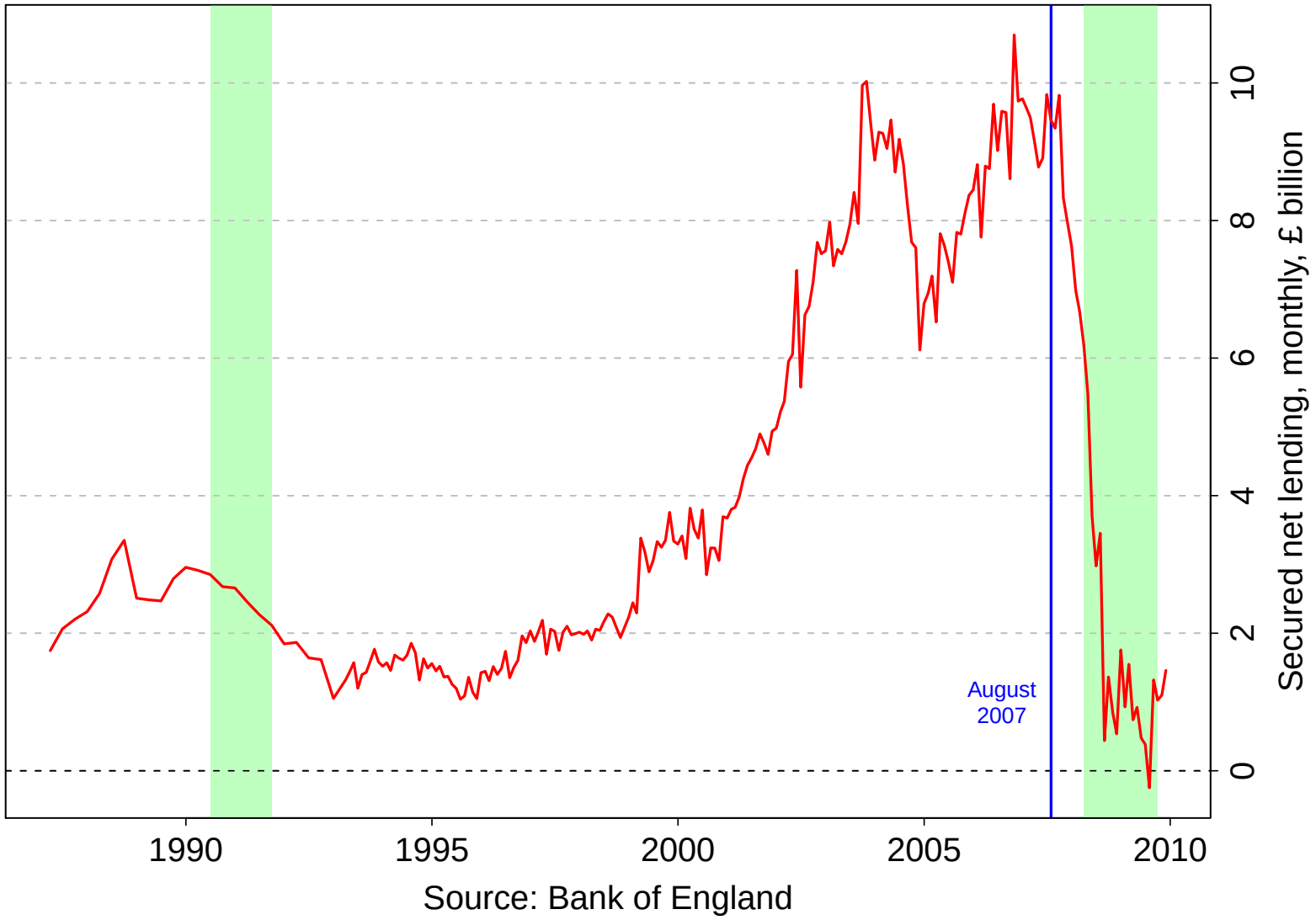
- official bank rate (or *base rate*) is lower than ever
  - harder to produce profitable products / services

Regulation / legislation

In no particular order

- Consumer Credit Directive
- Treating Customers Fairly
- Credit Card Default charges
- Payment Services Directive
- Current Account Market Study
- Credit & Store Card Consultation
- Product Sales Data
- Interchange investigation
- Payment Protection Insurance investigation
- Banking Conduct of Business
- Basel II and Capital Adequacy Directive
- ...

