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

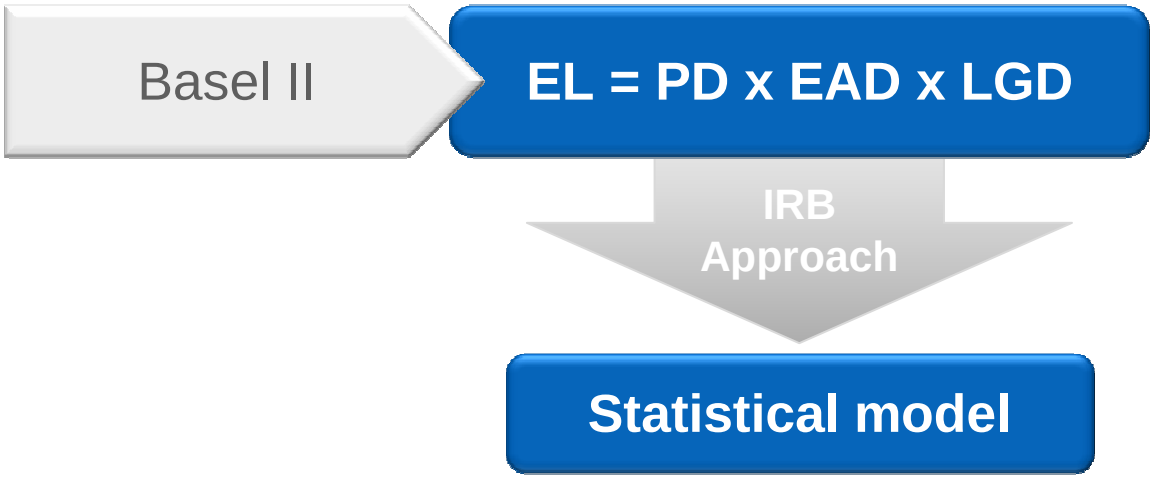
Serasa Experian

Analytics Brazil

Topics

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

Previous work



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

The problem

Common situation

- Vehicle financing
- Retail individuals
- Risk driver: Percentage of deposit

Percentage of deposit	Default rate	Risk increase
0%	30%	1.5x
1%-30%	20%	2x
31%-50%	10%	2x
51%-80%	5%	5x
81%-99%	1%	

LDP

- Truck financing
- Middle companies
- Risk driver: Percentage of deposit

Default rate	Risk increase
2.5%	2.9x
0.9%	0.75x
1.1%	4.7x
0.3%	NA
0.0%	

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

$$p_i = \left(\frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \right)$$

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

y_i = response variable (1 : default; 0 : non default)

x_i' = covariate's vetor

β = parameter vector

p_i = probability of default (i - th observation)

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

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

Posterior distribution

$$posterior(\beta \mid y) \propto L(y \mid \beta) \cdot prior(\beta)$$

Onde:

i = observation

n = sample size

y_i = response variable (1: default; 0: non default)

$prior(\beta_k) = Normal(\mu_k, \sigma_k^2)$

x_i' = covariates vector

β = 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

$$Y_i \sim Bernoulli(p_i)$$

Target-covariates link

$$p_i = \left(\frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \right) \qquad \beta_k \sim Normal(\mu_k, \sigma_k^2)$$

Posterior distribution

$$posterior(\beta \mid y) = \frac{L(y \mid \beta) \cdot prior(\beta)}{\int L(y \mid \beta) \cdot prior(\beta) d\beta}$$

Prior information of risk drivers

Data information

PDF of parameters for the model

Solved via Monte Carlo simulation (MCMC)

