Low default modelling: a comparison of techniques based on a real Brazilian corporate portfolio

MSc Guilherme Fernandes and MSc Carlos Rocha

Credit Scoring and Credit Control Conference XII August 2011
Topics

- The problem
- Modelling techniques:
  - Classical logistic regression
  - Bayesian logistic regression
  - Limited logistic regression
  - Oversampling combined with correction
- Model validation
- Results
- Conclusion
Previous work

\[ EL = PD \times EAD \times LGD \]

**Basel II**

**Statistical model**

- Hand and Henley (1997):
  - Statistical classification methods
  - Consumer credit scoring
  - Large data
- Pluto and Tasche (2006):
  - Low default portfolio
  - PD estimation
The problem

<table>
<thead>
<tr>
<th>Common situation</th>
<th>Percentage of deposit</th>
<th>Default rate</th>
<th>Risk increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle financing</td>
<td>0%</td>
<td>30%</td>
<td>1.5x</td>
</tr>
<tr>
<td>Retail individuals</td>
<td>1%-30%</td>
<td>20%</td>
<td>2x</td>
</tr>
<tr>
<td>Risk driver: Percentage of deposit</td>
<td>31%-50%</td>
<td>10%</td>
<td>2x</td>
</tr>
<tr>
<td></td>
<td>51%-80%</td>
<td>5%</td>
<td>5x</td>
</tr>
<tr>
<td></td>
<td>81%-99%</td>
<td>1%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default rate</td>
</tr>
<tr>
<td>2.5%</td>
</tr>
<tr>
<td>0.9%</td>
</tr>
<tr>
<td>1.1%</td>
</tr>
<tr>
<td>0.3%</td>
</tr>
<tr>
<td>0.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LDP</th>
<th>Percentage of deposit</th>
<th>Default rate</th>
<th>Risk increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truck financing</td>
<td>0%</td>
<td>2.5%</td>
<td>2.9x</td>
</tr>
<tr>
<td>Middle companies</td>
<td>1%-30%</td>
<td>0.9%</td>
<td>0.75x</td>
</tr>
<tr>
<td>Risk driver: Percentage of deposit</td>
<td>31%-50%</td>
<td>1.1%</td>
<td>4.7x</td>
</tr>
<tr>
<td></td>
<td>51%-80%</td>
<td>0.3%</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>81%-99%</td>
<td>0.0%</td>
<td></td>
</tr>
</tbody>
</table>

© 2011 Experian Limited. All rights reserved.
Topics

- Scenario and the problem
- Modelling techniques:
  - Classical logistic regression
  - Bayesian logistic regression
  - Limited logistic regression
  - Oversampling combined with correction
- Model validation
- Results
- Conclusion
Classical logistic regression

Model statement

\[ Y_i \sim Bernoulli(p_i) \]

Target-covariates link

\[ p_i = \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \]

Likelihood function

\[ \ln(L(\beta \mid y)) = -\sum_{i=1}^{n} \ln(1 + \exp((1 - 2y_i)x_i \beta)) \]

Where:

- \( i \) = observation
- \( n \) = sample size
- \( x_i \) = covariate's vector
- \( y_i \) = response variable (1: default; 0: non default)
- \( p_i \) = probability of default (i - th observation)
- \( \beta \) = parameter vector
Topics

- Scenario and the problem
- Modelling techniques:
  - Classical logistic regression
  - Bayesian logistic regression
  - Limited logistic regression
  - Oversampling combined with correction
- Model validation
- Results
- Conclusion
Bayesian logistic regression

Model statement

\[ Y_i \sim Bernoulli(p_i) \]

Target-covariates link

\[ p_i = \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \]
\[ \beta_k \sim Normal(\mu_k, \sigma_k^2) \]

Posterior distribution

\[ \text{posterior}(\beta \mid y) \propto L(y \mid \beta) \cdot \text{prior}(\beta) \]

Onde:

- \( i \) = observation
- \( n \) = sample size
- \( y_i \) = response variable (1: default; 0: non default)
- \( prior(\beta_i) = Normal(\mu_i, \sigma_i^2) \)
- \( x_i \) = covariates vector
- \( \beta \) = parameters vector
- \( p_i \) = probability of default
- \( L(\beta \mid y) = \exp \left( - \sum_{i=1}^{n} \ln(1 + \exp((1 - 2y_i)x_i \beta)) \right) \)
Bayesian logistic regression

