Class(ic) Scorecards

Selecting Characteristics and Attributes in Logistic Regression
Edinburgh Credit Scoring Conference - 25 August 2011

Gerard Scallan
gerard.scallan@scoreplus.com

Class(ic) Scorecards

Using the Statistics!

- What’s the Problem?
- Nested Dummy Variables
- Stepwise Method
- Selecting Characteristics
- Lessons Learned
**Example: Age Characteristic**  
**Typical Analysis Layout**

**CHARACTERISTIC: AGE**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>SAMPLE COUNTS</th>
<th>COLUMNS</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Goods</td>
<td>Bads</td>
<td>Total</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3608</td>
<td>1018</td>
<td>4645</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>19</td>
<td>22</td>
<td>19</td>
<td>41</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
<td>19</td>
<td>44</td>
</tr>
<tr>
<td>21</td>
<td>24</td>
<td>29</td>
<td>53</td>
</tr>
<tr>
<td>22</td>
<td>26</td>
<td>29</td>
<td>55</td>
</tr>
<tr>
<td>23</td>
<td>32</td>
<td>31</td>
<td>63</td>
</tr>
<tr>
<td>24</td>
<td>34</td>
<td>26</td>
<td>60</td>
</tr>
<tr>
<td>25</td>
<td>44</td>
<td>29</td>
<td>73</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>66+</td>
<td>18</td>
<td>1</td>
<td>19</td>
</tr>
</tbody>
</table>

Information Value: 0.373  
Chi²: 334.61  
DF: 47  
p-level: 5.04938E-45

**Goal of Classing → Maximise predictive power**

**WoE Graph: Show overall picture**

**WoE**  
\[ \text{WoE} = \ln(\text{Odds}(\text{attr})) - \ln(\text{Odds}(\text{popn})) \]

**IV**  
\[ \text{IV} = \text{Avg}_i(\text{WoE}) - \text{Avg}_b(\text{WoE}) \]

**Problem: Testing Wrong Hypothesis**
Current Practice: Classing

Current Practice

- “Fine” breakdowns on each predictive characteristic
- Manual or Automatic Classing
  - Based on Information Value
  - or Chi² measure
- 1 dummy variable per class
- Select model variables using stepwise Logistic Regression

And what’s wrong with it

- One characteristic at a time
  - Anomalies in one characteristic often explained by another
- Lots of predictors → Lots of time
  - 700 chars x 3 mins. = 35 hours
- Variable selection in model at attribute level
  - “gap toothed” models
  - Age 18-21, Age 25-29 in model
  - Age 22-24 not in model
- Stepwise measures certainty
  - Not distance

Good technical solutions – but wrong problem

Solution 1: Continuous Variables

Risk improves continuously with Age

- Simpler Hypothesis
  - 1 parameter vs. 15+
- Data do not contradict the linear hypothesis
  - In most cases
- But sample sliced into many small categories
  - Combine categories
  - → More reliable tests
- Slope changes ~ age 30
  - Again ~ age 50?

Better Starting Point
Why Discretise?

Non-Linearities

 Tradition – 1960s

- Scores calculated by hand
- No pocket calculators
- Multiplication less reliable than addition
- Coefficients – 2 digit integers

- Slope changes ~ age 30
- Again ~ age 50?

Not quite discrete ...

No longer justified

---

Class(ic) Scorecards

Using the Statistics!

- What’s the Problem?
- Nested Dummy Variables
- Stepwise Method
- Selecting Characteristics
- Lessons Learned
Partition Variables

a.k.a. Nested Dummy Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Age 18</th>
<th>Age 19</th>
<th>Age 20</th>
<th>Age 21</th>
<th>Age 22</th>
<th>Age 23</th>
</tr>
</thead>
<tbody>
<tr>
<td>P18</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P19</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P20</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- Partition variable for each fine class
- P18 = intercept – will not enter model
- Score for 22 year old = P18 + P19 + P20 + P21
- Coefficient P22 = incremental change for Age 22 compared to Age 21
- Partition model gives same score to each individual as Attribute model
- Partition and Attribute variables = two bases for same linear space
- Monotone increasing ↔ Partition Coefficients > 0

