Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC – Bank Indonesia International Workshop and Seminar on *"Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data"*

Bali, Indonesia, 23-26 July 2018

A robust machine learning approach for credit risk analysis of large loan-level datasets using deep learning and extreme gradient boosting[1]

Anastasios Petropoulos, Vasilis Siakoulis,
Evaggelos Stavroulakis and Aristotelis Klamargias,

Bank of Greece

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting

Anastasios Petropoulos

Vasilis Siakoulis

Evaggelos Stavroulakis

Aristotelis Klamargias

Abstract

In the aftermath of global financial crisis of 2007–2008, central banks have put forward data statistics initiatives in order to boost their supervisory and monetary policy functions which will lead to central banks possessing big databases increasing the need for robust data mining processes and financial statistical modelling to support more informed decision making. Conventional econometric methods fail to capture efficiently the information contained in the full spectrum of the datasets. To address these challenges, in this work we investigate the analysis of a corporate credit loans big dataset using cutting edge machine learning techniques and deep learning neural networks.

The novelty of our approach lies in the combination of a data mining algorithms that aim to reduce dimensionality in the data and increase accuracy in predicting the future behaviour of corporate loans, to facilitate a more effective micro and macro supervision of credit risk in the Greek banking system. Our analysis is based on a large dataset of loan level data, spanning a 10 year period of the Greek economy with the purpose of performing obligor credit quality classification and quantification of Probability of Default under a through the cycle setup.

We perform extensive comparisons of the classification and forecasting accuracy of the proposed methods, using a 3-years' period out-of-time sample. Our experimental results are benchmarked against other traditional methods, like logistic regression and discriminant analysis methods, yielding significantly superior performance. In the final stage of our analysis, a robust through the cycle financial credit rating is developed which can offer a proactive monitoring mechanism of the credit risk dynamics in a financial system. Finally the methodological framework introduced can support a more in depth analysis of database initiatives like ECB AnaCredit.

Keywords: Credit Risk, Neural Networks, Deep Learning, Extreme Gradient Boosting.

JEL classification: G24, C38, C45, C55

Contents

1. Introduction

In the aftermath of global financial crisis of 2007–2008, central banks have put forward data statistics initiatives in order to boost their supervisory and monetary policy functions. In the coming years central banks will possess big databases increasing the need for robust data mining processes and financial statistical modelling to support more informed decision making. Under this era, central banks should simultaneously enrich their statistical techniques in order to accommodate the increase availability of data, and to exploit all possible dimensions of information collected. Big financial datasets usually pose significant statistical challenges because they are characterized by increased noise, heavy-tailed distributions, nonlinear patterns and temporal dependencies. Conventional econometric methods fail to capture efficiently the information contained in the full spectrum of the datasets. To address these challenges, in this work we focus on the analysis of a corporate credit loans big dataset using cutting edge machine learning techniques, like Extreme Gradient Boosting (XGBoost) and deep learning neural networks (MXNET).

The novelty of our approach lies in the combination of a data mining algorithms that aim to reduce dimensionality in the data and increase accuracy in predicting the future behaviour of corporate loans, to facilitate a more effective micro and macro supervision of credit risk profile in the Greek banking system. Our analysis is based on a large dataset of loan level data, spanning in a 12 year period of the Greek economy. Data are collected by Bank of Greece for statistical and banking supervision activities. The dataset is comprised of more than 200k records of corporate and SME loans of the Greek banking system, with information related to the one-year-ahead delinquency behaviour. Features collected for analysis include companies' historical data of properly selected set of financial ratios, along with historical data of macro variables relevant to the Greek economy. To ameliorate the issue of high dimensionality in the data we used an advanced machine learning algorithm, called Boruta, to perform the variable importance selection in a multivariate holistic approach. Extreme gradient Boosting and Deep neural networks are used for performing obligor credit quality classification and quantification of Probability of Default under a through the cycle setup.