Model statement

\[ Y_i \sim \text{Bernoulli}(p_i) \]

Target-covariates link

\[ p_i = \left( \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \right) \quad \beta_k \sim \text{Normal}(\mu_k, \sigma_k^2) \]

Posterior distribution

\[ \text{posterior}(\beta | y) = \frac{L(y | \beta) \cdot \text{prior}(\beta)}{\int L(y | \beta) \cdot \text{prior}(\beta) d\beta} \]

Prior information of risk drivers

Data information

PDF of parameters for the model

Solved via Monte Carlo simulation (MCMC)
Bayesian logistic regression

Posterior distribution

\[ \text{posterior}(\beta | y) \propto L(y | \beta) \cdot \text{prior}(\beta) \]

Prior distributions

**Non Informative**

\[ \text{prior}(\beta) = \text{Normal}(\mu = 0, \sigma^2 = 1000) \]

**Informative**

\[ \text{prior}(\beta) = \text{Normal}(\mu = 1.5, \sigma^2 = 1.2) \]

Knowledge about the parameters prior to data information
Topics

- Scenario and the problem
- Modelling techniques:
  - Classical logistic regression
  - Bayesian logistic regression
  - Limited logistic regression
  - Oversampling combined with correction
- Model validation
- Results
- Conclusion
Limited logistic regression

Model statement

\[ Y_i \sim Bernoulli(p_i) \]

Target-covariates link

\[ p_i = \omega \cdot \left( \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \right) \]

Classical logistic regression

Limited logistic regression (w = 0.2)
Limited logistic regression

Model statement

\[ Y_i \sim Bernoulli(p_i) \]

Target-covariates link

\[ p_i = \omega \cdot \left( \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \right) \]

Log-Likelihood function

\[
\ln(L(\beta \mid y)) = -\sum_{i=1}^{n} \left( y_i \ln \left( \omega \cdot \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \right) + (1 - y_i) \ln \left( 1 - \omega \cdot \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \right) \right) I_{(0,1)}(\omega)
\]

Onde:

- \( i \) = observation
- \( n \) = sample size
- \( y_i \) = response variable (1: default; 0: non default)
- \( x_i \) = covariates vector
- \( \beta \) = parameters vector
- \( \omega \) = upper bounding parameter (0 < \omega < 1)
- \( p_i \) = probability of default
Topics

- Scenario and the problem
- Modelling techniques:
  - Classical logistic regression
  - Bayesian logistic regression
  - Limited logistic regression
  - Oversampling combined with correction
- Model validation
- Results
- Conclusion
Oversampling and state-dependent correction

Original LDP

Default observations artificially created

Biased estimators due to oversampling

Unbiased estimators (McCullagh e Nelder - 1989)
Oversampling technique: SMOTE*

- Artificial observations
  1. Select defaulted observations
  2. Randomly choose 2 observations
  3. Randomize one point within the space defined
  4. Estimate model with the inflated default rate database

*SMOTE: Chawla et al. 2002
State-dependent sample

Weighted Log-likelihood function

\[
\ln \left( L_w \left( \beta \mid y \right) \right) = \omega_1 \sum_{[Y_i = 1]} \ln \left( p_i \right) + \omega_0 \sum_{[Y_i = 0]} \ln \left( 1 - p_i \right) \\
= -\sum_{i=1}^{n} \omega_i \ln \left( 1 + \exp \left( \left( 1 - 2y_i \right)x_i \beta \right) \right)
\]

Parameters (\( \beta \)) estimated by WMLE are biased, even on large data. McCullagh e Nelder (1989) present the correction so the model will be unbiased.