Secured lending



Changes in the environment

I chose one of the more dramatic changes to illustrate the impact on the sector, although I could have shown others

Lending had been the main source of profit for the sector during the decade 2000 - 2009

August 2007 was the month before the first signs of trouble - Northern Rock had its first problems in September that year

Bank of England

- Base rate is lower than it has ever been
  . . . in more than 300 years
- £200 billion pumped into the economy
  (AKA ‘quantitative easing’)
- ‘NICE’ decade is over
  (**N**on **I**nflationary **C**onsistent **E**xpansion)

Outline

Data driven insight

Data driven insight is crucial

- there will be more and more data
- using data wisely will result in better decisions
- the right skills to undertake analysis will be essential
- businesses do not want the perfect (usually unattainable)
- something that works ***now***, not an unknown future date

Analysis projects

- 3 aspects - data, analysis and communication [Chambers 2008]
- must use the latest technology to communicate our findings
  - the internet
  - dynamic, interactive graphics
  - real time data updating - models and data
- if there are delays, opportunities might be missed

Statisticians in business (more correctly, the numerate)

Communication

- this is the most important aspect in business
- will be even more important in future
- as models and data sets become more and more complex
- therefore less understandable to the lay person?
- and with the democritisation of computing power and data

Who will provide the insight?

- engineers and medics may be as useful as statisticians
- they produce useable solutions
- they are more numerous
- may be less concerned with theoretical niceties
- particularly where the data are messy

Influencing skills

Business

- profit is the imperative, not publication
- it can be easy to take decisions based on hunch
- people generally want a quick answer
- with no uncertainty [sic]
- and in days rather than weeks or months

Academia

- enjoyment of the intellectual challenge for its own sake
- getting things ‘right’ more important than a swift answer
- will consider novel approaches
- academic timescales may revolve around PhD projects
- laughably long for most businesses

Persuasion

My approach

- present results face to face
- **ALWAYS** - insist if necessary
- follow up with written report
- will be different from a PhD thesis or academic paper
- use the structure described by David Hand [Hand 2010]

Some of the skills you'll need

- be a good, confident, presenter
- write clear, concise documents
- engage in debate
- involve the client at all stages (AKA networking)
- *be positive!*

The importance of language - an example

Hypothesis testing[1]

H_0 = null hypothesis, H_1 = alternative hypothesis

The final conclusion once the test has been carried out is always given in terms of the null hypothesis

We either 'reject H_0 in favour of H_1' or 'do not reject H_0', we never conclude 'reject H_1', or even 'accept H_1'

Use language carefully

- avoid phrases such as 'cannot do that'
- offer a good alternative before saying 'no'
- be positive at all times

[1] http://www.stats.gla.ac.uk/steps/glossary/hypothesis_testing.html#h1

Language . . . part II

Statistical language that might confuse

- normal
- error
  - as in ‘standard error’ or ‘Type I error’
- significance
- variance
- average
  - possibly meaningless (largely) for highly skewed distributions

Don’t necessarily avoid statistical terms, but . . .

- know your client - whether internal or external
- if from a different field, choose your language carefully
- use plain English if possible

The analysis challenge

Business requirements

- remember profit
- little interest in developing new methods - *unless they work*!
- speed to implement may be critical
- outcome may need to be explicable
- mandatory in the case of credit scoring

Issues

- models built on historic data, applied to future data
- models that evolve in real time?
- fusion of disparate data sources?
- expert systems developed for non-experts?
- statisticians' relationship with their clients

Changing client requirements

Be flexible

- a year may be a long time in business
- be prepared for your project to . . .
  - be changed
  - be cancelled
  - become higher / lower priority
- it will happen regularly

Be firm

- clients may try to make decisions not based on data
  - based on other knowledge, occasionally belief
- always offer something based on your knowledge
- this is too important not to use
- even if it takes you out of your comfort zone

Outline

Exploratory data analysis (EDA)

EDA is still essential

- we owe thanks to John Tukey for his classic 1977 book [Tukey 1977]
- there will be structures in the data of which we are unaware
- we need to find them
- distinguish between 'real' and process generated features

Visualisation is essential

- visualisation is 'a necessary part of data analysis' [Cleveland 1993]
- even more important given the size of modern data sets
- and the fact that we need to use a computer to examine them
- has been an active area of research in recent years

