Academy of Economic Studies
Doctoral School of Finance and Banking

Risk assessment using machine learning

MSc Student: Tarţa Alexandrina Alina
Supervisor: Prof. Dr. Moisă Altăr

Bucharest, June 2011

Topics

- 1. Motivation
- 2. Objectives
- 3. Literature review
- 4. Methodology and data input
- 5. Empirical results
- 6. Conclusions
- 7. References

1. Motivation

- Under the Basel II banks have the possibility to use their internal rating models to quantify the risks
- The core of the IRB approach is the use of banks' own estimates of the probability of default (PD) associated with an exposure
- Finding variables that can explain the default

2.Objectives

- Assessing risk using different models
- Estimating the probability of default
- Calibrating and validating the results

3. Literature review

- FitzPatrick (1932) compared 13 ratios of failed and successful companies

- The logit and probit models were introduced by Martin (1977) and Ohlson (1980)

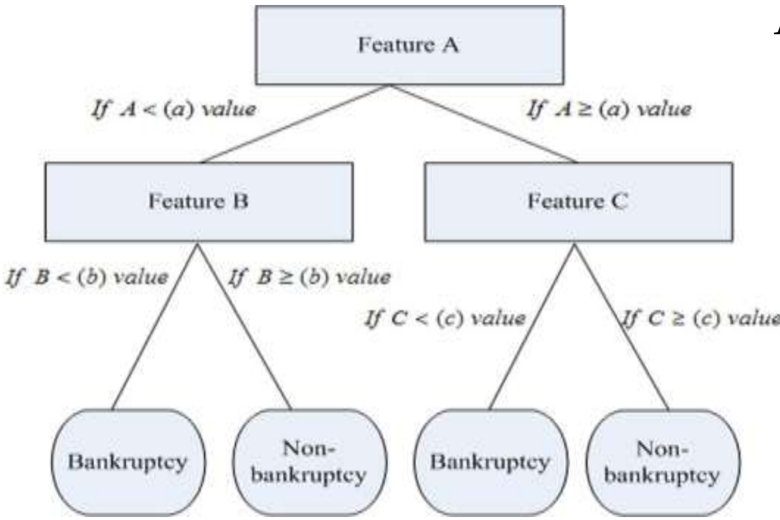
- West(2000) compared neural network models with other techniques such as logistic regression and decision trees

- Rodriguez, Kuncheva and Alonso (2006) introduced a new classifier called Rotation Forest that has very good performance.

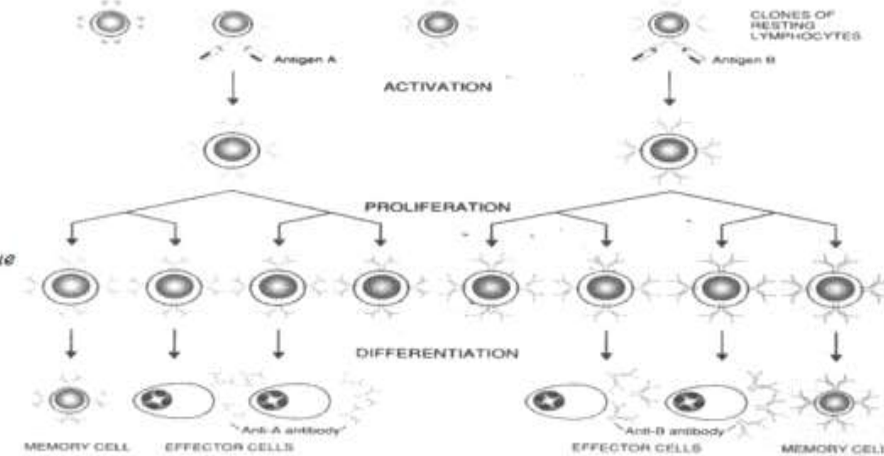
- Singh and Sengupta (2007) used a Clonal Selection Classification Algorithm for estimating probability of default.