Bayesian logistic regression

Posterior distribution

$$posterior(\beta \mid y) \propto L(y \mid \beta) \cdot prior(\beta)$$

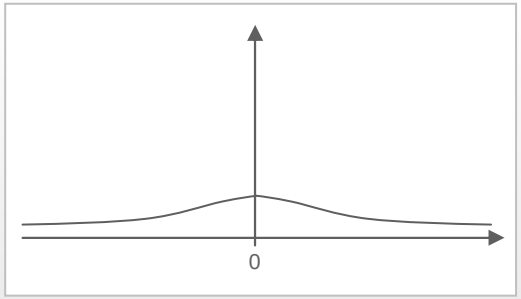
Prior distribution

Knowledge about the parameters prior to data information

Prior distributions

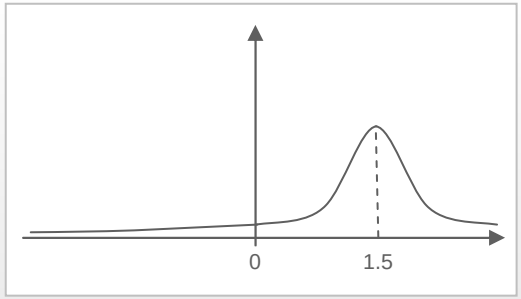
Non Informative

$$prior(\beta) = Normal(\mu = 0, \sigma^2 = 1000)$$



Informative

$$prior(\beta) = Normal(\mu = 1.5, \sigma^2 = 1.2)$$



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

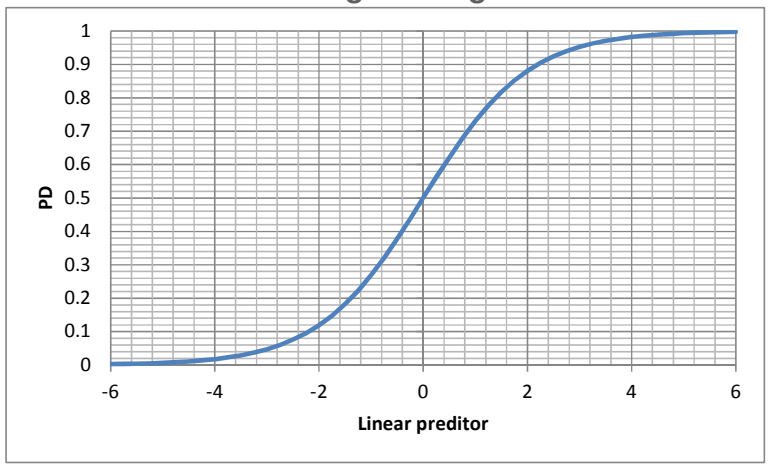
Model statement

$$Y_i \sim Bernoulli(p_i)$$

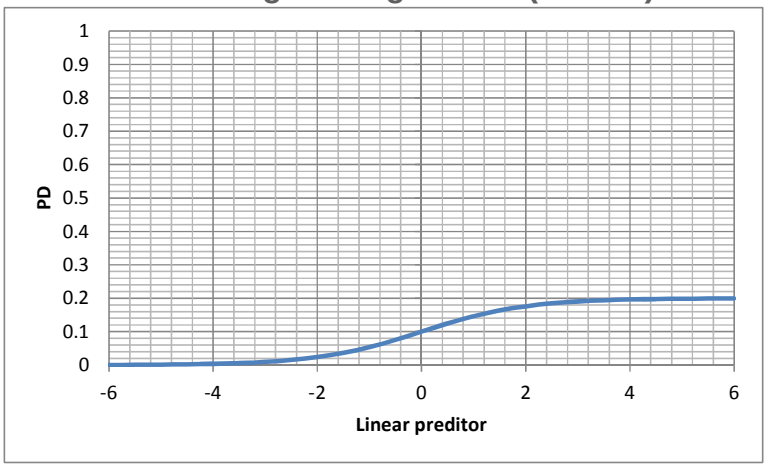
Target-covariates link

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

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

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

Onde:

i = observation

n = sample size

y_i = response variable (1 : default; 0 : non default)

ω = upper bounding parameter $(0 < \omega < 1)$

x_i' = covariates vector

β = parameters vector

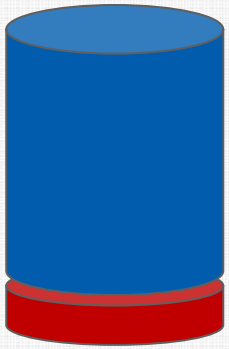
p_i = probability of default

Topics

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

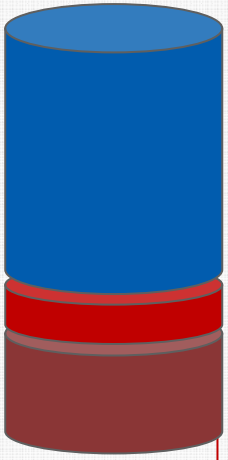
Oversampling and state-dependent correction