We perform extensive comparisons of the classification and forecasting accuracy of the proposed methods, using a 3-years' period out-of-time sample. Our experimental results are benchmarked against other traditional methods, like logistic regression, and discriminant analysis methods, yielding significantly superior performance. Furthermore, it is also found that the performance of deep neural-network models depend on the choice of activation function, the number and structure of the hidden layers, and the inclusion of dropout and batch normalization layers signalling increase flexibility in addressing complex datasets and potential increased classification capabilities. In the final stage of our analysis, a robust through the cycle financial credit rating scale is developed which can accommodate the efficient benchmarking of A-IRB models and offer a proactive mechanism of the credit risk dynamics in a financial system. In addition, it can support top down stress testing exercises offering a more risk sensitive and accurate forecasting framework.

In all the methodological framework introduced can support a more in depth analysis of database initiatives like ECB AnaCredit[1].

2. Literature Review

In the domain of credit risk modelling, more accurate and robust systems to drive expert decisions have been employed in recent years, exploring new statistical techniques especially from the field of machine and deep learning. In the last decades, a plethora of approaches has been developed to address the problem of modelling the credit quality of a company, using both quantitative and qualitative information.

Several studies have explored the utility of probit models (Mizen and Tsoukas, 2012) and linear regression models (Avery, et al., 2004). These models however, suffer from their clear inability to capture non-linear dynamics, which are prevalent in financial ratio data (Petr and Gurný, 2013). Another class of statistical models used for credit rating is hazard rate models. These models extend the time horizon of a rating system, by looking at the probability of default during the life cycle of the examined loan or portfolio (Chava and Jarrow, 2004 & Shumway, 2001).

A Bayesian inference-based analogous to support vector machines (SVMs) (Vapnik, 1998), namely Gaussian processes, has been considered by Huang (2011). A drawback of this approach is its high computational complexity, which is cubic to the number of available data points, combined with the assumption of normally distributed data. Yeh et al. (2012) applied Random Forests (Breiman, 2001) in credit corporate rating determination, Zhao et al, (2015) employed feed forward neural networks in the same domain whereas Petropoulos et al (2016) made use of Student's-t hidden Markov models

Addo et al. (2018) focus on credit risk scoring where they examine the impact of the choice of different machine learning and deep learning models in the identification of defaults of enterprises. They also study the stability of these models relative to a choice of subset of variables selected by the models. More specifically, they build binary classifiers based on machine and deep learning models on real data in predicting loan default probability. The top features from these models are selected and then used for testing the stability of binary classifiers by comparing their performance on separate data. They observe that the tree-based models are more stable than the models based on multilayer artificial neural networks.

Khandani et al. (2010) apply machine learning techniques (generalized classification and regression trees (CART)-like algorithm (Breiman et al., 1984)) to construct nonlinear nonparametric forecasting models of consumer credit risk. They combine customer transactions and credit bureau data from January 2005 to April 2009 for a sample of a major commercial bank's customers; thus, they are able to

[1] https://www.ecb.europa.eu/stats/money/aggregates/anacredit/shared/pdf/explanatorynoteanacreditregulation.en.pdf

construct out-of-sample forecasts that significantly improve the classification rates of credit-card-holder delinquencies and defaults.

Butaru et al. (2016) use account-level credit card data from six major commercial banks from January 2009 to December 2013; they combine consumer tradeline, credit bureau, and macroeconomic variables to predict delinquency, employing C4.5 decision trees, logistic regression and random forests. They find substantial heterogeneity in risk factors, sensitivities, and predictability of delinquency across banks, implying that no single model applies to all six institutions. The results suggest the need for a more customized approached to the supervision and regulation of financial institutions, in which capital ratios, loss reserves, and other parameters are specified individually for each institution according to its credit risk model exposures and forecasts.

Galindo and Tamayo (2000) test CART decision-tree models on mortgage-loan data to detect defaults. They also compare their results to the Neural Networks (ANN), the k-nearest neighbor (KNN) and probit models, showing that CART decision-tree models provide the best estimation. Huang et al. (2004) provides a survey of corporate credit rating models showing that Artificial Intelligence (AI) methods achieve better performance than traditional statistical methods. The article introduces a relatively new machine learning technique, support vector machines (SVM), to the problem in attempt to provide a model with better explanatory power. They used backpropagation neural network (BNN) as a benchmark and obtained prediction accuracy around 80% for both BNN and SVM methods for the United States and Taiwan markets.