4. Methodology and data input (1)

- Rotation forest



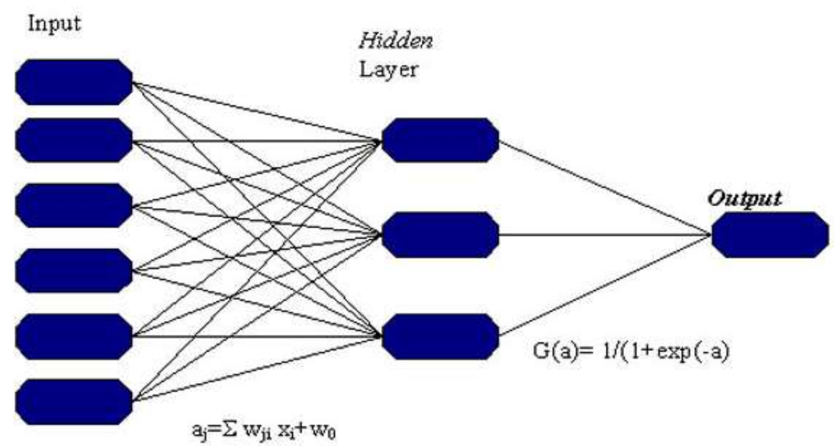
- Clonal Selection Classification Algorithm



- Logit

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

- Artificial Neural Network



4. Metodology and data input (2)

- The default definition is set acording to Basel II (90 days overdue)
- The database represents non-financial companies with bank loans
- The model was developed on 2008 balance-sheet data and in development sample are taken only companies with loan in 2009 that were not in default 12 month before
- A stratified sampling is employed. The strata are the main sectors the companies are active in
- Optimal allocation is used to allocate companies in development sample. The number of instances chosen for each stratum is computed as:

$$n_h = \frac{n * N_h * \sigma_h}{\sum_i N_i * \sigma_i}$$

4. Methodology and data input (3)
Performance measures

		Predicted class	
		Defaulter	Non-defaulter
Actual class	Defaulter	TP	FN *(Type I error)*
	Non-defaulter	FP *(Type II error)*	TN

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \text{Recall}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

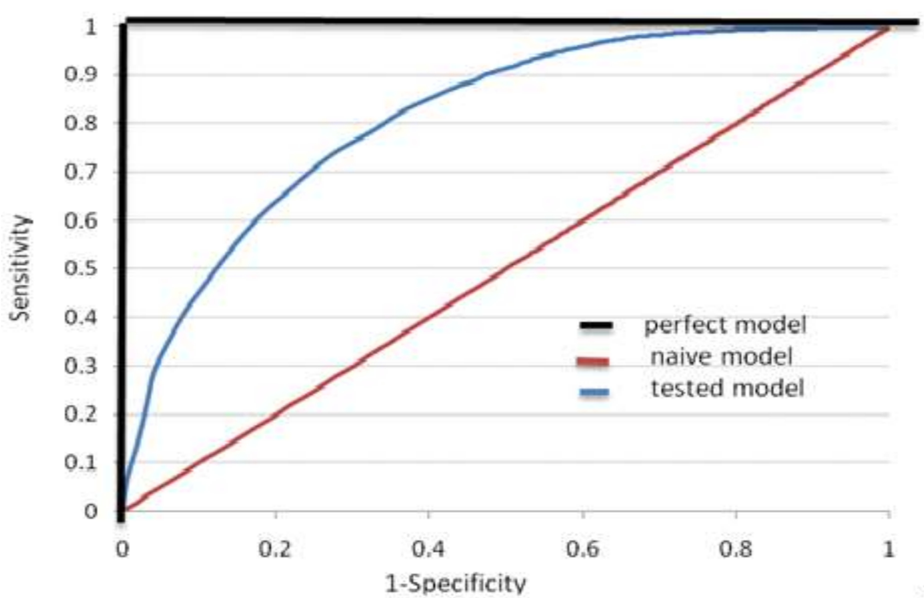
$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Methodology and data input (4) Performance measures

$$AR = \frac{TP + TN}{TP + FP + TN + FN}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Receiver Operator Characteristic (ROC) curve



- ROC>0.9 –exceptionally
- 0.8<ROC<0.9 excellent
- 0.7<ROC<0.8 acceptable

4. Methodology and data input (5) Variables selection –nonlinear models

- Filters used
  - _Kolmogorov Smirnov test
    - Compare the distribution of values of defaulters and nodefaulters for each variable
  - Multicolinearity
    - The variables with correlation coefficient higher then 0.7 are dropped
  - Correlation Feature Selection

$$M = \frac{k * r_{cf}}{\sqrt{k + k*(k-1)*r_{ff}}}$$

    - Only variables that are highly correlated with the class and uncorrelated with each other are kept
    - The acceptance of a variable will depend on the extend to which it predicts classes in areas of the sample not already predicted by the other variables

