Credit risk prediction for corporate clients

MSc student: Dicu Bogdan

Contents

1. Introduction
2. Literature Overview
3. Methodology
4. Empirical Study and analysis
5. Conclusions
6. References

1. Introduction

- Financial institutions and portfolio managers seek efficient methods of financial analysis to make the difference, providing higher profitability to investors.
- Given the weight loans to firms have on financial institutions assets predicting corporate credit risk is a hot topic for making correct business decision.
- Various models of Logistic regression have been developed in order to analyze corporate credit risk with classification of high precision.
- Furthermore AI approaching nonlinearity systems behavior deploys vigorous methods of Neural Networks with ensemble algorithms optimization in corporate finance.
- The objective of this research is to determine the most efficient methods in corporate credit risk analysis.

2. Literature Overview

- The past reveals a variety of statistical methods, and shows that the most commonly used are LDA (Durand 1941). and LRA (West 2000) studies indicated that both models work when the relationships between variables are linear and hence are reported to be lacking in sufficient prediction accuracy.
- The most common AI methods in corporate credit risk prediction are ANN (West 2000) and SVM (Baesens et al. 2003). West 2000, Desai et al.1996 have compared the statistical methods against ANN and claimed that ANN shows a promise when a percentage of bad loans are accurately classified.
- Ensemble methods train multiple classifiers to solve the same problems. Some of these methods that have a good rate of accuracy are: Boosting (Zhou 2009), Random subspace (Ho 1998), RS-Boosting (Wang, Ma 2011).

3. Methodology

- Romanian banks have a default rate around10%
- LRA doesn't estimate the output directly but the log-odds.

$$Log_e\left[\frac{P\left(Y=1\middle|X_1,\dots,X_p\right)}{1-P\left(Y=1\middle|X_1,\dots,X_p\right)}\right]=Log_e\left[\frac{\pi}{1-\pi}\right]$$

$$=\alpha+\beta_1X_1+\cdots+\beta_pX_p=\alpha+\sum_{j=1}^{p}\beta_jX_j$$

- First method applied is Multi Nominal Logistic Regression where the Quasi-Newton Method is implemented to seek the optimized values for beta, aiming to minimize the log-likelihood.

$$L=-\sum_{i=1}^{n}[\sum_{j=1}^{k-1}(Y_{ij}\times\ln(P_j(X_i)))+(1-(\sum_{j=1}^{k-1}Y_{ij})\times\ln(1-\sum_{j=1}^{k-1}P_j(X_i))]+ridge\times B^2$$

3. Methodology – Logistic Regression

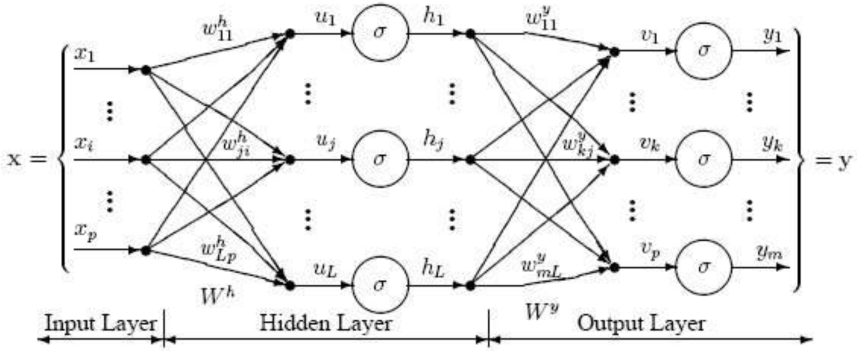
- Simple Logistic
- I used the Simple Logistic function as a Linear Logistic Regression that uses a LogitBoost algorithm for implementing ordinal regression functions as base learners to fit the logistic models and to estimate α and β.
- LogitBoost is form of Additive Logistic Regression that boosts schemes for numeric prediction being able to create a combined classifier that predicts a categorical class. The classification process is implemented through a regression scheme as the base learner in multi-class problems.

$$L = \prod_{i=1}^{n} P\left(Y_i \middle| X_{i1}, \dots, X_{ip}\right) = \prod_{i=1}^{n} \left[\left(\frac{e^{\alpha + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\alpha + \sum_{j=1}^{p} \beta_j X_j}} \right)^{Y_i} \times \left(\frac{1}{1 + e^{\alpha + \sum_{j=1}^{p} \beta_j X_j}} \right)^{1 - Y_i} \right]$$

- Another function used is Bayesian Logistic Regression. Bayesian network models are widely used for discriminative prediction tasks such as classification. The parameters of such models are often determined using 'unsupervised' methods such as maximization of the joint likelihood.

