Benchmarking State-Of-The-Art Regression Algorithms For Loss Given Default Modelling

Gert Loterman[1], Iain Brown[2]
David Martens[1,3], Christophe Mues[2], Bart Baesens[2,3]

[1]University College Ghent (BE)
[2]University of Southampton (UK)
[3]Catholic University of Leuven (BE)

Overview

Credit Risk Modelling with Basel II (1)

What?

Basel II is an international accord between banks to protect the international financial system.

How?

Basel II regulates risk and capital management requirements to ensure that a bank holds enough capital reserves proportional to the exposed risk of its lending.

'Your card is fine. I'm just checking that your bank hasn't expired'

(figure from http://www.worldbank.org)

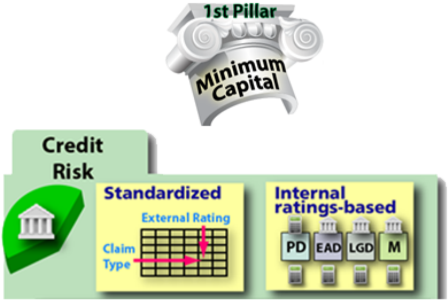
Credit Risk Modelling in Basel II (2)

Key components

PD: Probability of Default
LGD: Loss Given Default
EAD: Exposure At Default

Approaches

1. **Standardized**
2. **Internal Ratings Based**
   - Foundation
   - Advanced



(figure from http://www.bionicturtle.com)

Credit Risk Modelling with Basel II (3)

LGD estimation?

Foundation

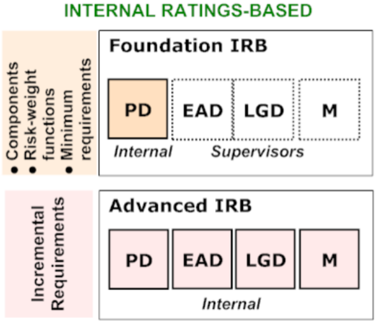
$\Downarrow$

regulator's estimate of LGD

Advanced

$\Downarrow$

internal estimate of LGD



(figure from http://www.bionicturtle.com)

Credit Risk Modelling with Basel II (4)

LGD errors are more expensive than PD errors!

Basel II Capital Requirement: K(LGD,PD).EAD with

$$K = LGD \times \left(\Phi\left(\sqrt{\frac{1}{1-\rho}}\Phi^{-1}(PD) + \sqrt{\frac{\rho}{1-\rho}}\Phi^{-1}(0.999)\right) - PD\right)$$

Example: PD = 0.03, LGD = 0.50, EAD = $10000

K(0.03,0.50)(10000) = $34.37

- 10% over estimate on PD means capital required is K(0.033,0.50)(10000) = $36.73
- 10% over estimate on LGD means capital required is K(0.03,0.55)(10000) = $37.80

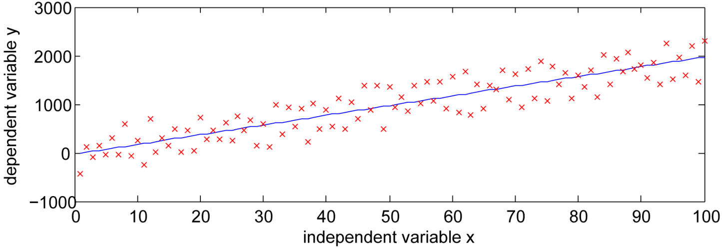
(example from Marfintel workshop)

Regression analysis

How is LGD estimated given a customer's personal loan data?

$$\widehat{\text{LGD}} = \mathbf{f(}\text{creditinfo, customerinfo, warrantyinfo, ...}\mathbf{)}$$

Simplest case: $\widehat{y} = \mathbf{b_0} + \mathbf{b_1}.x$



General case: $\widehat{y} = \mathbf{f(}x_1, x_2, ..., x_n\mathbf{)}$

Goal

To what degree can LGD be predicted using regression models?

In this study 24 regression techniques are used in the prediction of LGD on 5 real-life data sets from major international banking institutions.

The predictive performance of these models are evaluated and compared with each other.

Experimental set-up (1)

1. Training and testing

Each data set is split in 2/3 training and 1/3 test. The regression models are then constructed on the training set.

2. Validation

Where a regression model requires parameters to be optimized, 10-fold cross validation is used on the training set. The optimal parameters are estimated by minimizing the squared difference between predicted and observed values from the training set.

Experimental set-up (2)

3. Evaluation

After building models with the different regression techniques, their predictive performances are measured on the test set with appropriate metrics for evaluation and comparison.

4. Significance testing

A Friedman test is performed to determine if at least one technique is significantly better or worse than another. Further a post-hoc Nemenyi test is calculated in order to report any significant differences between the techniques.