Different coding – Same model

Variance of Coefficients and Significance Testing

| No. | Characteristic | Variable | Estimate | Std. Error | z-value | P>|z| | Significance | [95% Conf. Interval] |
|-----|----------------|----------|----------|------------|---------|--------|---------------|------------------|
| 0   | (Intercept)    | 0        | 0.54343  | 0.17772    | 3.058   | 0.00223 | ***            | 0.19510 0.89176 |
| 1   | TmBooks       | 2y6m+    | 0.82928  | 0.09972    | 8.316   | < 2e-16 | ***            | 0.63383 1.02473 |
| 2   | TmBooks       | 7y1m+    | 0.68709  | 0.12361    | 5.558   | 2.72E-08 | ***            | 0.44481 0.92937 |
| 3   | TmBooks       | 14y1m+   | 0.56779  | 0.1673     | 3.394   | 0.000689 | ***           | 0.23988 0.89570 |
| 4   | DaysXS        | Any      | -0.68069 | 0.13247    | -5.138  | 2.77E-07 | ***            | -0.94033 -0.42105 |
| 5   | DaysXS        | 11+      | -0.45509 | 0.18657    | -2.439  | 0.01472 | *             | -0.82077 -0.08941 |
| 6   | DaysXS        | 16+      | -0.08821 | 0.17508    | -0.504  | 0.614396 |                | -0.43137 0.25495 |
| 7   | DaysXS        | 61+      | -0.45783 | 0.11057    | -4.141  | 3.47E-05 | ***            | -0.67455 -0.24111 |
| 8   | Bounce        | 1m+      | 0.39119  | 0.13214    | 2.96    | 0.003072 | **             | 0.13220 0.65018 |
| 9   | Bounce        | 41m+     | -0.06494 | 0.28124    | -0.231  | 0.81739 | -              | -0.61617 0.48829 |
| 10  | Bounce        | Never    | 1.02127  | 0.28125    | 3.631   | 0.000282 | **             | 0.47002 1.57252 |
| 11  | AutoCredit    | Any      | 0.41368  | 0.12795    | 3.233   | 0.001225 | **             | 0.16290 0.66446 |
| 12  | AutoCredit    | 4000+    | 0.44995  | 0.11074    | 4.065   | 4.84E-05 | ***            | 0.23290 0.66700 |

- Maximum Likelihood Estimates
- Std. Error from Covariance Matrix of Estimates
- Z-value = Estimate/Std. Error
- OR Wald Statistic = Z²
Z-test and Wald Chi² Test: Is this variable necessary?

Z-test
- Z-value = Estimate/Std. Error
- If “true” value of Coefficient = 0
  - Null Hypothesis
  - then sample value of Z has Normal distribution
    - Mean = 0, Variance = 1
  - (From theory of Max Likelihood)
- If Null Hypothesis is true, then unlikely to get this big |z| OR
- If |z| is “large”, data are not consistent with NH

Wald Chi² Test
- Z² = Estimate²/Variance
- Under Null Hypothesis Z² has Chi² Distribution w/ 1 DF
  - Square of N(0,1)
- Same test!
  - Test at 10%, 5%, 1%, .1%
    - *** p < 0.1%
    - ** p < 1%
    - * p < 5%
    - . p < 10%

Large sample approximation – easy to apply

Hypothesis Tests with Partition Variables

Attribute Dummy Variables
- “Reference Attribute” on every characteristic
  - Receives 0 score
  - Avoids linear indeterminacy
  - Usually last attribute
  - E.g. Age 60+
- Coefficient = 0
  - Risk same as Reference Attribute
- E.g. Risk on Age 22-25 = Risk on Age 60+
- Useless hypothesis

Ignore statistics

Partition Dummy Variables
- Coefficient = 0 ↔ Risk same as neighbour to left
- E.g. No difference in risk between Age 22-25 and Age 20-21
- What are key turning points in risk pattern?