3. Methodology – ANN

- For ANN classification I used the Multi Layer Perceptron algorithm model that maps sets of input data onto a set of appropriate output. Simple Perceptrons consist of a layer of input neurons, coupled with a layer of output neurons, and a single layer of weights between them.
- The development of the network relies heavily on the qualitative data that are solicited from the applicants to specify the interactions among all characteristics. The attributes are linearly combined and by the activation of the sigmoid function they become subject to a non-linear transformation, then fed as input to the next layer for similar manipulation.



- MLP uses a supervised learning technique called Back Propagation in order to train the network

3. Methodology – Ensemble Methods

- Boosting is a kind of ensemble methods which produces a strong classifier that is capable of making very accurate predictions by combining rough and moderately inaccurate base classifiers.
- The method is applied using AdaBoost algorithm with Decision Stump as base classifier .
- The random subspace method is an ensemble construction technique, where the training dataset is also modified as in Bagging. However, this modification is performed in the feature space (rather than instance space). For computing I used as base classifier Random Forest.
- Combined decision of such base classifiers may be superior to a single classifier constructed on the original training dataset in the complete feature space.
- RS-Boosting introduces random subspace strategy into each boosting iteration. The base classifiers were trained using random subspace method. After that, these base classifiers were used to reweight the instances. With integrating boosting and random subspace, RS-Boosting combines the two different partitioning methods above. As there are two different ensemble strategies encouraging diversity in RS-Boosting, it would be advantageous to get more accuracy than Boosting and Random Subspace individually.

4. Empirical Study and analysis

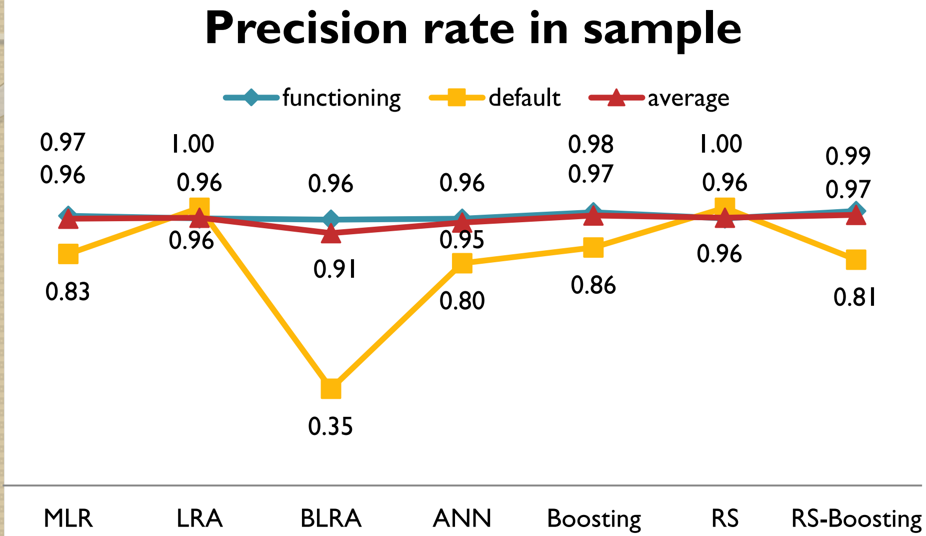
- As training set for these models I used the financial data for 186 companies and have tested them with data drawn from potential clients.

