Credit Scoring and Data Mining

Ross Gayler
What is credit scoring?
What is credit scoring?

• Predictive modelling of operational outcomes in mass-market credit
• Used to automate operational decision making
  – Make decisions on the basis of predicted outcomes
Why use credit scoring?

• Better decision making
  – More accurate prediction of outcomes
• Automation
  – Cost saving
  – Faster service
• Objectivity
• Repeatability
• Controllability
Potential value of scoring

Consider credit card application processing for a large Australian lender (order of magnitude estimate)

- **Total new exposure**
  - 1k applications / day
  - 70% approval
  - $5k average credit limit
  - ~ $900M p.a. new exposure

- **Total credit loss**
  - 2.5% loss rate
  - ~ $23M p.a. loss

- **Value of improvement**
  - 1% decrease in loss ~ $230k p.a.
History of credit scoring
The first scorecards

• First used in 1946
• Lightly used in the 1950s and early 1960s
• Took off with credit cards and computing in the late 1960s and early 1970s
• First applied to application processing
  – Predictors from application form and credit bureau
  – Predicted outcome is failure to repay
  – Decision is accept/reject the application
Constraints on early scorecards

- Minimal computing power available to build models
  - Simple modelling techniques
    - Linear regression or discriminant analysis

- Models applied manually
  - Simple form of models (scorecards)
### Example application scorecard

<table>
<thead>
<tr>
<th>Time at Job (yrs)</th>
<th>&lt; 0.5</th>
<th>0.5 to 1.4</th>
<th>1.5 to 6.4</th>
<th>6.5 to 10.5</th>
<th>&gt; 10.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>14</td>
<td>20</td>
<td>27</td>
<td>39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time at Address (yrs)</th>
<th>&lt; 1</th>
<th>1 to 2.4</th>
<th>2.5 to 6</th>
<th>&gt; 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-11</td>
<td>1</td>
<td>8</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Residential Status</th>
<th>Own or Buy</th>
<th>Rent</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40</td>
<td>19</td>
<td>26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Retired</th>
<th>Professional</th>
<th>Clerical</th>
<th>Sales</th>
<th>Service</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>41</td>
<td>36</td>
<td>27</td>
<td>18</td>
<td>12</td>
<td>27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age of Applicant (yrs)</th>
<th>18 to 25</th>
<th>26 to 31</th>
<th>32 to 34</th>
<th>35 to 51</th>
<th>52 to 61</th>
<th>&gt; 61</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
<td>14</td>
<td>22</td>
<td>26</td>
<td>34</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bureau Enquiries</th>
<th>No record</th>
<th>1 to 2</th>
<th>3 to 5</th>
<th>&gt; 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-12</td>
<td>1</td>
<td>-7</td>
<td>-32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bureau Defaults</th>
<th>No record</th>
<th>0</th>
<th>1</th>
<th>&gt;1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>-57</td>
<td>-126</td>
</tr>
</tbody>
</table>

Adapted from E.M. Lewis (1992) “An introduction to credit scoring”
Later scorecards

- Different outcomes
  - Response to mailed offer (marketing)
  - Account closure (account management)
- Different predictors
  - Geodemographics
  - Account transaction history
- Modelling method has not changed much
How well does scoring work?
Relationship between score and outcome
Distributions of application scores

\[ d' = \text{separation (in standard deviations)} \]
How well does scoring work?

• More predictive than subjective estimates for application processing
  – Typically 15% - 25% lower default rate at the same accept rate

• Typical predictive performance
  – Application scoring
    • Extreme odds ratio ~ 50:1
    • AUC ~ 0.75
    • $d'$ ~ 1 sd
  – Behaviour scoring
    • Extreme odds ratio ~ 200:1
    • AUC ~ 0.85
    • $d'$ ~ 1.5 sd
A data mining opportunity?
Argument for data mining

• Credit scoring is simple and modestly predictive

• Why not use:
  – Neural networks
  – Random forests
  – Support vector machines
  – …

• Academic research has concentrated on more sophisticated modelling techniques
  – Rarely adopted in practice
Why credit scoring hasn’t changed much

• Methodological inertia
• Systems inertia
• Lender conservatism
  – Transparency for confidence
  – Regulatory requirements
  – Staff expertise

• Mostly - lack of benefit
Is data mining becoming more predictive?
No improvement over 10 years

Error rates vs paper publication date
Pima Indian data (UCI ML repository)
Hand (2003) Banff Credit Scoring Workshop
Any theoretical benefit of data mining likely to be dominated by other issues
Issues

- Functional form of scorecards
  - Standard regression is not really inflexible
  - Every surface looks flat with enough noise
  - Diminishing returns in incremental models
  - Flat maximum effect is valuable

- Data issues
  - Scoring models are not scientific models
    - Not causal models
    - Population drift & jump
  - Poor quality data (modelling the defects)
  - Arbitrary outcome definitions
Flat maximum effect gives flexibility
What is the flat maximum effect?