PD: 1%



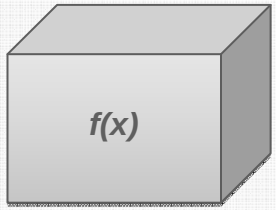
Original LDP

PD: 15%



Default observations artificially created

Estimated model

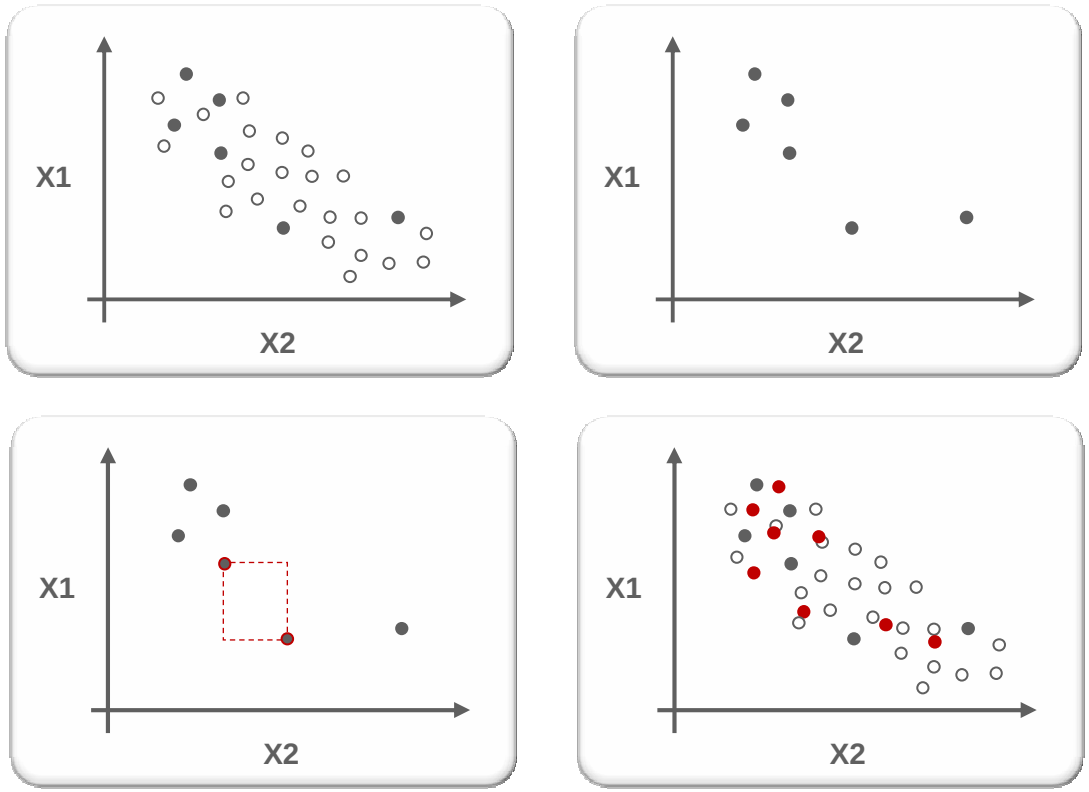


Biased estimators due to oversampling

State-dependent correction

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

$$\ln\left(L_w(\beta \mid y)\right) = \omega_1 \sum_{[Y_i=1]} \ln\left(p_i\right) + \omega_0 \sum_{[Y_i=0]} \ln\left(1-p_i\right)$$

$$= -\sum_{i=1}^{n} \omega_i \ln\left(1+\exp\left((1-2y_i)x_i'\beta\right)\right)$$

Parameters (β) 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
- ω_i = weight of ith observation = $\omega_1 Y_i + \omega_0(1-Y_i)$
- τ = default rate in population
- y_i = response variable (1: default; 0: non default)
- β = parameters vector
- $\omega_1 = \tau/\bar{y}$
- $\omega_0 = (1-\tau)/(1-\bar{y})$
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$\xi_i = 0.5 Q_{ii}\left[(1+\omega_1)\pi_i - \omega_1\right]$$

$$Q_{ii} = elemento_{ii}\left[X\left(X'WX\right)^{-1}X'\right]$$

$$vi\acute{e}s\left(\hat{\beta}\right) = \left(X'WX\right)^{-1}X'W\xi_i$$

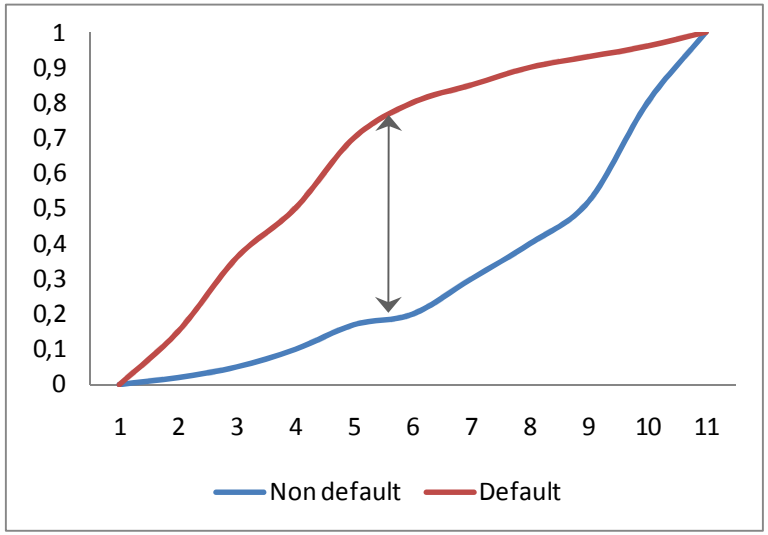
$$\tilde{\beta} = \hat{\beta} - vi\acute{e}s\left(\hat{\beta}\right)$$