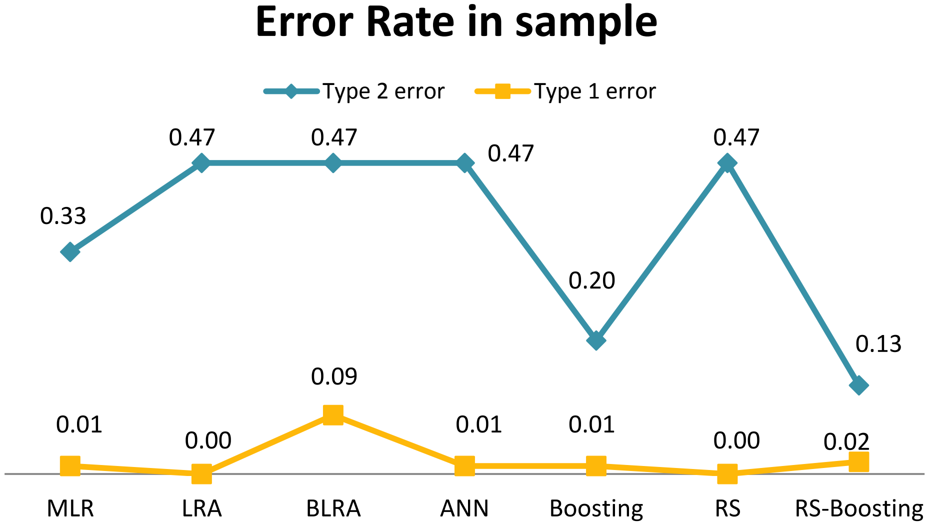
- To identify the best method for classifying corporate credit risk we study the results in sample and out of sample, using 2000 potential clients that have credit at other financial institutions.

- I compared type 1 error, type 2 error, precision and ROC area.

- For this study I used Weka 3, a data mining software in Java, a collection of machine learning algorithms. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization.

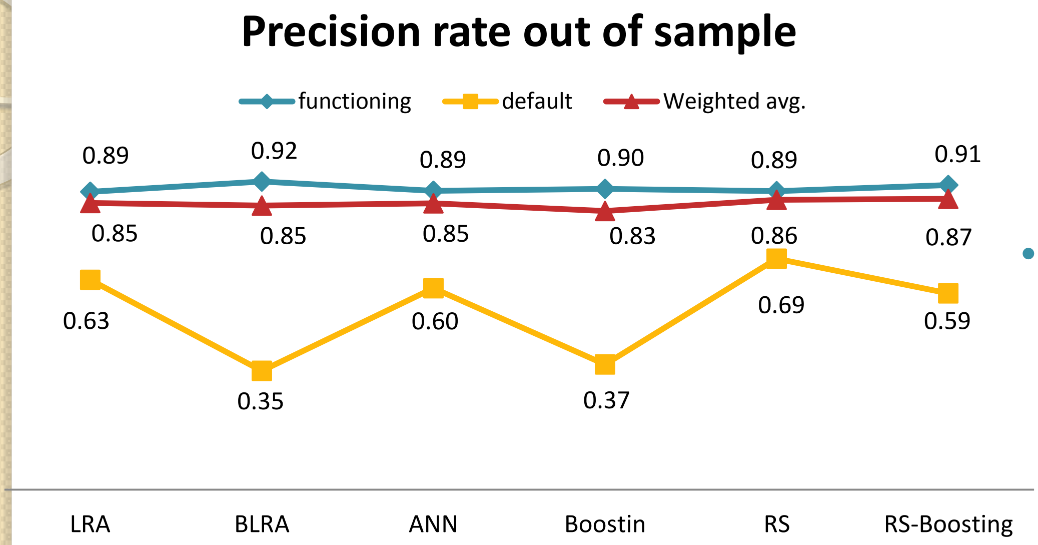
4. Empirical Study and analysis – In Sample Analysis

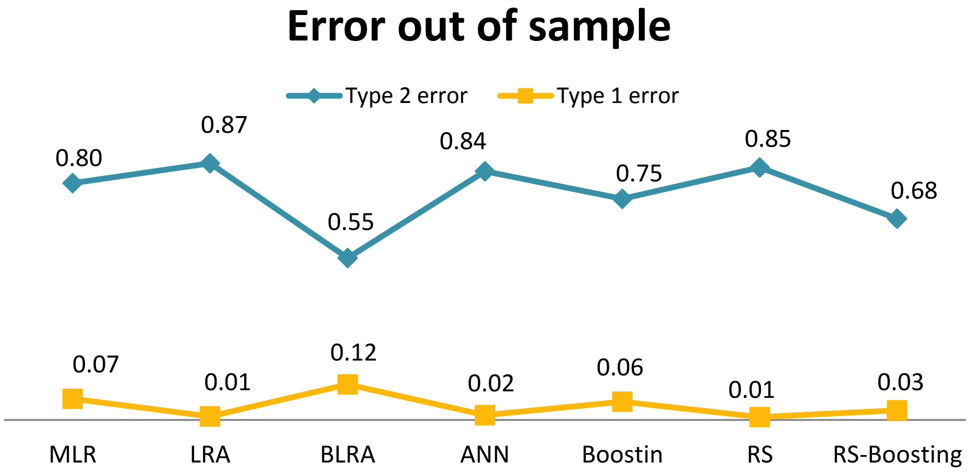




- PR: Due to the low default rate we have a high precision rate for all models, over 80%. The best model is LRA which acts like RS with 100% in examples which truly have class functioning among all classified as functioning.
- Error: Type 2 error indicates that RS-Boosting is the most accurate method in sample tests.