- Distribution of goodness of fit in model parameter space
- Many approximately equivalent models
Properties of flat maximum effect

- Arises from
  - Additive form of standard regression
  - Conditional monotone relationship between predictors and outcome

- Benefits
  - Models don’t fail suddenly
  - Models can be cross-applied
  - Models can be selected on a basis other than goodness of fit
Assumptions of predictive modelling
Basic assumptions of predictive modelling

- Similar cases behave similarly
- The future is like the past
- Practically, your data are the past

Caveats

- Available data does not address the (stable) causal mechanisms of customer behaviour
- Important predictors are not available
- Important predictors do not vary in the data
- Data is always out of date
The future is not like the past

- New types of customers
- New systems
- New operational procedures
- New competitors
- …
Time frame of modelling

Model based on data 2 ~ 3 years old

Observation period (12 months)

Applications obtained in this period ...

… to predict outcomes at these points

Outcome period (12 – 24 months)
The truth is in the data?

- Poor data quality
  - The data are from operational systems not designed to support statistical data collection

- Data not representative of the future
  - Introducing a new credit product, so no data available for that specific credit product
  - The world will change after data collection

- Puts a premium on being able to understand the models

- Puts a premium on being able to subjectively modify the models
Operational changes
Portfolio acquisitions

- Small off-trend groups at 5th & 97th percentiles
- Identified as cases from an acquired portfolio
- Exclude them?
- Correct for them?
System changes
System changes

- Encoding of dates was changed
- Location of cut-over point will change over time
- Better to see the issue and deal with it than to lose it in an automated process
Reject inference
Reject inference

- Want to apply the model to all applications
- Have outcome data for accepted applications
- Accepts are a systematically biased sample
- Model will be biased if built on accepts
- Need to infer the outcome for the rejects
- Two broad approaches to reject inference:
  1. Extrapolate from the accepts (no new information)
  2. Use proxy outcome information for all applications
- Uncertainty of inferred outcome may dominate change in model due to improved techniques
Comparison of actual and inferred outcomes

![Graph showing comparison of actual and inferred outcomes. The x-axis represents lender old score, and the y-axis represents P(Lender Good). The graph includes curves for: rejects: Bureau inference, rejects: Lender inference, and accepts. The reject quantiles are indicated.]
Account management actions
Impact of effort on outcome

- Collections outcome probably depends on collections effort
- Collections effort is systematically allocated via collection plans
- Collection plans are systematically allocated based on predictive characteristics
- Modelling the raw outcome may be modelling the effects of past plan allocations

- Aim for “all other things being equal” model
Maximum effort function

\[ \log_b(amt\text{.orig}, 2) \]
Population volatility
Volatility of creditor mix

Debt-collection company

• Debt from new creditors can be loaded at the drop of a hat
• Relationship between characteristics and outcome may vary
• Creditors vary widely in size (and impact on portfolio)

• Try to exclude characteristics that show a volatile relationship
Volatility of complex effects
Volatility of complex effects

• Advanced techniques get increased prediction from modelling complex effects (nonlinearities and interactions)

• Credit scorers believe that unmotivated complex effects are more likely to be volatile

• Do not include complex effects unless there is external evidence for their reality

• 11 data sets from a range of countries, type of credit provider and credit product
Volatility study

• 11 data sets from a range of countries, type of credit provider and credit product
• Three predictor variables (Age, Time in Employment, Time at Address) taken 2 at a time
• Three regression models fitted to each combination of data set and predictor variables
  – LINEAR = additive, no further transformation of predictors
  – NONLINEAR = nonparametric optimal transformation of predictors (GAM)
  – INTERACTIVE = locally weighted regression (loess; span = 0.5, degree = 2)
Cross-testing models

![Box plot showing discrimination difference relative to linear on-sample P(A) vs. conditional on portfolio for different cross-test types: Interactive On, Nonlinear On, Linear Off, Nonlinear Off, Interactive Off.](image)
Implementation pragmatics - granularity
Granularity of score distributions

- More fine-grained scores are easier to control
  - Decision cut-offs can be placed anywhere
- Direct classification output is not controllable
Predictions are inherently probabilistic

- Scoring is **not** a classification problem

\[
P(\text{Bad}) = 10%
\]

- Time at address < 1yr
  - Prior defaults: 80%
  - No prior defaults: 15%

- Time at address > 1yr
  - Prior defaults: 30%
  - No prior defaults: 5%
Implementation pragmatics – resistance to gaming
Application fraud detection

Much hand-crafting, including:

• Emphasise predictors that are harder to fake
  – Gender of applicant at point of sale
  – Bureau inquiries > 12 months ago

• Emphasise predictors that work in the lender’s favour if gamed
  – Applicant income
Comparison of fraud models

- Hand-crafted model
  - Nominal predictor df of 3 models: ~400
  - Effective predictor df: < ~60
- Alternative “standard practice” model
  - Nominal predictor df of 1 model: ~100
- Performance
  - Equivalent predictive power at development
  - Alternative model significantly worse after a few months
  - Crafted model’s predictive power unaltered after a few months
  - Crafted model still in use > 5 years after implementation!
Conclusion
Conclusions

Credit scoring is not just fitting a model to data. There are many reasons why the available data do not represent the future population. Techniques for better fitting the data at hand are unlikely to yield practical benefits.

Principled incorporation of external knowledge
Optimisation for robustness
Action-conditional predictions in the absence of true experimental design
Dealing with causal loops
Possible objectives for data mining

• Principled incorporation of external knowledge
• Optimisation for robustness
• Action-conditional predictions in the absence of true experimental design
• Dealing with causal loops