Key information
Automated classing  
Provisional Solution

Algorithm

◆ Partition Vars. for “fine” classes
  ◆ Must be ordered “sensibly”
  ◆ Natural order or WoE
  ◆ Possibly 20-30 variables/characteristic
  ◆ All characteristics in model
◆ Candidates in stepwise Logistic
◆ Stepwise algorithm identifies “significant” breakpoints
  ◆ Partition variable enters iff “significant” difference between neighboring attributes

Advantages

◆ Less work for analyst!
◆ Classing adapts to sample size
  ◆ Small sample → Coarser
  ◆ Large sample → Finer
◆ Accounts for interactions between characteristics
  ◆ Fewer classes/characteristic
  ◆ Multivariate approach
◆ Equivalent to systematic use of Marginal Chi²
  ◆ But approximations are better!
◆ Avoids gap-toothed scorecards

Get minimal classing needed for predictive structure

Continuous Variables  
Piecewise Linear

Idea

◆ Analogous idea for continuous predictors
◆ Family of spline variables
◆ E.g. Age
  ◆ (Age – 20)⁺ = max(0, Age-20)
  ◆ (Age – 22)⁺ = max(0, Age-22)
  ◆ (Age – 24)⁺ = max(0, Age-24)
  ◆ … etc.
◆ Candidates in stepwise Logistic
◆ Terms entering correspond to significant changes in slope
◆ a.k.a. MARS
  ◆ Multivariate Adaptive Regression Splines

Example

Score = .2 x Age  
-.06 x (Age – 22)⁺  
-.04 x (Age – 30)⁺  
-.03 x (Age – 38)⁺  
-.02 x (Age – 46)⁺
Class(ic) Scorecards

Using the Statistics!

✓ What’s the Problem?
✓ Nested Dummy Variables
→ Stepwise Method
✓ Selecting Characteristics
✓ Lessons Learned

Stepwise Approach

3 variants

◆ Forward Selection
  ◆ Start with null model
  ◆ Add variables
  ◆ Until no further variable adds significant predictive power

◆ Backward Elimination
  ◆ Start with all variables
  ◆ Drop variable which makes least contribution to likelihood
  ◆ Until no further variable can be dropped without significant loss of predictive power

◆ Bidirectional
  ◆ Start with null model
  ◆ Add variables
  ◆ At each step, check to see if variables can be dropped
  ◆ Then check to see if any variable can be added
  ◆ Until no variable to be dropped AND
  ◆ No variable to be added

Computation: Forward < Backward < Bidirectional
What’s wrong with Stepwise?

“If this method had just been proposed ... it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing”
– Harrell 2001 “Regression Modeling Strategies”, p. 56

- Parameters estimates too large
  - Selects “overestimated” coefficients
- Overestimates precision
  - Because underestimates variance
- Collinearity makes variable selection arbitrary

“It allows us not to think about the problem”

Stepwise Logistic on Random Numbers
Simulated Example

- Similar to Flom & Cassell (2007)
- 1000 Goods
- Bads from 100 to 1000
- 100 candidate variables
- All “white noise”
  - Random from Normal Distribution
  - Real predictive power = 0
- 100 replications for each sample size
- Entry/Exit criterion: p < 0.1

- Results on estimation sample
- Won’t validate (we hope!)
- All models have Deviance statistics w/ p-level < 0.1%
- 2/3 of variables significant at 5% p-level

Adds noise to model
Class(ic) Scorecards

Using the Statistics!

- What’s the Problem?
- Nested Dummy Variables
- Stepwise Method
- Selecting Characteristics
- Lessons Learned

Goal: Minimal Sufficient Model

- Bring in enough variables to explain the variation in outcome across the sample
- But no more …
- Tell a (sensible) story

End point: predictive power of sample is exhausted
Marginal Information and Delta Scores

- Weight of Evidence (WoE) = \log (\text{Attribute Odds}) – \log (\text{Population Odds})
- One-dimensional score coefficients
- Delta Score = Observed WoE – Expected WoE
- Approximation to score coeffts needed to line up expected with observed
- Marginal Information Value = Avg\text{Good}(\text{Delta Score}) – Avg\text{Bad}(\text{Delta Score})
- Similar to Kullback-Liebler Information Value
- Increased spread between average score of goods and bads
- … if this characteristic brought into model