4. Methodology and data input (6) Variables selection –nonlinear models

- Gain Ratio (GR)

$$GR = {}^{IG}\!/_{IV}$$

$$IG(X,a) = H(X) - \sum_{f \in values(a)} \frac{|X_f|}{|X|} H(X_f)$$

$$IV = \sum_i \{\%non - defaulters - \%defaulters\} * WOE_i$$

$$WOE = \ln \frac{\%non - defaulters}{\%defaulters}$$

Selected variables

Variables	Mean	Standard deviation	CSF Evaluation	Gain Ratio Average merit
Cost of goods sold to stock	6.4	5.5	100%	0.218
Debt to equity	8.3	5.8	100%	0.201
Interest to total assets	0.03	0.02	100%	0.404
Interest coverage ratio	3.1	4.4	100%	0.353

4. Methodology and data input (7) Variables selection –Logit

- Linearity and monotony
  - the sample is divided in several subsamples that contain the same number of observation and for each group the default rate is computed
  - A linear regression of the historical defaults rate on the mean value of the variable is run
- Selected variables

variables	coefficient	t-stat	Std. error
constant	-1.92	-16.0331	0.08
Debt to equity	0.06	11.1994	1.7
Debt to value added	0.08	4.6177	0.53
Receivable cash conversion days	0.006	9.3864	13.3
Short term bank loan to total assets	4.38	17.2598	0.04
Asset turnover	-0.25	-7.1485	0.2
Dummy1 (overdue payment 0-15 days)	1.77	15.3777	0.09
Dummy2 (overdue payment 15-30 days)	2.71	11.5874	0.03
Dummy3 (overdue payment 30-60 days)	3.46	14.0495	0.01
Dummy4 (overdue payment 60-90 days)	4.37	8.2335	0.53

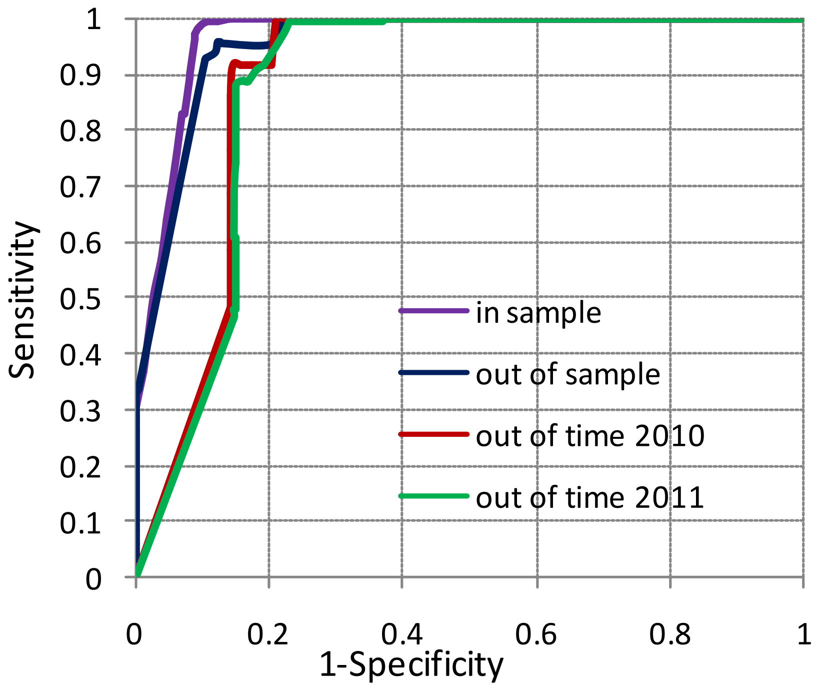
5. Empirical Results(1)

- Structure of companies with bank loans by sector of activity

	Dec. 2009		Dec. 2010		Jun. 2011
	Obs.	defaults	Obs.	defaults	Obs.
Agriculture	5.1%	4.4%	5.7%	4.0%	6.3%
Mining	0.3%	0.5%	0.3%	0.3%	0.3%
Manufacturing	16.2%	15.6%	16.3%	15.1%	17.7%
Energy	0.8%	0.6%	0.8%	0.7%	1.0%
Construction	9.4%	14.5%	8.9%	13.6%	9.2%
Trade	39.6%	36.4%	39.5%	39.5%	40.6%
Services	25.7%	25.3%	25.5%	22.8%	22.7%
Real estate	3.9%	2.7%	3.0%	4.1%	2.2%