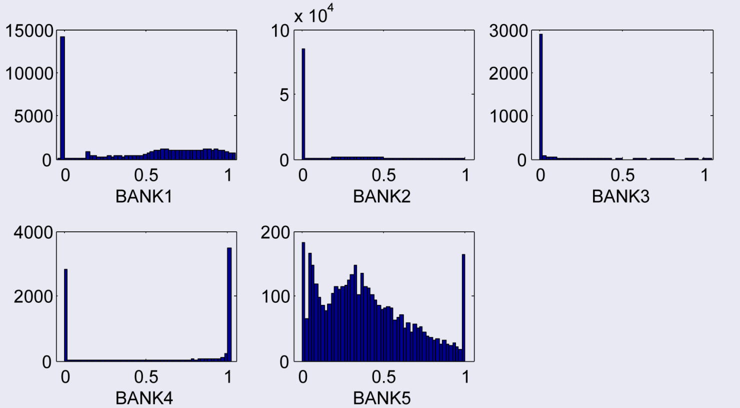
Data sets (1)

Characteristics

Dataset	Inputs	Data size	Training size	Test size
BANK1	44	47853	31905	15948
BANK2	18	119210	79479	39732
BANK3	14	3351	2232	1119
BANK4	12	7889	5260	2629
BANK5	35	4097	2733	1364

Data sets (2)

LGD distributions



Regression models (1)

Overview

1. **One stage models**
   - Linear models
   - Linear models with transformation
   - Nonlinear models
2. **Two stage models**
   - Logaritmic + (non)linear models
   - Linear + nonlinear models

Regression models (2)

Linear regression model

$$\widehat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Techniques

Ordinary Least Squares (OLS)
Robust Regression (RoR)
Ridge Regression (RiR)

Regression models (3)

Linear regression model with transformation

$$\widehat{y}_t = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

$$\text{with } \widehat{y}_t = f(\widehat{y}) \text{ and } \widehat{y} = f^{-1}(\widehat{y}_t)$$

Techniques

OLS + Beta Transformation (B-OLS)
OLS + Box-Cox Transformation (BC-OLS)
Beta Regression (BR)

Regression models (4)

Nonlinear regression model

$$\widehat{y} = f(x_1, x_2, ..., x_n)$$

Techniques

Classification And Regression Trees (CART)
Multivariate Adaptive Regression Splines (MARS)
Least Squares Support Vector Machines (LSSVM)
Artificial Neural Networks (ANN)

Regression models (5)

Logistic + (non)linear regression model

$$\widehat{y} = \underbrace{P[in\ peak].\bar{y}_{peak}}_{=\ 0} + \underbrace{(1 - P[in\ peak])}_{logistic\ \ regression}\ \underbrace{f(x_1, x_2, ..., x_n)}_{(non)linear\ \ regression}$$

$$with\ P[in\ peak] = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n)}}$$

Techniques

LOG + OLS/RoR/RiR
LOG + B-OLS/BC-OLS/BR
LOG + CART/MARS/LSSVM/ANN

Regression models (6)

Linear + nonlinear regression model

$$\hat{y} = \underbrace{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n}_{linear\ \ regression} + \underbrace{e}_{nonlinear\ \ regression}$$

$$with\ \hat{e} = f(x_1, x_2, \ldots, x_n)$$

Techniques

OLS + CART/MARS/LSSVM/ANN

Performance metrics (1)

Definition

A performance metric evaluates to what degree the predicted values $\hat{y}$ differ from the actual values y. Each metric has its own method to express the predicted performance of a model as a quantitative value.

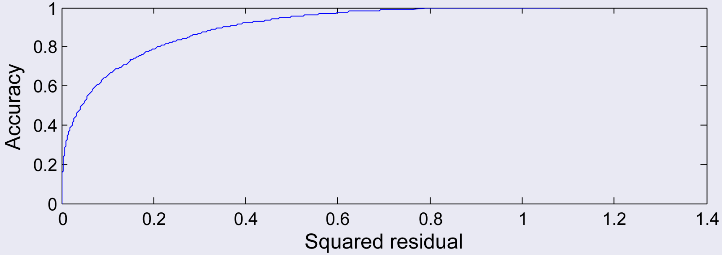
Techniques

Mean Absolute Error *MAE*
Root Mean Square Error *RMSE*
Receiver Operating Characteristics *ROC*
Regression Error Characteristics *REC*
Coefficient of Determination R^2
Correlation Coefficients r (Pearson), ρ (Spearman) and τ (Kendall)

Performance metrics (2)

Regression Error Characteristics

The REC curve plots the error tolerance on the x-axis versus the percentage of points predicted within the tolerance (or accuracy) on the y-axis. The area above the curve is an estimate of the error.



Significance tests

Is there at least one technique that outperforms the others?

The **Friedman test** is used to determine significant differences in the technique's performances over all data sets. When the p-value of the test is small (<0.05), there is evidence to assume there are differences in performance.

Which technique's performances differ significantly?

The **post-hoc Nemenyi test** is used to determine which technique's performances differ significantly from one another. Two or more techniques are significantly different if their average ranks differ by at least the critical difference.