4. Empirical Study and analysis – Out of Sample analysis case 1

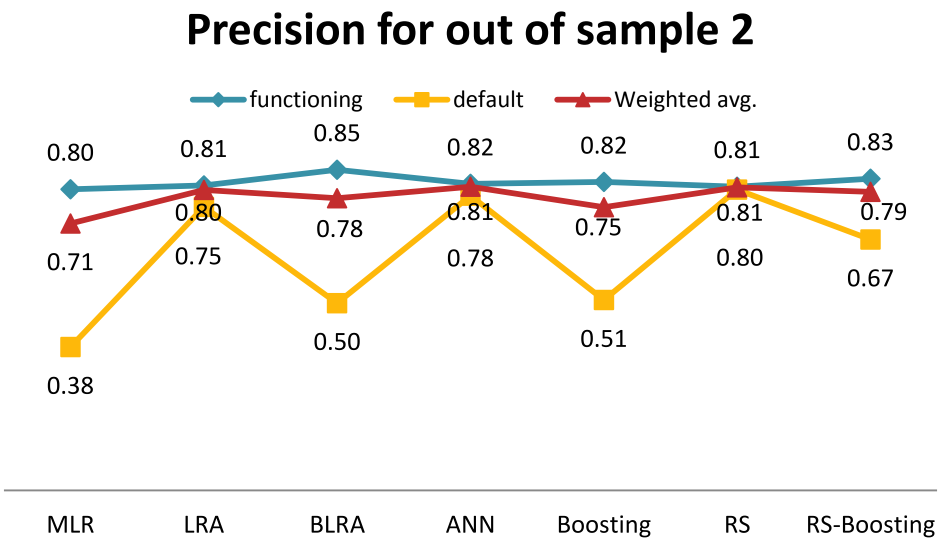


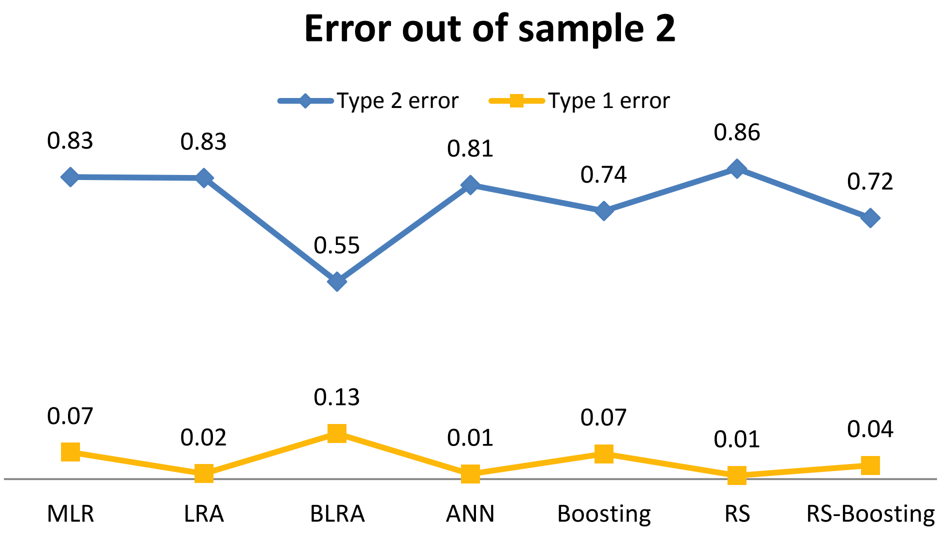


- PR: same as the in sample test due to the low default rate we have a high average precision rate for all models, over 80%. RS have 68.8% rate of companies that truly default among all those that were classified to default.

- Error: All methods have a high vale for type 2 error, lowest error reached 0.547 (bayesian LR).