In many ways, data mining is EDA on large data sets

‘End user’ EDA

Analysis vs business needs

- data analysis needs ‘iteration and experimentation’ [Cleveland 1984]
- often conflicts with business need for instant results
- we must develop our skills in client handling and negotiation
- the ‘quick & dirty’ will often work as well as the complex

Democritisation of data and analysis

- how do we ensure non-experts use appropriate methods?
- are able to visualise and model data appropriately
- so that they can learn from their own data
- there is more to life than bar and pie charts!
- we must work with our non-statistician colleagues

Nature of the data

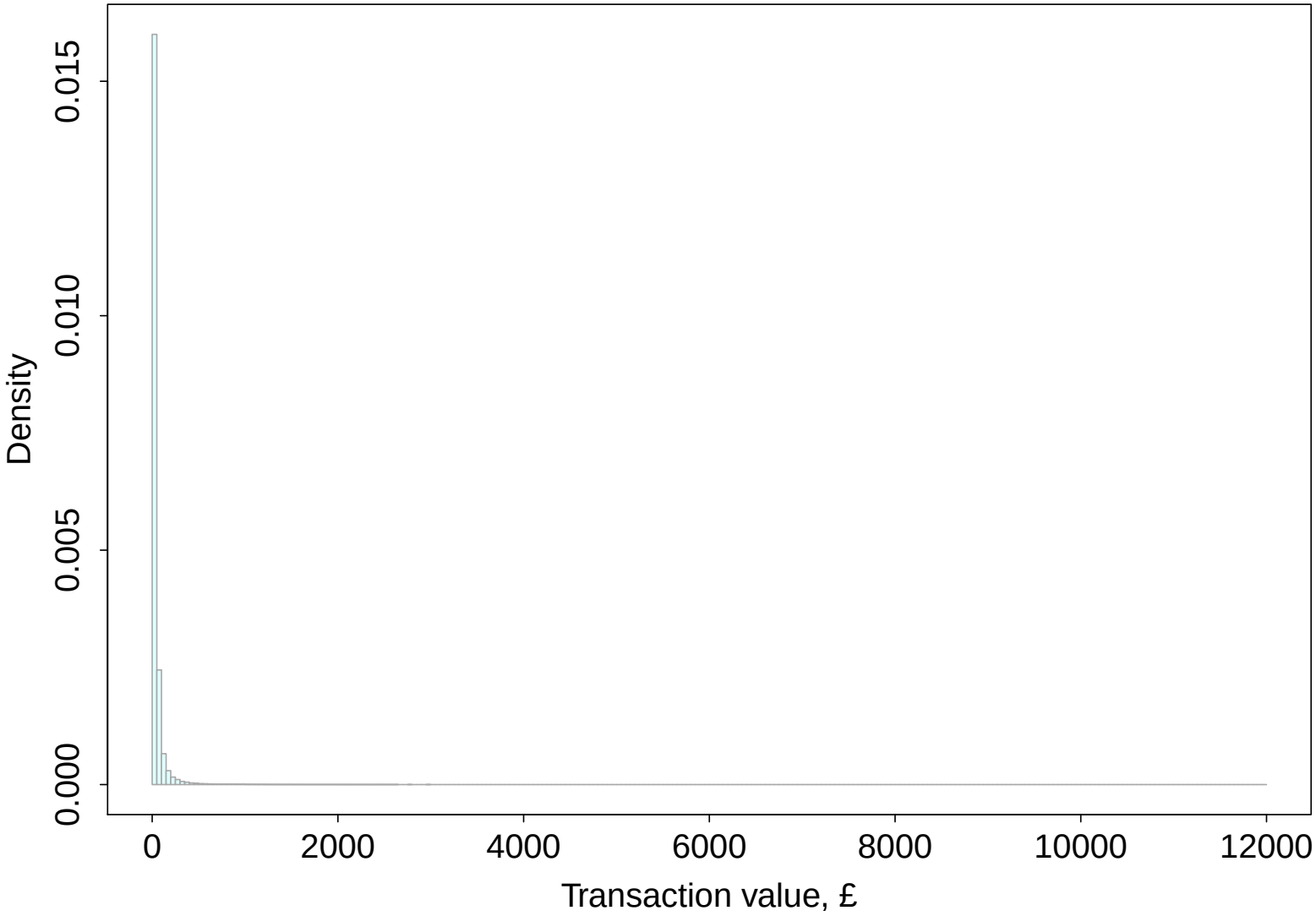
Opportunistic

- data are often a by-product of operational processes
- they are not drawn from a properly constructed sample
- probably contain missing or incorrect fields