Top performance results

Metric	BANK1	BANK2	BANK3
MAE	0.3054 (ANN)	0.0937 (ANN)	**0.0340 (BC-OLS)**
RMSE	0.3618 (ANN)	0.1459 (ANN)	0.1219 (OLS+ANN)
ROC	0.6824 (OLS+LSSVM)	0.8430 (LOG+ANN)	0.7216 (OLS+MARS)
REC	0.1309 (ANN)	0.0213 (ANN)	0.0133 (LOG+ANN)
R^2	0.1437 (ANN)	0.3743 (ANN)	0.2634 (LOG+ANN)
r	0.3802 (ANN)	0.6118 (ANN)	0.5381 (LOG+ANN)
ρ	0.3661 (ANN)	0.6118 (ANN)	0.2312 (BC-OLS)
τ	0.2594 (ANN)	0.4463 (ANN)	0.1765 (BC-OLS)

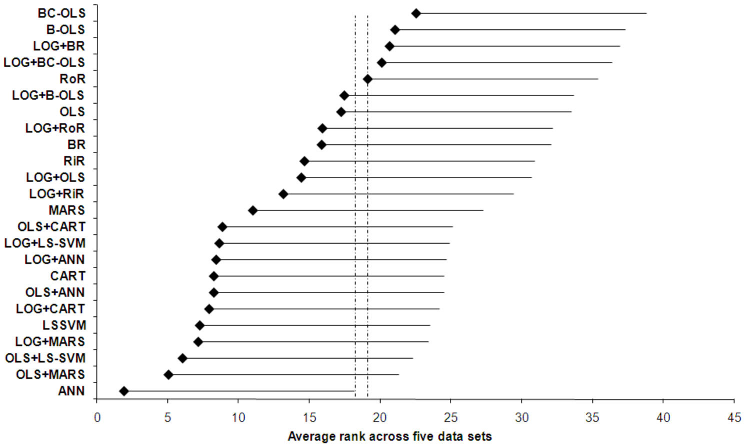
Metric	BANK4	BANK5
MAE	0.2190 (ANN)	0.0767 (ANN)
RMSE	0.3225 (ANN)	**0.1123 (ANN)**
ROC	0.8718 (OLS+LSSVM)	**0.9560 (ANN)**
REC	0.1038 (ANN)	**0.0124 (ANN)**
R^2	0.5197 (ANN)	**0.8261 (ANN)**
r	0.7226 (ANN)	**0.9100 (ANN)**
ρ	0.6339 (OLS+LSSVM)	**0.9100 (ANN)**
τ	0.4827 (CART)	**0.7539 (ANN)**

REC performance results

Technique	BANK1	BANK2	BANK3	BANK4	BANK5	AR
ANN	0,1309	0,0213	0,0152	0,1038	0,0124	1,9
OLS+MARS	0,1347	0,0234	0,0136	0,1074	0,0228	5
OLS+LS-SVM	0,1335	0,0219	0,0178	0,1038	0,0387	6
LOG+MARS	0,1360	0,0236	0,0149	0,1134	0,0196	7,1
LSSVM	0,1365	0,0222	0,0175	0,1063	0,0385	7,2
LOG+CART	0,1342	0,0236	0,0146	0,1169	0,0359	7,9
OLS+ANN	0,1368	0,0225	0,0170	0,1054	0,0415	8,2
CART	0,1392	0,0224	0,0154	0,1095	0,0346	8,2
LOG+ANN	0,1392	0,0234	0,0133	0,1163	0,0180	8,4
LOG+LS-SVM	0,1356	0,0230	0,0176	0,1136	0,0381	8,6
OLS+CART	0,1392	0,0224	0,0157	0,1120	0,0346	8,8
MARS	0,1372	0,0234	0,0212	0,1097	0,0193	11
LOG+RiR	0,1362	0,0263	0,0178	0,1130	0,0535	13,1
LOG+OLS	0,1366	0,0255	0,0179	0,1165	0,0541	14,4
RiR	0,1377	0,0258	0,0178	0,1192	0,0533	14,6
BR	0,1363	0,0275	0,0169	0,1425	0,0540	15,8
LOG+RoR	0,1373	0,0259	0,0185	0,1205	0,0531	15,9
OLS	0,1380	0,0259	0,0178	0,1197	0,0552	17,2
LOG+B-OLS	0,1406	0,0245	0,0182	0,1232	0,0538	17,4
RoR	0,1379	0,0276	0,0189	0,1304	0,0533	19,1
LOG+BC-OLS	0,2590	0,0273	0,0188	0,1186	0,0542	20,1
LOG+BR	0,1715	0,0285	0,0180	0,1265	0,0544	20,6
B-OLS	0,1843	0,0262	0,0188	0,1385	0,0546	21
BC-OLS	0,2096	0,0262	0,0190	0,1802	0,0553	22,5

Significance results

Demsar significance diagram



Conclusion (1)

Summary

- 24 techniques have been benchmarked on 5 real-life data sets from major international banking institutions.
- The LGD predicting performance of the regression models is low.
- Nonlinear techniques as ANN and MARS generally outperform linear techniques.
- The combination of OLS with nonlinear techniques gives more or less equal performances as the nonlinear techniques.
- Transforming the LGD before applying OLS does not increase performance.

Conclusion (2)

Future research

- Which independent variables are significant to predict LGD?
- How do corporate and personal loans differ in modelling LGD?
- What is the effect of the size of an LGD data set on a techniques predictive power?