5. Empirical Results(2)

➢ Rotation Forest method

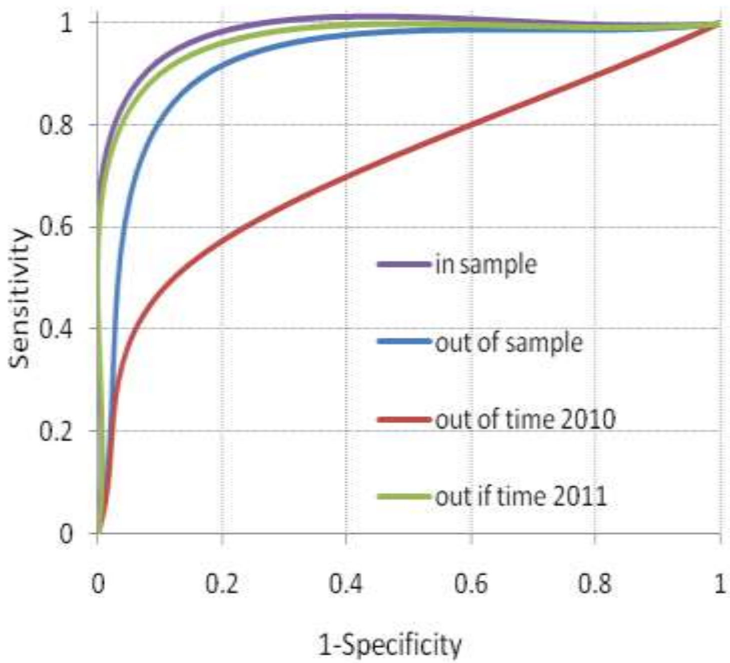


- ❖ ROC >0.9-exceptional discriminatory power
- ❖ 0.8<ROC<0.9 – excellent discriminatory power
- ❖ AR>0.8-exceptional discriminatory power
- ❖F-measure decreases due to decrease in precision -but still showing a good model
- ❖MCC- decreases but still showing an acceptable model (-1 the worst value, 0-no better then a random guess)

	ROC	F-measure	AR	MCC
In sample	0.96	0.94	0.94	0.91
Out of sample	0.95	0.88	0.87	0.88
Out of time 2010	0.88	0.59	0.86	0.29
Out of time 2011	0.87	0.44	0.85	0.46

5. Empirical results (3)

➢ Clonal Selection Classification Algorithm



❖The model works exceptional for in sample and out of sample

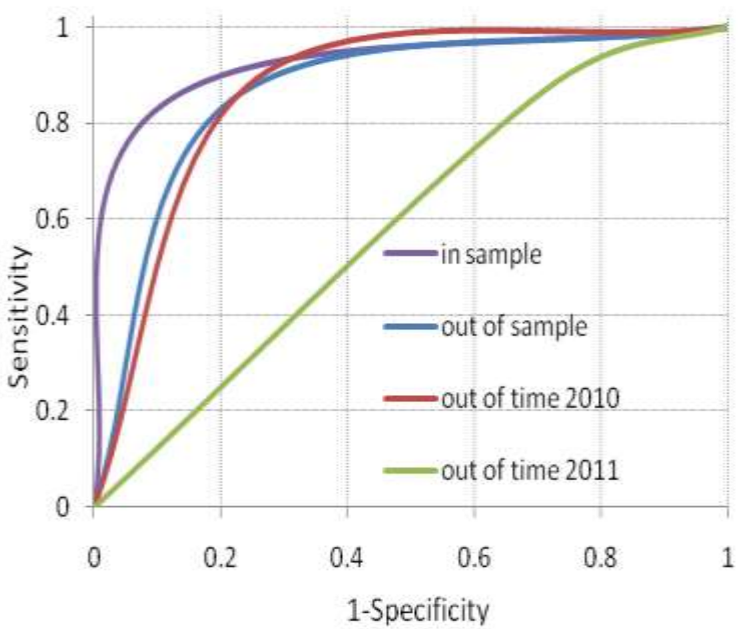
❖However, ROC curve declines in 2010 to near acceptable value due to the economic decline in 2009 (by 7.1% in real terms)

❖The discriminatory power goes up to excellent for 2011 sample. In the end of 2010 the economy declined by 1.3%