Large data sets

- millions of cases, thousands of variables
  (at an extreme Yahoo's 25 terabytes ***every day***) [Fayyad 2009]
- problems with use of 'standard' statistical techniques
- often will not fit 'standard' distributions
- any statistical test likely to prove significant

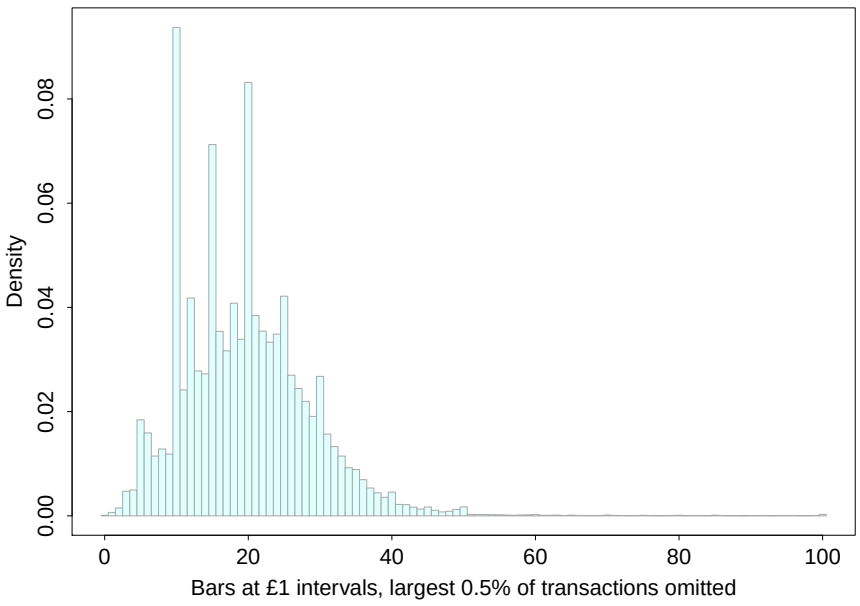
Credit card sales transactions



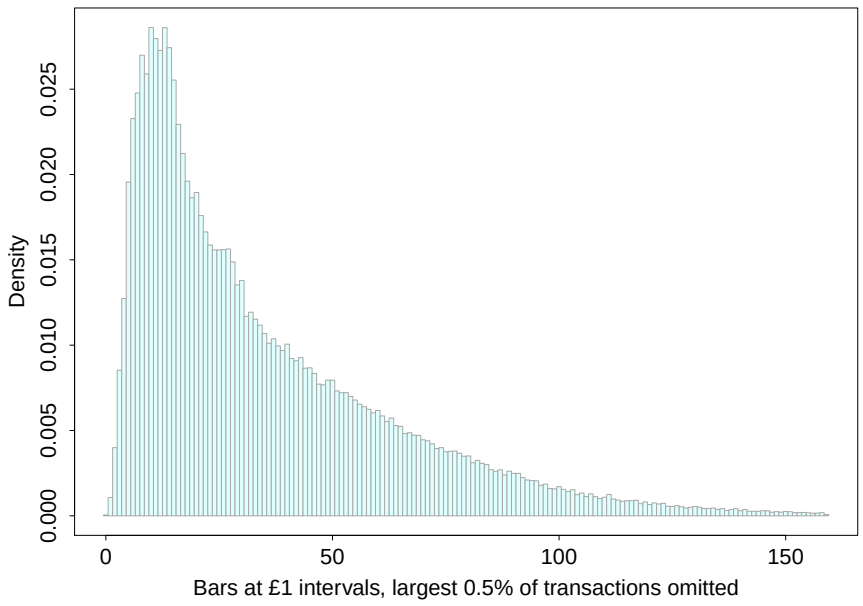
Data first reported in, with more detail, *Prospecting for gems in credit card data* [Hand and Blunt 2001]

Credit card transactions in two sectors

Petrol stations



Supermarkets



Spikes at multiples of £5
and of £6 too . . .