Topics

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

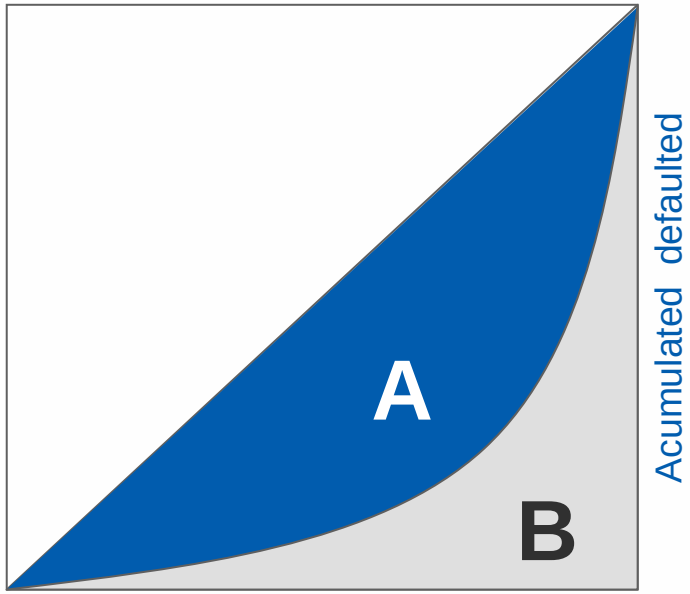
Model validation

KS



KS = Max(|NDAcum – Dacum|)

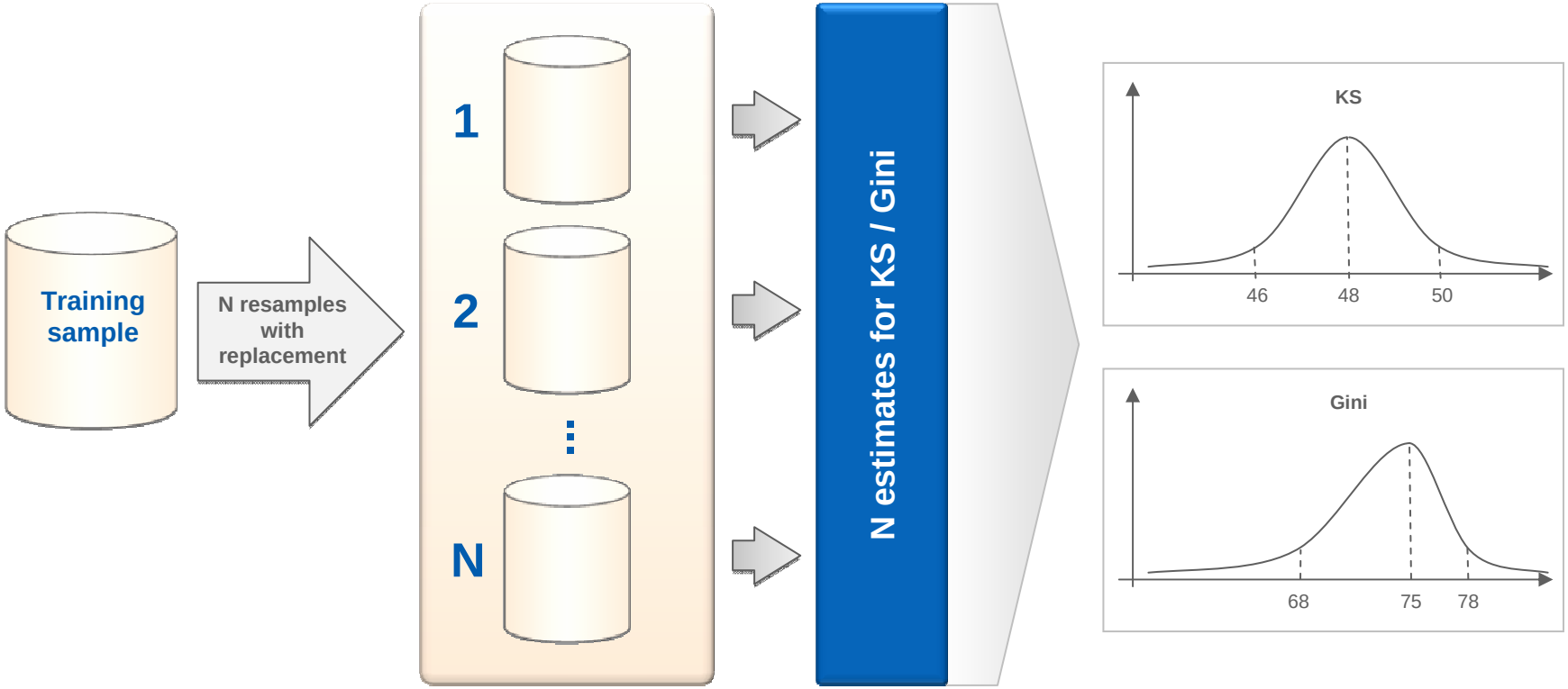
Gini



Gini = A / (A+B)