4. Empirical Study and analysis – Out of Sample analysis case 2


Precision for out of sample 2


Error out of sample 2

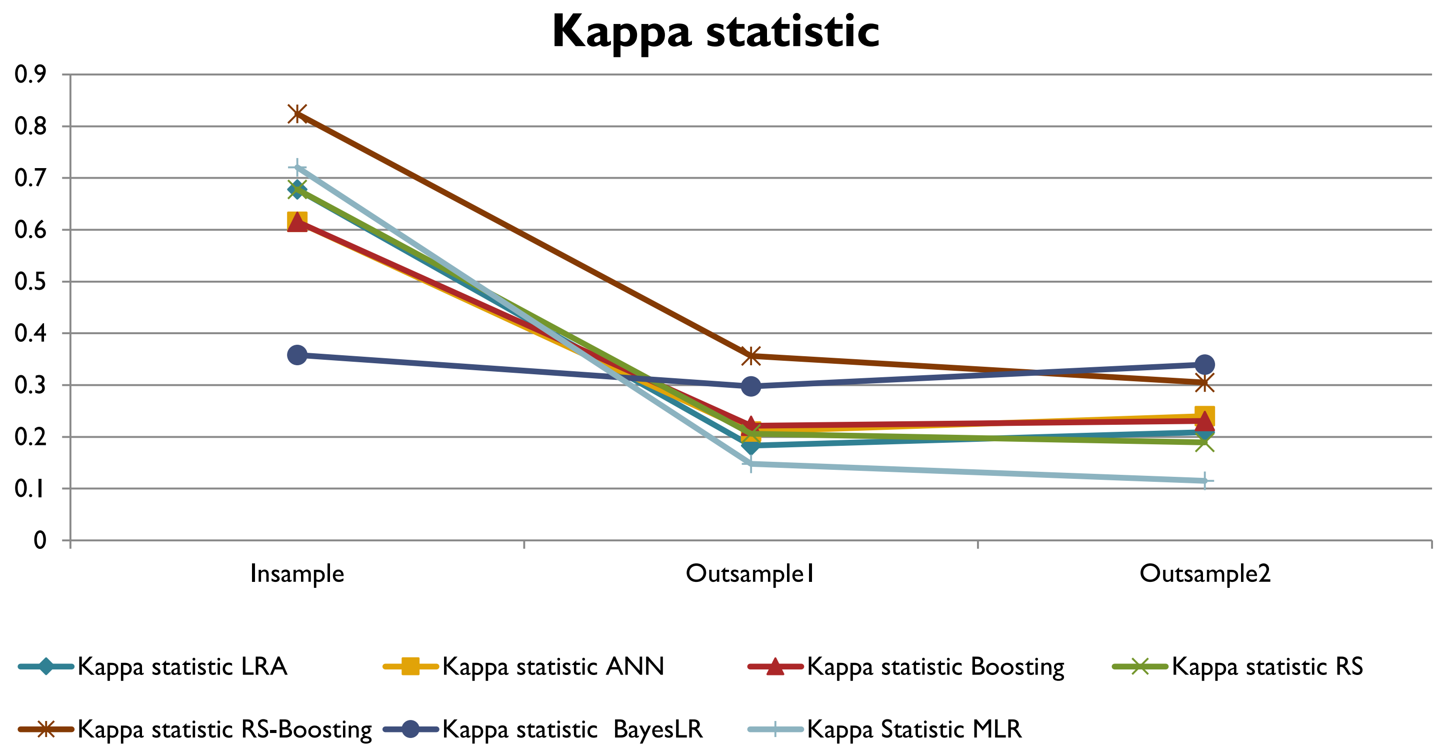
- PR: The second test confirm the loose of accuracy from in sample test. We can see that some models compensate the loose of precision rate in non-defaulting companies with a gain in the precision rate for correctly identify the default.
- Error: After analyzing the errors in out of sample we can say that Bayesian LRA outclass the other models in reducing Type 2 error, but have the highest value for Type I error. This could indicate that the method have the tendency to be more restrictive, and that is transforming in loosing good clients. The second method with the best performance in reducing Type 2 error is RS-Boosting, that also manage to maintain a low rate for Type I error.