A much smoother distribution
but there are some small peaks

Both of these sectors have frequent, relatively low value, transactions

Data quality

Some issues

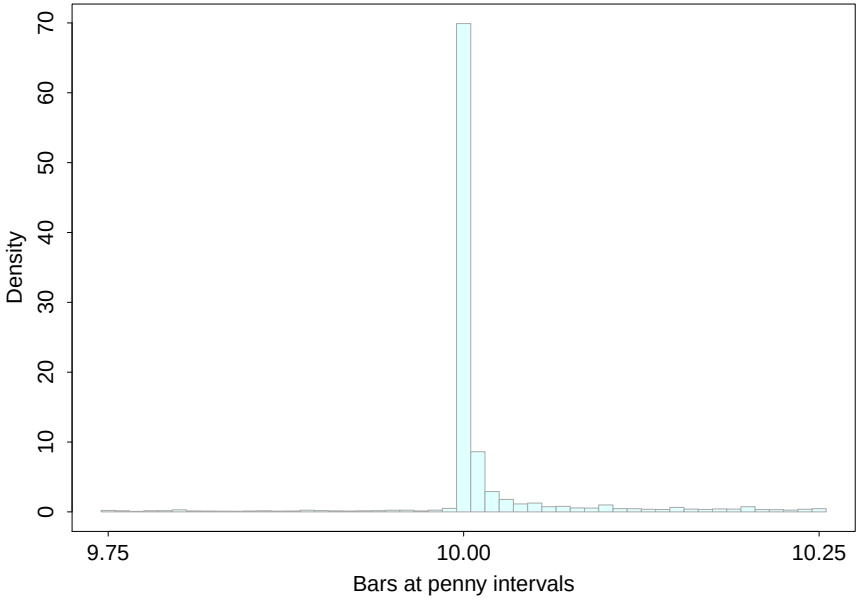
- spikes in account balances at £1, £5, £10 . . .
- customers with more than one ‘unique record number’
- overdrawn savings accounts
- interest charged on credit balances
- mortgage customers who are 4 years old

Data quality

- distrust a ‘clean’ dataset (unless you’ve done the cleaning!)
- cleaning the data can take 80% - 90% of a project’s time
- automatic fault removal may remove real features
- distinguishing ‘real’ from systemic patterns may not be trivial

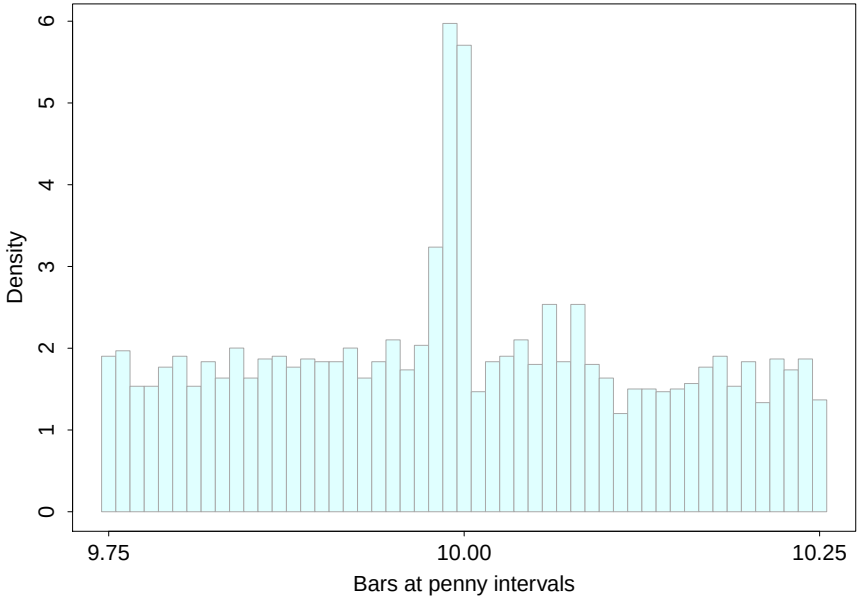
Local patterns - around £10

Petrol stations



Peak at *exactly* £10, decay to right
few transactions below £10

Supermarkets



Peak now at £9.99
slightly smaller one at £10

Models must take account of these structures, of course

Outline

Three broad areas

Operational

- manufacturing
- customer databases
- stock management etc etc . . .

Statistical - using all the power now available to us

- credit risk
- marketing
- empirical finance etc etc . . .

‘Business desktop’ - the majority of people today?