	ROC	F-measure	AR	MCC
In sample	**0.952**	**0.95**	**0.95**	**0.91**
Out of sample	**0.936**	**0.94**	**0.94**	**0.88**
Out of time 2010	**0.691**	**0.39**	**0.79**	**0.28**
Out of time 2011	**0.865**	**0.44**	**0.85**	**0.46**

5. Empirical results (4)

➢ Artificial Neural Network

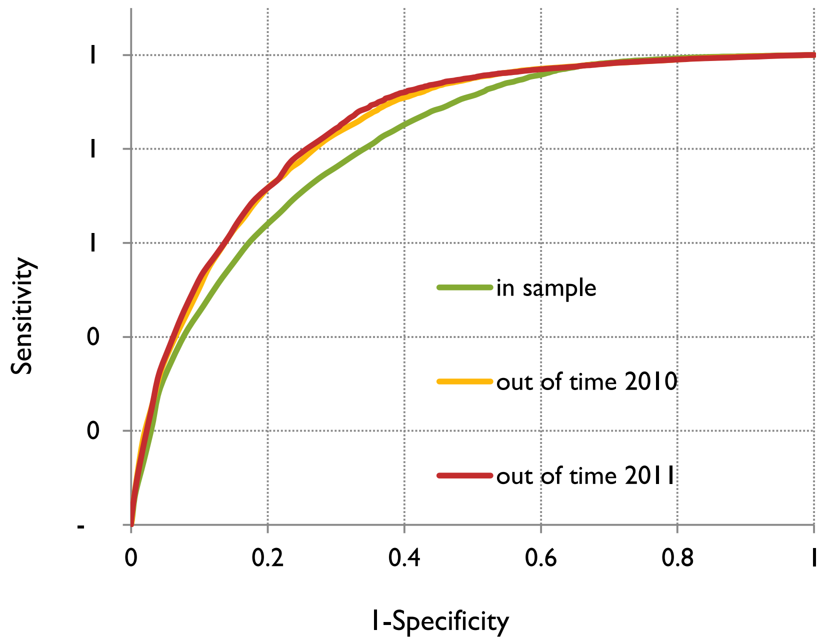


❖ Looking at the performance measures we can conclude that the model is unstable in the long run, for 2011 sample having a discriminatory power no better that a random guess.

	ROC	F-measure	AR	MCC
In sample	0.801	0.81	0.80	0.60
Out of sample	0.815	0.82	0.82	0.63
Out of time 2010	0.818	0.45	0.76	0.42
Out of time 2011	0.579	0.15	0.33	0.09

5. Empirical Results (5)

➢ Logit



❖ Logit model shows a excellent discriminatory power for in sample and out of time samples

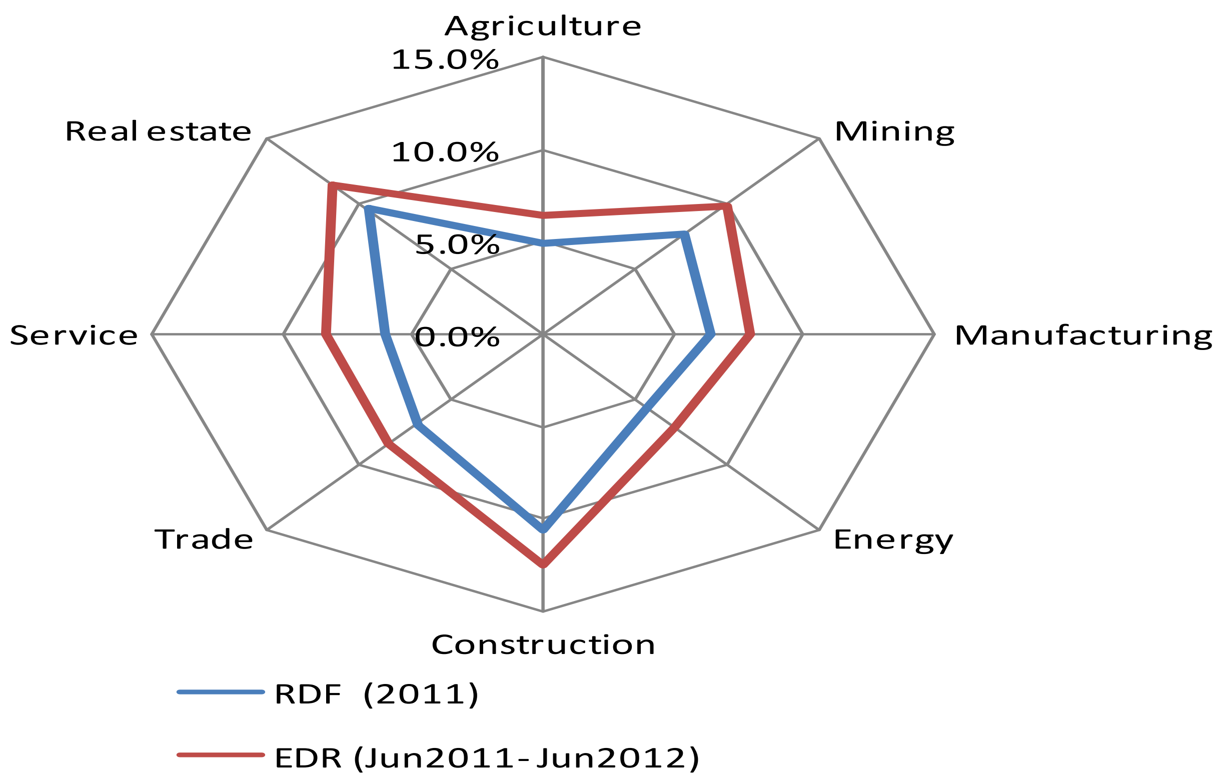
The results for out of sample are comparable with result obtained with Rotation Forest model