### Debit Turnover

<table>
<thead>
<tr>
<th></th>
<th>OBSERVED</th>
<th></th>
<th>EXPECTED</th>
<th></th>
<th>(\Delta)-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goods</td>
<td>436</td>
<td>174</td>
<td>-1.17</td>
<td>487.7</td>
<td>-0.70</td>
</tr>
<tr>
<td>Goods</td>
<td>178</td>
<td>38</td>
<td>-0.54</td>
<td>184.6</td>
<td>-0.32</td>
</tr>
<tr>
<td>Goods</td>
<td>84</td>
<td>17</td>
<td>-0.49</td>
<td>86.2</td>
<td>-0.33</td>
</tr>
<tr>
<td>Goods</td>
<td>263</td>
<td>46</td>
<td>-0.34</td>
<td>263.1</td>
<td>-0.34</td>
</tr>
<tr>
<td>Goods</td>
<td>6240</td>
<td>618</td>
<td>0.22</td>
<td>6179.4</td>
<td>0.12</td>
</tr>
<tr>
<td>Goods</td>
<td>7201</td>
<td>893</td>
<td>0.00</td>
<td>7201</td>
<td>0.00</td>
</tr>
<tr>
<td>Bads</td>
<td>192</td>
<td>38</td>
<td>-0.54</td>
<td>184.6</td>
<td>-0.32</td>
</tr>
<tr>
<td>Bads</td>
<td>178</td>
<td>38</td>
<td>-0.54</td>
<td>184.6</td>
<td>-0.32</td>
</tr>
<tr>
<td>Bads</td>
<td>84</td>
<td>17</td>
<td>-0.49</td>
<td>86.2</td>
<td>-0.33</td>
</tr>
<tr>
<td>Bads</td>
<td>263</td>
<td>46</td>
<td>-0.34</td>
<td>263.1</td>
<td>-0.34</td>
</tr>
<tr>
<td>Bads</td>
<td>6240</td>
<td>618</td>
<td>0.22</td>
<td>6179.4</td>
<td>0.12</td>
</tr>
<tr>
<td>Bads</td>
<td>7201</td>
<td>893</td>
<td>0.00</td>
<td>7201</td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7201</td>
<td>893</td>
<td>0.00</td>
<td>7201</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Chi² = 33.06**  
**D.F. = 4**  
**p-value = 0.00012%**

### Selecting Scorecard Characteristics

- Rank characteristics by Marginal IV
- Characteristic with maximum MIV enters model ...
- … i.e. partition variables become candidates for entry to model

### Scores

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>DaysXsL6m</th>
<th>ToB</th>
<th>SinceDish</th>
<th>AutoCr</th>
<th>CurDaysXs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CurBal</td>
<td>0.032</td>
<td>0.19</td>
<td>0.017</td>
<td>0.013</td>
<td>0.010</td>
</tr>
<tr>
<td>CurCTO</td>
<td>0.185</td>
<td>0.12</td>
<td>0.086</td>
<td>0.089</td>
<td>0.007</td>
</tr>
<tr>
<td>CurDaysXs</td>
<td>0.616</td>
<td>0.12</td>
<td>0.110</td>
<td>0.093</td>
<td>0.060</td>
</tr>
<tr>
<td>CurDTO</td>
<td>0.215</td>
<td>0.17</td>
<td>0.087</td>
<td>0.093</td>
<td>0.026</td>
</tr>
<tr>
<td>CurValXs</td>
<td>0.515</td>
<td>0.12</td>
<td>0.110</td>
<td>0.093</td>
<td>0.060</td>
</tr>
<tr>
<td>ToB</td>
<td>0.692</td>
<td>0.526</td>
<td>0.010</td>
<td>0.026</td>
<td>0.025</td>
</tr>
<tr>
<td>MthsInact</td>
<td>0.012</td>
<td>0.005</td>
<td>0.001</td>
<td>0.004</td>
<td>-0.002</td>
</tr>
<tr>
<td>MthsNoCTO</td>
<td>0.077</td>
<td>0.066</td>
<td>0.043</td>
<td>0.045</td>
<td>0.001</td>
</tr>
<tr>
<td>NetTO</td>
<td>0.074</td>
<td>0.028</td>
<td>0.007</td>
<td>0.010</td>
<td>0.002</td>
</tr>
<tr>
<td>DaysDbL3m</td>
<td>0.055</td>
<td>0.008</td>
<td>0.013</td>
<td>0.008</td>
<td>0.005</td>
</tr>
<tr>
<td>DaysXsL6m</td>
<td>0.856</td>
<td>0.000</td>
<td>0.008</td>
<td>0.011</td>
<td>0.015</td>
</tr>
<tr>
<td>CurMxBal</td>
<td>0.033</td>
<td>0.015</td>
<td>0.018</td>
<td>0.013</td>
<td>0.005</td>
</tr>
<tr>
<td>DisHL3m</td>
<td>0.291</td>
<td>0.090</td>
<td>0.084</td>
<td>-0.006</td>
<td>-0.008</td>
</tr>
<tr>
<td>SinceDish</td>
<td>0.810</td>
<td>0.397</td>
<td>0.299</td>
<td>0.057</td>
<td>0.050</td>
</tr>
<tr>
<td>InterCTO</td>
<td>0.017</td>
<td>0.004</td>
<td>-0.003</td>
<td>-0.004</td>
<td>-0.001</td>
</tr>
<tr>
<td>InterDTO</td>
<td>0.003</td>
<td>0.001</td>
<td>0.000</td>
<td>-0.002</td>
<td>-0.002</td>
</tr>
<tr>
<td>AutoCr</td>
<td>0.209</td>
<td>0.143</td>
<td>0.108</td>
<td>0.106</td>
<td>0.005</td>
</tr>
<tr>
<td>ValDishL6m</td>
<td>0.468</td>
<td>0.145</td>
<td>0.137</td>
<td>-0.001</td>
<td>-0.001</td>
</tr>
</tbody>
</table>
Marginal IV and Collinearity