4. Empirical Study and analysis – Preserving Accuracy

- For better analyzing the capacity of the methods to keep their accuracy in out of sample data, I have chosen kappa statistic and ROC area.
- The kappa statistic measures the agreement of prediction with the true class
- The area under an ROC Curve measures discrimination, that is, the ability of the test to correctly classify those companies with and without the default.
- ROC area interpretation

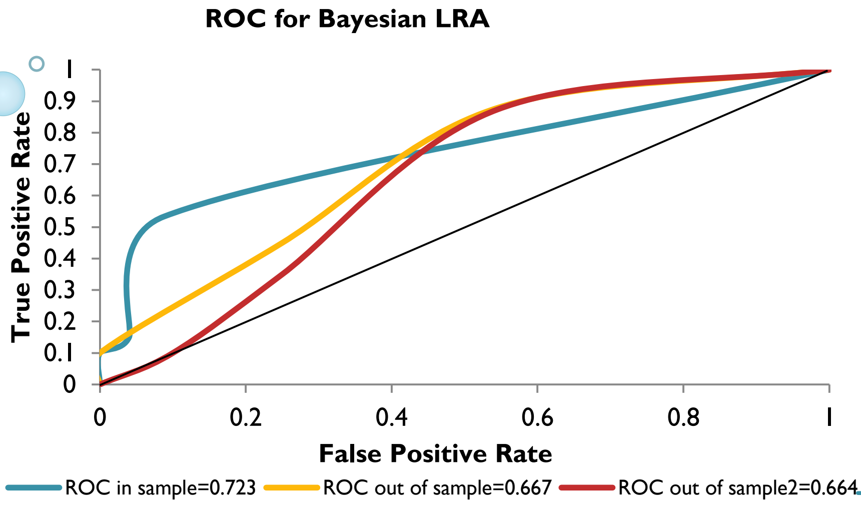
a. .90-1 = excellent (A)
b. .80-.90 = good (B)
c. .70-.80 = fair (C)
d. .60-.70 = poor (D)
e. .50-.60 = fail (F)

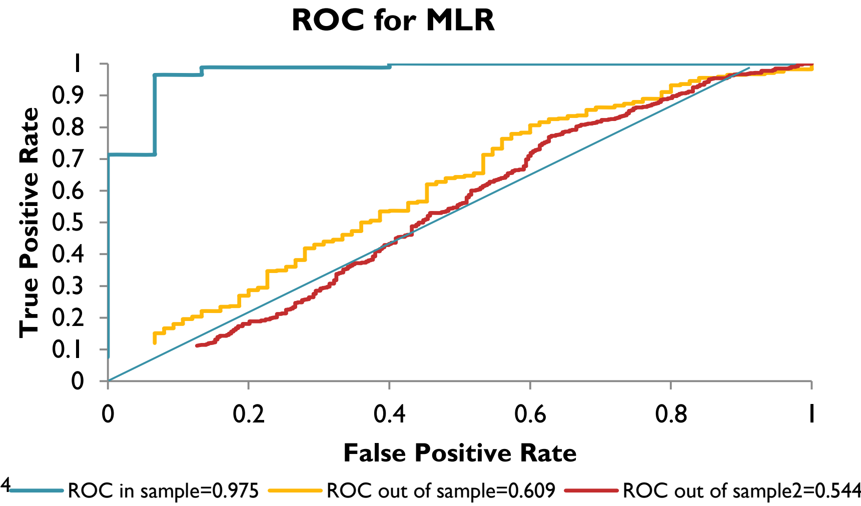
4. Empirical Study and analysis – Preserving Accuracy Tests

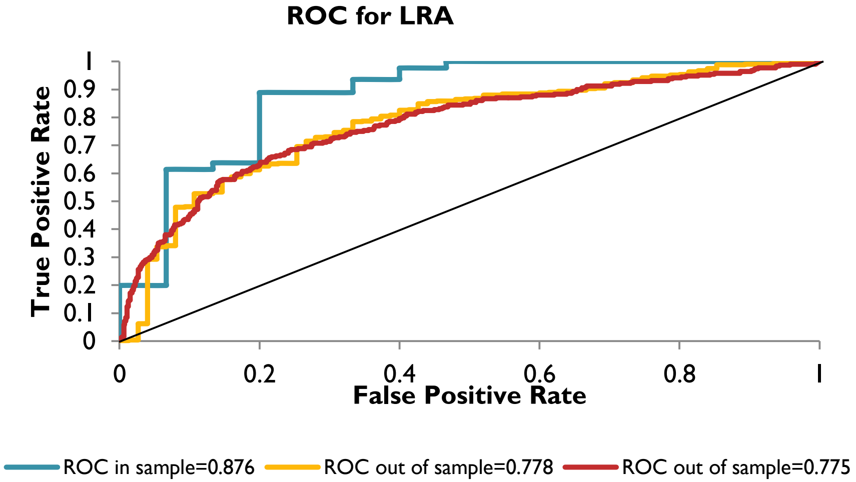


Kappa statistic indicates that the best model in preserving its accuracy is Bayes LR but with a value below 0.4. The test approves that the RS-Boosting ensemble has the best performance in the test sample.There is no statistical significant different performance from Bayes LR in Out of Sample tests.