- every other analysis?
- Microsoft® dominates, particularly Excel® and PowerPoint®
- ***other software is better suited to modern statistical graphics and techniques***

Systems constraints

Operational

- the need for ‘regression testing’
- new systems must work from day 1
- millions of customers, hundreds of millions of transactions
- too big to allow to fail

Legacy systems

- some systems may be based on those written 30 years ago
- the origins may be lost in the mists of time somewhere . . .
- anything new must work with them
- and ideally what comes in the future

Software . . . ermm, well . . .

The choice might be fairly limited

SAS® is the *de facto* standard in retail financial services, trying to introduce other software may be difficult

My preferred option would be to use *R*, but most of my clients do not allow executable files to be downloaded and installed

Outline

Careers

Opportunities

- marketing, credit risk, fraud etc etc
  - and, of course, not restricted to retail banking

Good things

- large, interesting, data sets
  - possibly only one step removed from consumers' behaviour
- huge potential for the technically able communicator
- wide variety of techniques and advice you can give

Not so good things

- less freedom than academia
  - may not be able to publish
- may be tedious at times . . . 'not another b***** scorecard!!'
- may be given tasks where we have no skills

Outline

Final thoughts - how to succeed

With the presence of a computer and data on most desktops, we risk losing involvement with much data analysis

We may also have to modify our culture. Any statistician who has worked in other data related fields is struck by their "culture gap" with statistics.

[Friedman 1997]

We must engage with non-statistician colleagues, to avoid some of the communication problems David Hand described in his RSS Presidential Address

[Hand 2009]

Opportunities . . .

- for us to improve the quality of data based decisions
- to have access to a wide variety of interesting data sets
- **communicate, communicate, communicate**

Outline

References

Chambers JM.
Comment on The Future of Statistical Computing.
Technometrics, **50** 4:435-437, 2008.

Cleveland WS.
The Elements of Graphing Data.
Wadsworth, 1984.

Cleveland WS.
Visualizing Data.
Hobart Press, 1993.

Fayyad U.
Evolution of web search, social networking and online marketing. Towards inventing new sciences of the internet.
Imperial College, 2009.

Friedman J.
Data Mining and Statistics: What's the connection?
Proc. 29th Symposium on the Interface Between Computing Science and Statistics, 3-9, 1997.

Hand DJ.
Statistical Consultancy
LTCC course, Institute for Mathematical Sciences, 2010.

Hand DJ.
Modern statistics: the myth and the magic?
J R Statist Soc , **172**:287-306, 2009.

Hand DJ, Blunt G.
Prospecting for gems in credit card data.
IMA Journal of Management Mathematics, **12**:173-200, 2001.

Tukey JW.
Exploratory Data Analysis.
Addison Wesley, 1977.

Wilkinson L.
The Future of Statistical Computing.
Technometrics, **50** 4:418-435, 2008.

Suggested reading

Chen C, Härdle W and Unwin A. (Eds.)
Handbook of Data Visualization.
Springer, 2008.

Gentle JE.
Computational Statistics.
Springer, 2009.

Gentle JE, Härdle W, Mori Y.
Handbook of Computational Statistics.
Springer, 2004.

Hand DJ, Mannila H, Smyth P.
Principles of Data Mining.
MIT Press, 2001.

Hastie T, Tibshirani R, Friedman J.
The Elements of Statistical Learning: Data Mining, Inference and Prediction.
Springer, 2001.

Unwin A, Theus M, Hofmann H.
Graphics of Large Datasets: Visualizing a Million.
Springer, 2006.

Young FW, Valero-Mora, PM, Friendly M.
Visual Statistics: Seeing Data with Dynamic Interactive Graphics.
Wiley, 2006.

Box GEP.
Improving Almost Anything.
Wiley, 2006.

Few S.
Show Me the Numbers: Designing Tables and Graphs to Enlighten.
Analytics Press, 2004.

Hoerl R and Snee R.
Statistical thinking: improving business performance.
Duxbury, 2002.

Robbins NB.
Creating More Effective Graphs.
Wiley, 2005.

Tufte ER.
The Visual Display of Quantitative Information.
Graphics Press, Connecticut, 2001.

Tufte ER.
Beautiful Evidence.
Graphics Press, Connecticut, 2006.

van Belle G.
Statistical Rules of Thumb (2nd ed).
Wiley, New York, 2008.

The books on the left are technical, the ones on the right will be useful in influencing colleagues and clients