Model validation

- Alternative method for out-of-time sample



Topics

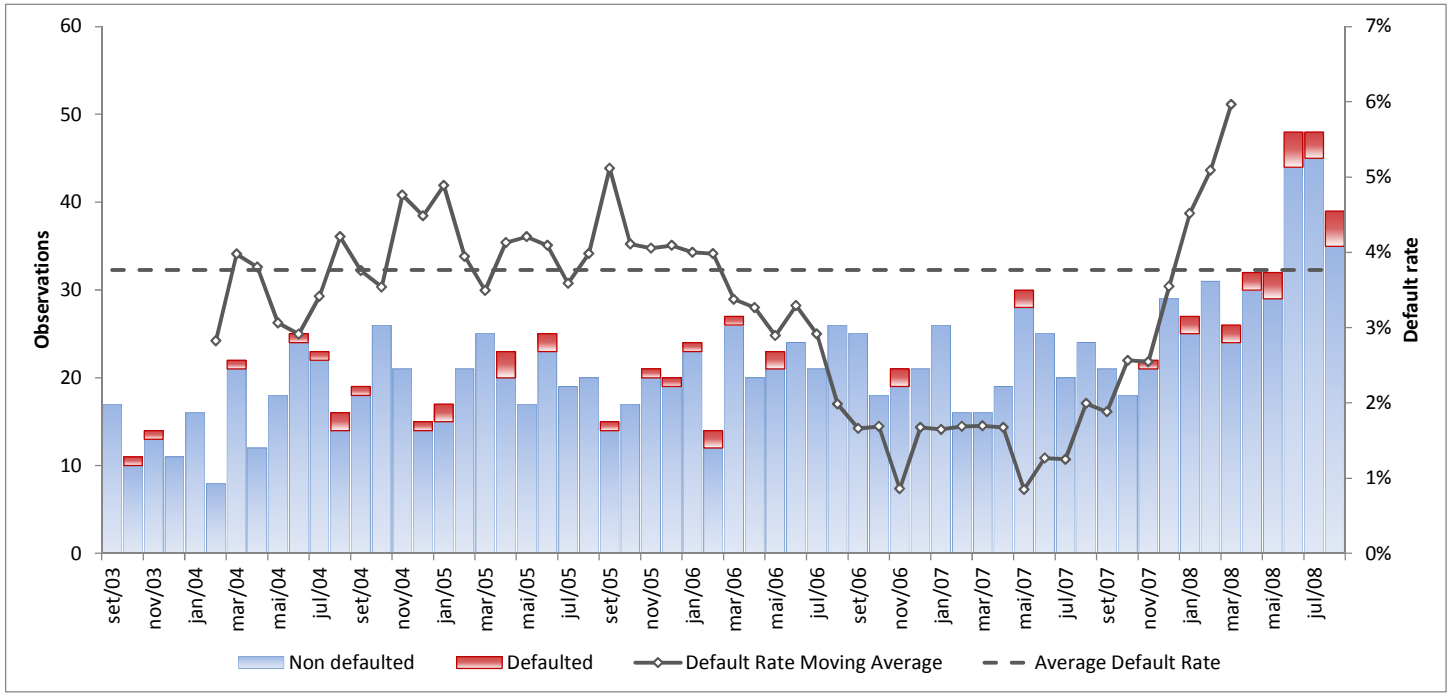
- Scenario and the problem
- Modelling techniques:
  - Classical logistic regression
  - Bayesian logistic regression
  - Limited logistic regression
  - Oversampling combined with correction
- Model validation
- Results
- Conclusion