- As each variable enters MIV on remaining characteristics reduces
- Reduction measures collinearity
  - “overlap” in predictive power
  - Improperly called “correlation”
- Understand relationships between characteristics through MIV decay
- Frequently identify “families”
  - Or “Factors”
  - If one member enters model,
  - MIV drops severely on other members
- Choice of member is arbitrary

Zero Marginal Information = Sufficient Statistic

Automated Classing with Marginal IV
Customer Age Example

- Compute Marginal Info Value for each partition
- Select partition with max. MIV
- Check Significance $\rightarrow$ Deviance Test
- Rebuild model w/ new variable
- Re-estimate MIVs
- Continue until no significant MIV left
- All characteristics processed simultaneously
Automated Classing with Marginal IV

Customer Age Example - Completion

- Continue until all MIVs < .020
- 5 variables – 6 classes
- -ve MIVs → Wrong direction
- In real life, do all chars simultaneously

End of process: “Zero” Marginal Information

Actual vs. Fitted WoE

- “Few” significant differences between fitted and actual
- Differences in neighbouring groups all significant at 95%
Triple Test
Bottom Line

- Marginal Information Value = Importance
  - Distance measure
  - Rule of Thumb: -.020 < MIV > +.020
  - Negative value indicates over-fitting
  - Re-examine history of MIV to drop variable from model
- Marginal Chi² = Reliability
  - Measure of certainty
  - Thousands of tests - beware of false positives
  - Sensitive to classing used for analysis
  - More robust to use Stepwise approach for classing
- Business sense = Coherence
  - Does characteristic tell a believable story?
  - Does the model make sense

Model complete when no further variable satisfies these 3 criteria

Class(ic) Scorecards
Using the Statistics!

- What’s the Problem?
- Nested Dummy Variables
- Stepwise Method
- Selecting Characteristics
- Lessons Learned
Conclusions

- Standard statistical tools can be used better
  - Corollary: We don’t need lots of special-purpose analysis software
- No statistical tool can take over the burden of sense-checking models

Outstanding Issues

Topics for Research

Marginal Analysis
- Confidence intervals on
  - Delta scores (easy)
  - Marginal Information values (hard)
- Re-design characteristic analysis to focus on partition variables
- Characteristic Analysis for Continuous Characteristics
  - Splines
  - Cf. Ross Gayler

Scorecard Estimation
- “Stepwise” type algorithm using Marginal IV
  - rather than Deviance measures
  - but also using significance checks
- Logistic Regression with constraints
  - Monotonicity ↔ Sign constraint
  - Would eliminate much over-fitting through stepwise

MORE POWER FROM STANDARD TOOLS
USE THE STATISTICS!
References

- Gerard SCALLAN (2011) “Building Better Scorecards” (Scoreplus, Course Notes, 2011 edition – Sections 5, 7, 8; Sections 8, 11 in older editions)