	ROC	F-measure	AR	MCC
In sample	0.815	0.73	0.73	0.47
Out of time 2010	0.847	0.76	0.96	0.74
Out of time 2011	0.851	0.76	0.97	0.75

5. Empirical results (6)

- Calibration

$$P(D/c_i)=\frac{P(D|c_i,s)*P(D)*(1-P(D|s)}{P(D|c_i,s)*P(D)*(1-P(D|s))+(1-P(D|c_i,s))*P(D|s)*P(D)}$$



5. Empirical results (5)

➢ Binomial Test

	PD	Default rates	p -value
Agriculture	7%	5%	1.00
Mining	10%	8%	0.95
Manufacturing	8%	6%	1.00
Energy	7%	6%	0.87
Construction	13%	11%	1.00
Trade	8%	7%	1.00
Service	8%	6%	1.00
Real estate	11%	10%	1.00
Economy	9%	7%	1.00

H0: PD is not underestimating the true probability default rate
H1: PD is underestimating the true probability default rate

6. Conclusions (1)

- The determinants of default at economy level are:
  - cost of goods sold to stock,
  - debt to equity,
  - interest payments to total assets
  - interest coverage ratio.
- Inventory turnover is a measure of the number of times inventory is sold in a time period. A low turnover rate may point deficiencies and a high turnover rate may indicate inadequate inventory level. Both may lead to a loos in business
- High debt to equity means that the company could potentially generate more earnings than it would have without this outside financing. However, the cost of this debt financing may outweigh the return that the company generates on the debt through investment and business activities and become too much for the company to handle. This can lead to bankruptcy.
- Interest expenses in total assets rate indicate the burden of a company
- Interest coverage ratio is used to determine how easily a company can pay interest expenses on its outstanding debt.

6. Conclusion (2)

- Construction and real estate sectors are the most risky, in line with economic evidence. The decline in property prices affected both sectors. Empirical evidences (WEO 2008) show that the financial crises generate a fall in property prices for 17 quarters in average.

- The relatively simple approach for modeling credit risk that was employed in this study presents both pros and cons. Advantages refer to the fact that these models are non-parametric and nonlinear models. However, most of them are black box models and is difficult to understand them.

- An improvement can be considered- using a bootstrap sample for selecting the variables, since the variables are very sensitive to the sample chosen.

- Also a stress test can be done.

- Futures work is also related with predicting the hard default, since the number of insolvent companies increased systematically in the last five years. Trade arrears, short term liquidity and receivable cash conversion days are very important factors in these cases.