Results

Data set

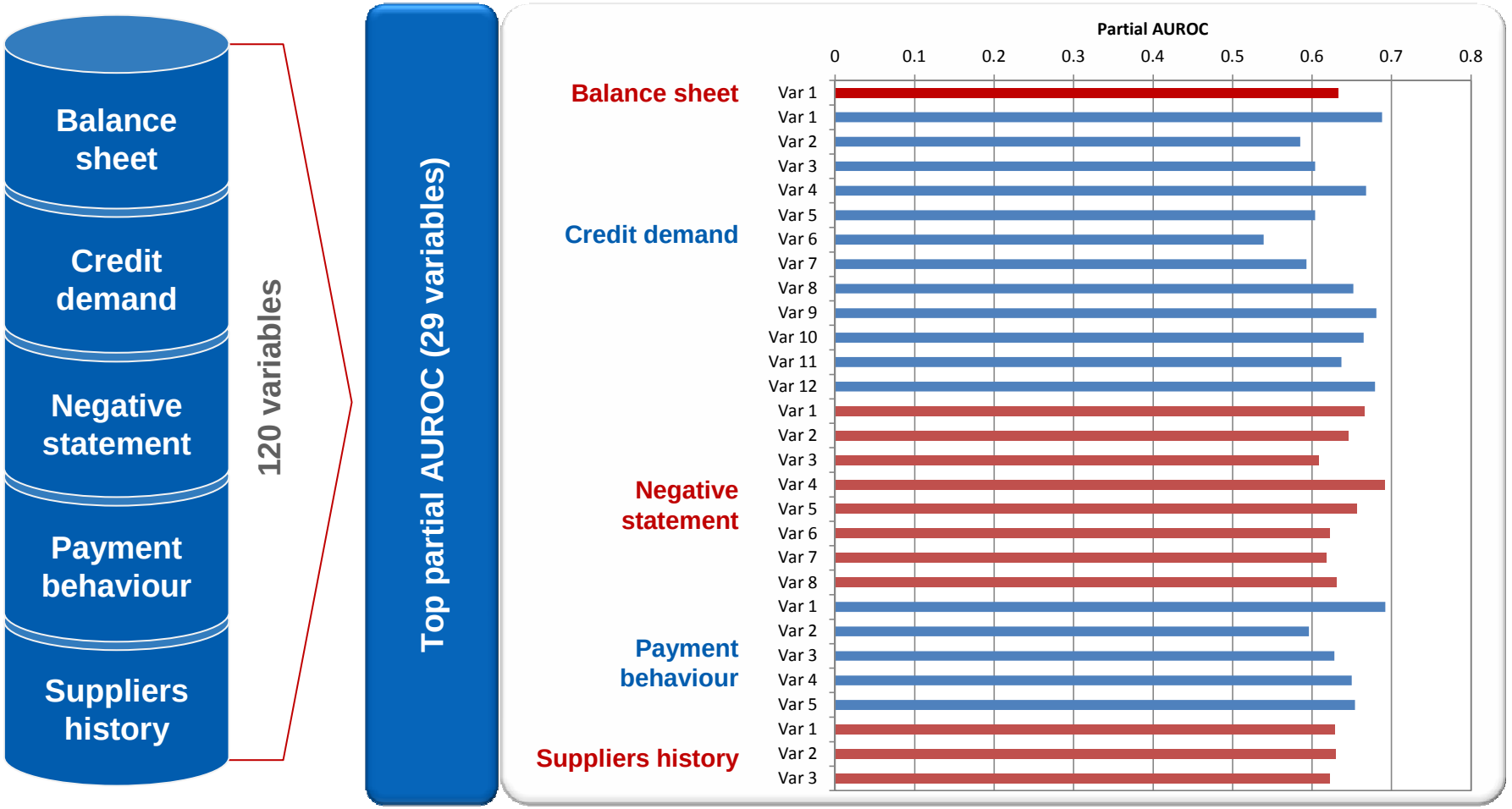
Corporate portfolio

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



Results

Variables available and selection



*** Spearman correlation index < 0.5**

Results

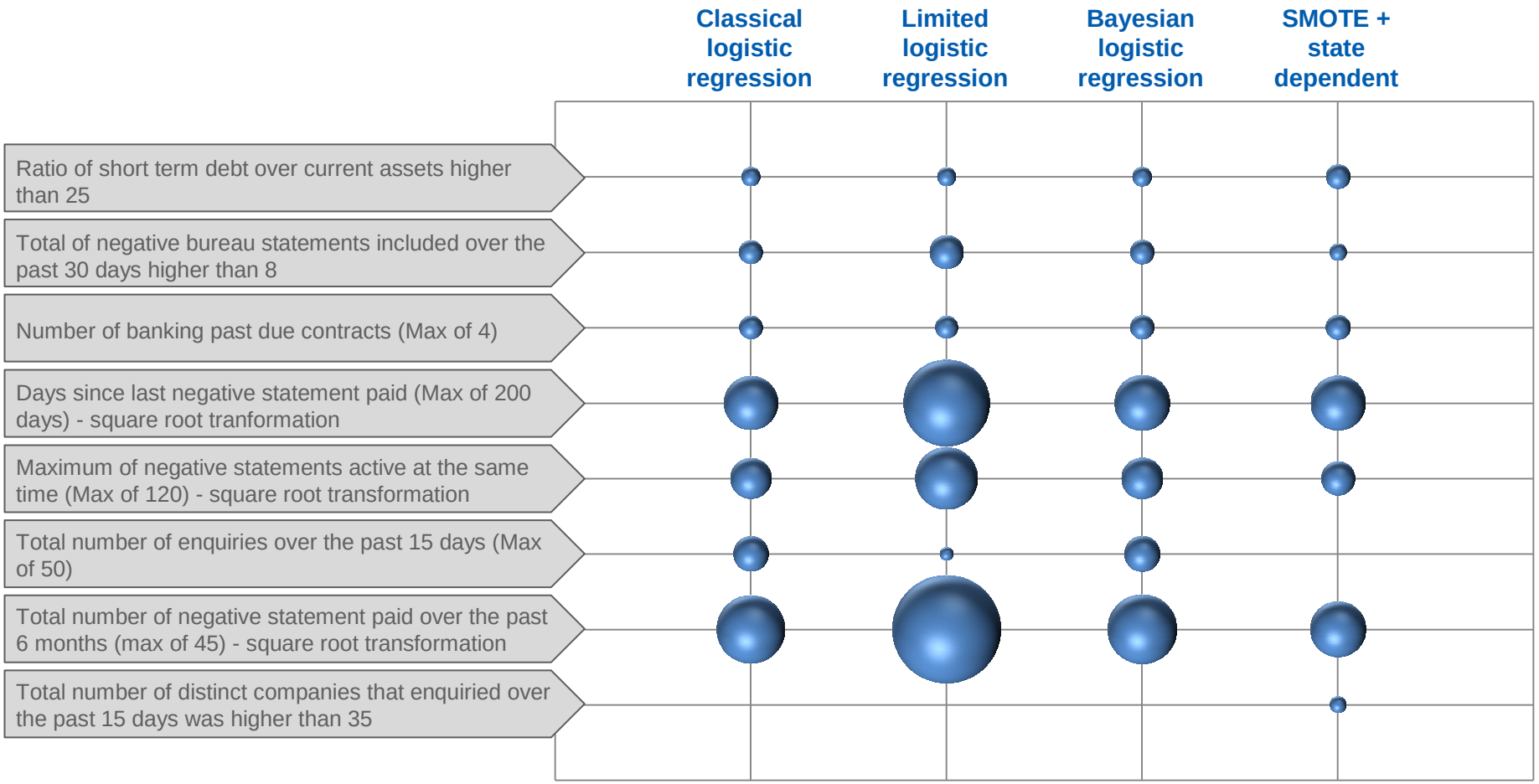
Parameters per model