Onde:
- \( i \) = observation
- \( n \) = sample size
- \( x_i \) = covariates vector
- \( p_i \) = probability of default
- \( \omega_i \) = weight of \( i^{th} \) observation = \( \omega_1 Y_i + \omega_0 (1 - Y_i) \)
- \( \tau \) = default rate in population
- \( y_i \) = response variable (1 : default; 0 : non default)
- \( \beta \) = parameters vector
- \( \omega_1 = \tau / \bar{y} \)
- \( \omega_0 = (1 - \tau) / (1 - \bar{y}) \)
- \( \bar{y} = \frac{1}{n} \sum y_i \)

\[
\xi_i = 0.5Q_{ii} \left[ (1 + \omega_i) \pi_i - \omega_i \right] \\
Q_{ii} = \text{elemento}_i \left[ X (X'WX)^{-1} X' \right]
\]

\[
\hat{\nu}_i = \hat{\beta} - \hat{\nu}_i = (X'WX)^{-1} X'W \hat{\xi}_i
\]

© 2011 Experian Limited. All rights reserved.
Topics

- Scenario and the problem
- Modelling techniques:
  - Classical logistic regression
  - Bayesian logistic regression
  - Limited logistic regression
  - Oversampling combined with correction
- Model validation
- Results
- Conclusion
Model validation

**KS**

\[ KS = \text{Max}(|\text{NDAcum} - \text{Dacum}|) \]

**Gini**

\[ \text{Gini} = \frac{A}{A+B} \]
Model validation

- Alternate method for out-of-time sample

Diagram:
- Training sample
- N resamples with replacement
- N estimates for KS / Gini
- KS distribution
- Gini distribution

KS

Gini
Topics

- Scenario and the problem
- Modelling techniques:
  - Classical logistic regression
  - Bayesian logistic regression
  - Limited logistic regression
  - Oversampling combined with correction
- Model validation
- Results
- Conclusion
Results

Data set

- Revenue > US$ 130 per year
- 1.327 obs
- Period: 2003 – 2008
- 50 defaulted observation
Results

Variables available and selection

- Balance sheet
- Credit demand
- Negative statement
- Payment behaviour
- Suppliers history

Top partial AUROC (29 variables)

<table>
<thead>
<tr>
<th>Balance sheet</th>
<th>Partial AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var 1</td>
<td>0.6</td>
</tr>
<tr>
<td>Var 2</td>
<td>0.5</td>
</tr>
<tr>
<td>Var 3</td>
<td>0.4</td>
</tr>
<tr>
<td>Var 4</td>
<td>0.3</td>
</tr>
<tr>
<td>Var 5</td>
<td>0.2</td>
</tr>
<tr>
<td>Var 6</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Credit demand</th>
<th>Partial AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var 1</td>
<td>0.6</td>
</tr>
<tr>
<td>Var 2</td>
<td>0.5</td>
</tr>
<tr>
<td>Var 3</td>
<td>0.4</td>
</tr>
<tr>
<td>Var 4</td>
<td>0.3</td>
</tr>
<tr>
<td>Var 5</td>
<td>0.2</td>
</tr>
<tr>
<td>Var 6</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative statement</th>
<th>Partial AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var 1</td>
<td>0.6</td>
</tr>
<tr>
<td>Var 2</td>
<td>0.5</td>
</tr>
<tr>
<td>Var 3</td>
<td>0.4</td>
</tr>
<tr>
<td>Var 4</td>
<td>0.3</td>
</tr>
<tr>
<td>Var 5</td>
<td>0.2</td>
</tr>
<tr>
<td>Var 6</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Payment behaviour</th>
<th>Partial AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var 1</td>
<td>0.6</td>
</tr>
<tr>
<td>Var 2</td>
<td>0.5</td>
</tr>
<tr>
<td>Var 3</td>
<td>0.4</td>
</tr>
<tr>
<td>Var 4</td>
<td>0.3</td>
</tr>
<tr>
<td>Var 5</td>
<td>0.2</td>
</tr>
<tr>
<td>Var 6</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Suppliers history</th>
<th>Partial AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var 1</td>
<td>0.6</td>
</tr>
<tr>
<td>Var 2</td>
<td>0.5</td>
</tr>
<tr>
<td>Var 3</td>
<td>0.4</td>
</tr>
<tr>
<td>Var 4</td>
<td>0.3</td>
</tr>
<tr>
<td>Var 5</td>
<td>0.2</td>
</tr>
<tr>
<td>Var 6</td>
<td>0.1</td>
</tr>
</tbody>
</table>