7. References (1)

- Al-Enezi, J.R., M.F. Abbod and S. Alsharhan (2010), "Artificial Immune Systems-Models, Algorithms and Applications", IJRRAS 3 (2), 118-131
- Alfaro-Cid, E., K. Sharman, and A. Esparcia-Alcazar (2007), "A genetic programming approach for bankruptcy prediction using a highly unbalanced database".EvoWorkshops, Springer, 4448 of LNCS, 169–178
- Atya F. (2001), "Bankruptcy Prediction for Credit Risk Using Neural networks: A Survey and New Results",IEEE Transactions on Neural Networks, 12(4), 929-935
- Altman, E.(1968), "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", The Journal of Finance, 23, 589-609
- Beaver, W., 1966, "Financial ratios as predictors of failure", Journal of Accounting Research, 4, 71-111
- Bellovary, J., D. Giacomino, and M. Akers (2007), "A Review of Bankruptcy Prediction Studies: 1930 to Present", Journal of Financial Education, Winter 2007, 33, 1-43
- Bhaduri, A. (2009), "Credit Scoring UsingArtificial Immune System Algorithms: A comparative study", Proceeding of 2009 World Congress on Nature & Biologically Inspired Computing, 1540-1543
- BIS (1999), Credit Risk Modelling: Current Practices and Applications, Basel
- Bishop, C. (1995), "Neural Networks for Pattern Recognition", Clarendon Press,Oxford
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984),"Classification and Regression Trees", Wadsworth, Belmont, CA.
- Dasgupta,D. and L.F. Niño (2008), "Immunological Computations. Theory and Applications", CRC Press
- De Castro, L.N. (2006), "Fundamental of Natural Computing: Basic Concepts, Algorithm and Applications", Chapman & Hall, Boca Raton, Fl
- De Castro, Leandro N.; Timmis, Jonathan (2002), " Artificial Immune Systems: A New Computational Intelligence Approach", Springer. pp. 57–58

7. References (2)

- FitzPatrick, Paul J. (1932), "A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies". The Certified Public Accountant Beaver 1968. Journal of Accounting Research.(In three issues: October, 1932, p. 598-605; November, 1932, p. 656-662; December, 1932, p. 727-731.
- Guyon I. and A. Elisseeff (2003), "An Introduction to Variable and Future Selection", Journal of Machine Learning Research 3, 1157-1182
- Hekanaho, J., B. Back, K. Sere and T. Laitinen (1998), "Analyzing Bankruptcy Data with Multiple Methods", American Association for Artificial Intelligence
- Hofmeyr, S.A. and S. Forest (2000), "Architecture of an Artificial Immune System, Journal of Evolutionary Computation", 8 (4), 443-473
- Komorád K.(2002), "On Credit Scoring Estimation" Master's Thesis, Institute for Statistics and Econometrics, Humboldt University, Berlin
- Kuncheva L.I. and J.J. Rodriguez (2007), "An Experimental Study on Rotation Forest Ensembles", LNCS, Springer-Verlag, 459-468
- Leung, K., F. Cheong, C. Cheong, S. O'Farrell, and R. Tissington (2008), "Developing a Scorecard using a Simple Artificial Immune System (SAIS) Algorithm and a Real-World Unbalanced Dataset", Proceedings of the 7th International Conference on Computational Intelligence in Economics and Finance, Taiwan, 5-7
- Loukeris N. and N. Matsatsinis, (2006), "Corporate Financial Evaluation and Bankruptcy Prediction Implementing Artificial Intelligence Methods", Proceedings of the 10th WSEAS International Conference on Computers, Vouliagmeni, Athens, Greece, July 13-15
- Leung, K.,F. Cheong, and C. Cheong (2007a), "Consumer Credit Scoring using an Artificial Immune System Algorithm", Proceedings of 2007 IEEE Congress on Evolutionary Computation (CEC2007), Singapore, 25-28 September 2007
- (2008b) "Developing a Scorecard using Simple Artificial Immune System (SAIS) Algorithm and a Real World Unbalanced Dataset ", Proceedings of the 7th International Conference on Computational Intelligence in Economics and Finance, Taiwan, 5-7
- Löffler G. and P. N. Posch (2007), " Credit risk modeling using Excel and VBA", John Wiley and Sons Ltd, Chichester, West Sussex, England.
- Marsland, S. (2009), "Machine Learning. An Algorithmic Perspective", CRC Press
- Martin, D. (1977), "Early warning of bank failures: A logit regression approach" Journal of Banking and Finance 1: 249-276.
- Merwin, C. (1942), "Financing small corporations in five manufacturing industries", 1926-1936 New York: National Bureau of Economic Research.
- Ohlson, J. (1980), "Financial ratios and the probabilistic prediction of bankruptcy", Journal of Accounting Research, 18, 109-131
- Quinlan, J. R. (1979a), "Discovering rules by induction from large collections of examples", Expert Systems in the Micro Electronic Age, Edinburgh University Press, pp. 168–201.

Thank you!