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

*** Spearman correlation index < 0.5**

Results

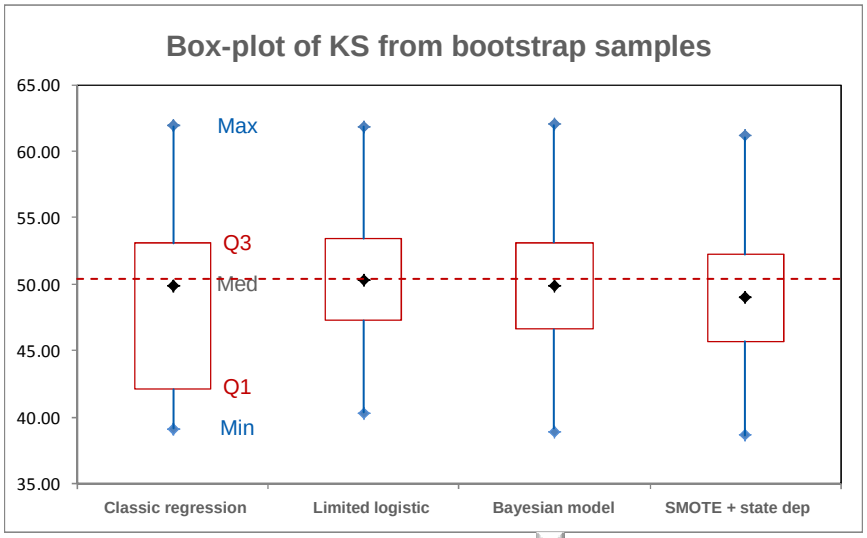
Importance of variables

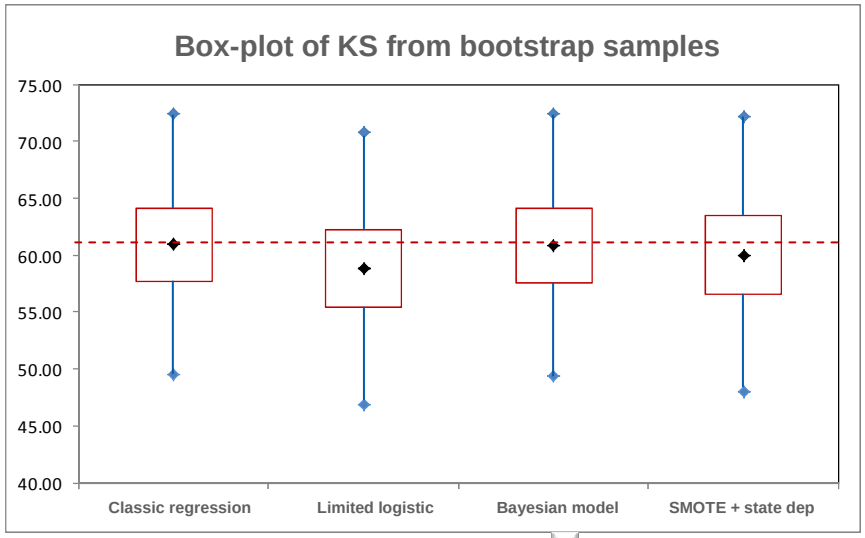


Results

Performance measures

10.000 bootstrap samples