* Spearman correlation index < 0.5
# Results

## Parameters per model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Domain</th>
<th>Classical model</th>
<th>Limited logistic</th>
<th>Bayesian model</th>
<th>SMOTE + state dep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td>-3.62280</td>
<td>-3.69740</td>
<td>-3.13471</td>
<td></td>
</tr>
<tr>
<td>$w$ (upper bound limit)</td>
<td></td>
<td>0.14641</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio of short term debt over current assets higher than 25</td>
<td>$(0;1); \text{Dummy for values higher than 25}$</td>
<td>0.84510</td>
<td>0.96934</td>
<td>0.86510</td>
<td>1.09750</td>
</tr>
<tr>
<td>Total of negative bureau statements included over the past 30 days higher than 8</td>
<td>$(0;1); \text{Dummy for values higher than 8}$</td>
<td>1.07010</td>
<td>2.12650</td>
<td>1.09010</td>
<td>0.77100</td>
</tr>
<tr>
<td>Total number of distinct companies that enquired over the past 15 days was higher than 35</td>
<td>$(0;1); \text{Dummy for values higher than 35}$</td>
<td></td>
<td></td>
<td></td>
<td>0.73950</td>
</tr>
<tr>
<td>Number of banking past due contracts (Max of 4)</td>
<td>Integer; limited in 4</td>
<td>0.26380</td>
<td>0.30292</td>
<td>0.26860</td>
<td>0.27870</td>
</tr>
<tr>
<td>Days since last negative statement paid (Max of 200 days) - square root tranformation</td>
<td>Real; limited in 14.14</td>
<td>-0.17440</td>
<td>-0.33405</td>
<td>-0.17930</td>
<td>-0.17714</td>
</tr>
<tr>
<td>Maximum of negative statements active at the same time (Max of 120) - square root transformation</td>
<td>Real; limited in 10.95</td>
<td>0.17130</td>
<td>0.33579</td>
<td>0.17270</td>
<td>0.14201</td>
</tr>
<tr>
<td>Total number of enquiries over the past 15 days (Max of 50)</td>
<td>Integer; limited in 50</td>
<td>0.03200</td>
<td>0.01352</td>
<td>0.03240</td>
<td></td>
</tr>
<tr>
<td>Total number of negative statement paid over the past 6 months (max of 45) - square root transformation</td>
<td>Real; limited in 6.70</td>
<td>-0.46370</td>
<td>-0.86208</td>
<td>-0.47180</td>
<td>-0.37800</td>
</tr>
</tbody>
</table>

* Spearman correlation index < 0.5
## Results

### Importance of variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Classical logistic regression</th>
<th>Limited logistic regression</th>
<th>Bayesian logistic regression</th>
<th>SMOTE + state dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio of short term debt over current assets higher than 25</td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
</tr>
<tr>
<td>Total of negative bureau statements included over the past 30 days higher than 8</td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
</tr>
<tr>
<td>Number of banking past due contracts (Max of 4)</td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
</tr>
<tr>
<td>Days since last negative statement paid (Max of 200 days) - square root transformation</td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
</tr>
<tr>
<td>Maximum of negative statements active at the same time (Max of 120) - square root transformation</td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
</tr>
<tr>
<td>Total number of enquiries over the past 15 days (Max of 50)</td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
</tr>
<tr>
<td>Total number of negative statement paid over the past 6 months (max of 45) - square root transformation</td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
</tr>
<tr>
<td>Total number of distinct companies that enquired over the past 15 days was higher than 35</td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
<td><img src="image" alt="Value" /></td>
</tr>
</tbody>
</table>
Results

Performance measures

10.000 bootstrap samples

Box-plot KS from bootstrap samples

Box-plot Gini from bootstrap samples

Bayesian model: High performance and low variability
Results

Default rate assortment

Bayesian model - Default rate assortment
Conclusion

- Pluto & Tasche: estimates PD for a pre-existing rating grade
- Approaches presented:
  - Limited logistic regression: Best KS, worst Gini;
  - SMOTE + State dependent: incorporates different variable;
  - Classic logistic regression: High KS and Gini, but higher variability on bootstrap;
  - Bayesian logistic regression: High KS and Gini.
- Bayesian model gives reasonably assorted default rate even on LDP
- Future research: how to incorporate informative prior using specialist’s knowledge and bureau information
Questions