ROC For Statistical Models

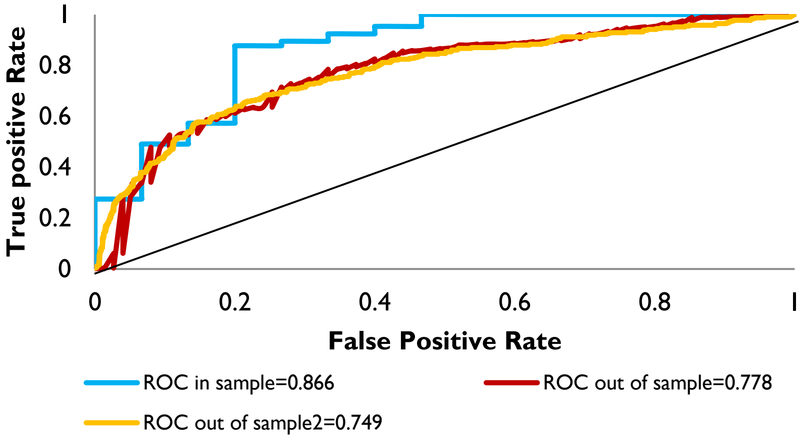

ROC for Bayesian LRA


ROC for MLR

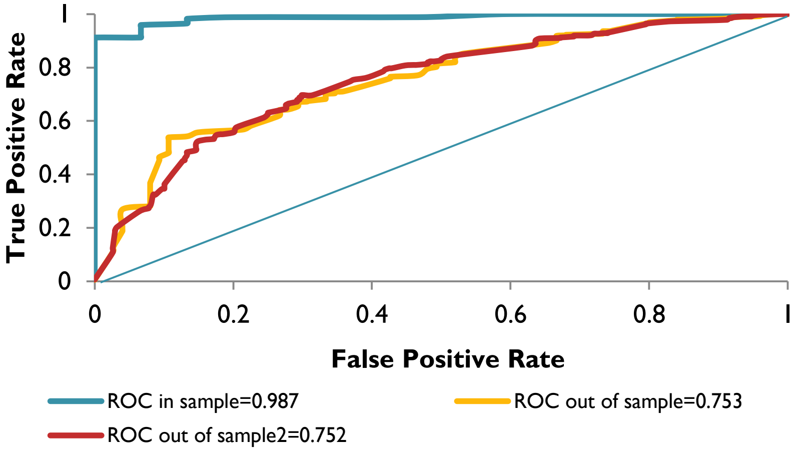

ROC for LRA

ROC For AI Models

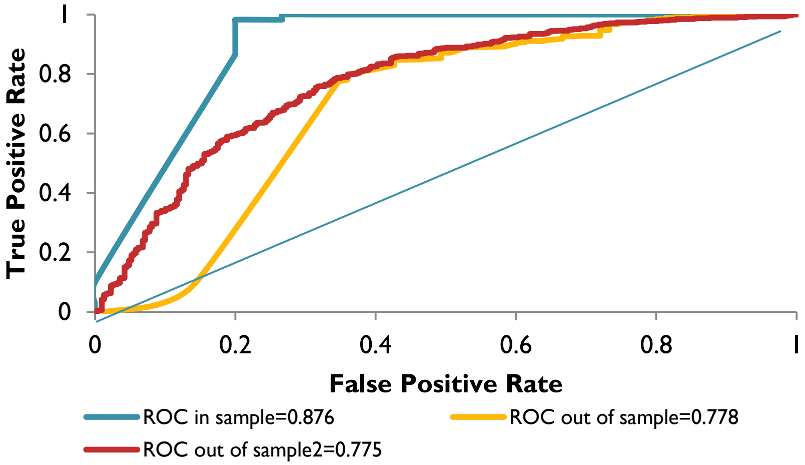
ROC for ANN



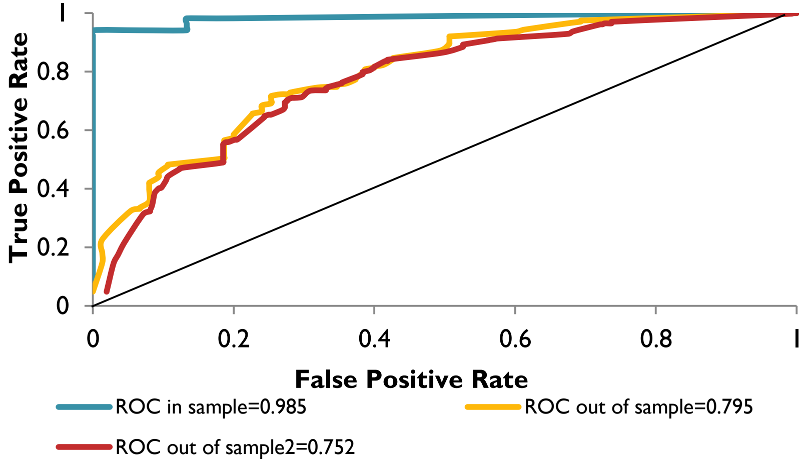
ROC for Boosting



ROC for Random subspace



ROC for RS Boosting



- By comparing the ROC areas we find that even AI methods reach excellent efficiency in sample test they are acting fair, as LRA method, in out of sample tests., turning out not to be statistically different, however.

5. Conclusions

- A slight improvement in the accuracy of identifying the default might translate into significant future savings.
- Traditional methods perform well only when essential assumptions are satisfied.
- In contrast, AI methods do not require to make assumptions of the underlying relationships between input and output.
- The AI methods can lose precision when are tested out of sample. Building the model on a large data set can prevent that from happening.

5. Conclusions

RS-Boosting analyze

RS-Boosting in sample		RS-Boosting out of sample		RS-Boosting out of sample 2		
TP	FP	TP	FP	TP	FP	Class
0.982	0.133	0.967	0.68	0.962	0.721	functioning
0.867	0.018	0.32	0.033	0.279	0.038	default

The study shows that if a financial institution would apply RS-Boosting method in order to decide which companies receive credit, the default rate might drop with 30% and functioning credit could lower with 3.5%.

With a rate of default at 5 years of 10% this will turnout in future savings from the reduction of recovery cost.

For future research AI methods should be tested with more exploration of credit data structures in order further validate the conclusions of this studies.