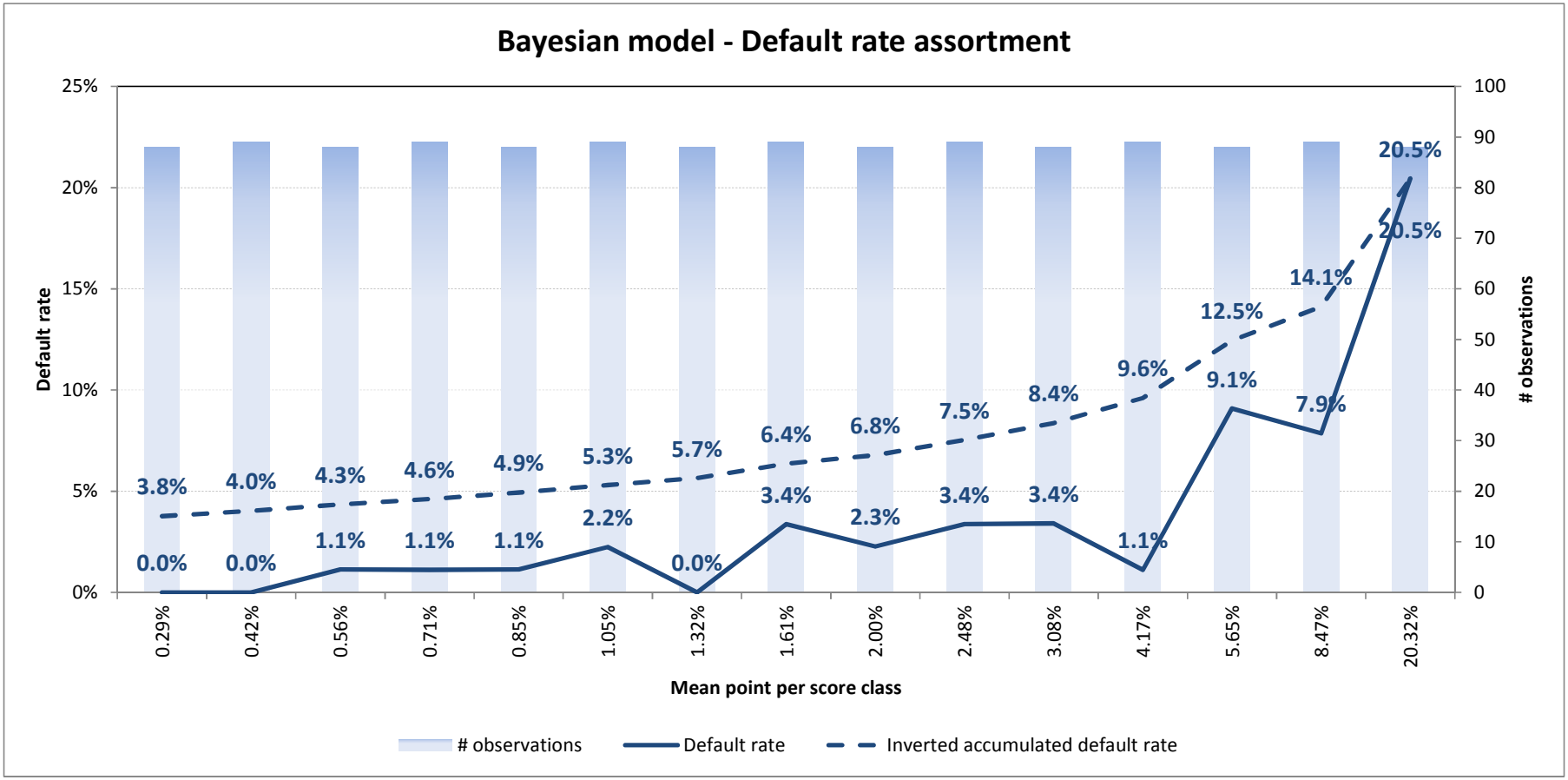




Bayesian model: High performance and low variability

Results

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

Guilherme B Fernandes
guilherme.fernandes@br.experian.com
Carlos A. Rocha
carlos.rocha@br.experian.com

Questions

Serasa Experian