6. References

- [1] Basel Committee on Banking Supervision, Principles for the Management of Credit Risk, Basel, Switzerland, 2000, p. 11, http://www.bis.org/publ/bcbs75.pdf?noframes=1.
- [2] Basel Committee on Banking Supervision, International Convergence of Capital Management and Capital standards: A Revised Framework, Basel, Switzerland, 2005, p. 15, http://www.bis.org/publ/bcbs118.htm.
- [3] Loukeris N. (2007), "Corporate Financial Analysis with efficient Logistic Regressions and Hybrids of Neuro-Genetic networks", Wivenhoe park, CO4 3SQ, UK
- [4] C. Mitchell Dayton (1992), "Logistic Regression Analysis", Department of Measurement, Statistics & Evaluation, University of Maryland
- [5] Adnan Khashman(2010), "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes", Expert Systems with Applications xxx
- [4] Thomas, L. C. (2000), "A survey of credit and behavioral scoring: Forecasting financial risks of lending to customers", International Journal of Forecasting, 16(2), 149–172.
- [6] Lean Yu, Shouyang Wang and Kin Keung (2008), "Lai Credit risk assessment with a multistage neural network ensemble learning approach", Expert Systems with Applications, 34, 1434–1444
- [7] Khashman A (2011), "Credit risk evaluation using neural networks: Emotional versus conventional", Models Applied Soft Computing 11 5477–5484
- [8] Nicolas Garcıa-Pedrajas and Domingo Ortiz-Boyer, (2008), "Boosting random subspace method", Neural Networks, 21, 1344-1362
- [9] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald and David Scuse, (2011), "WEKA Manual for Version 3-6-6",
- [10] Wang G. and Jian Ma, (2011), "Study of corporate credit risk prediction based on integrating boosting and random subspace", Expert Systems with Applications, 38, 13871–13878