Motivated from all the aforementioned research endeavours we revisit the issue of credit risk modelling following a different venue. We explore two state of the art techniques namely Extreme Gradient Boosting (XGBoost) and deep learning neural networks in order to obtain at first maximum information gain from a loan level large size data source and secondly to create a useful, from a regulatory scope, credit rating grade system measuring credit risk in supervised banks portfolios.

3. Data collection processing and variable selection

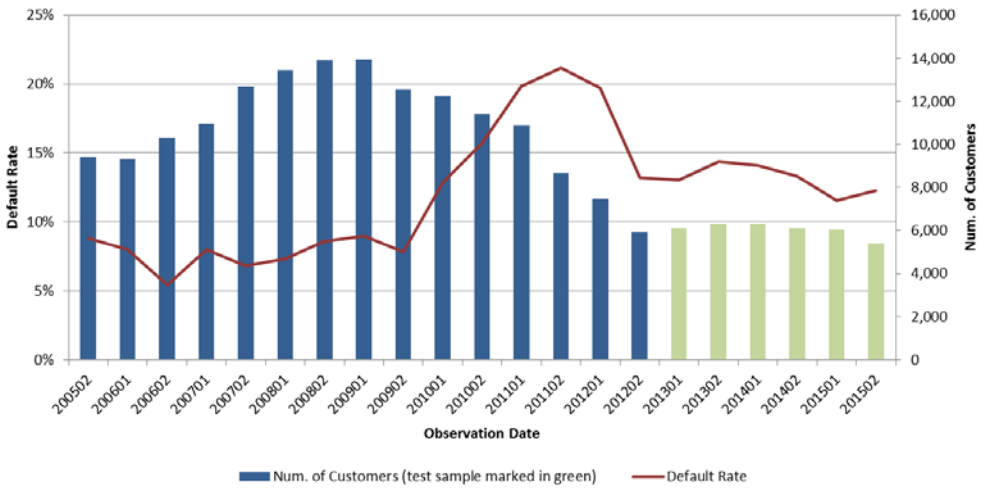
We have collected loan level information on Corporate and SME loans of the Greek banking system, from the supervisory database of the Central Bank of Greece. The data collection procedure excludes special cases of obligors from the financial sector, including banks, insurance, leasing, and factoring companies, due to the very unique nature of their business models, which deviate quite a lot from the business models of commercial companies.

The adopted definition of a default event in this dataset is in line with the rules of the Credit Risk Regulation (CRR). Specifically, a loan is flagged as delinquent if it is either 90 days past due or it gets rated as delinquent based on each bank's internal rating rules. At each observation snapshot, all performing loans are considered. At the end of the 12-month observation period, each obligor is categorized as either good (i.e., performing) or bad (i.e., non-performing). At the end the dependent variable in our dataset is a binary indicator, with the value of

one flagging a default event (i.e., the obligor is categorized as bad at the end of the 12-month observation period).

The dataset covers the 2005-2015 period; a 10 years' period with semi-annual information (i.e. semi-annual snapshots). The selected time period, seems to approximate a full economic cycle, in terms of the default rate evolution. Figure 1, shows the number of customers included in each snapshot and the corresponding default rate. The overall dataset includes approximately 200.000 unique customers, resulting in even more records on a facility level, as one customer may have more than one facility in one or more than one banks with different risk characteristics (for example the average facility number in the credit risk supervisory database reaches 120.000 records per quarter). It is clear that the default rates have elevated in the most recent period, i.e. from the second half of 2010 and onwards, compared to the older observations, i.e. up to 2010. Specifically, the default rates follow an increasing trend in the 2010-2011 periods, where they peak at 21.2% in the second half of third quarter of 2011. Thereafter, they follow a decreasing trend. The default rates seem to have flattened out since 2013, remaining stable at around 12%-13%.

Figure 1: Greek banking system business portfolio metric evolution



In order to perform the modelling and prediction methodology, our approach incorporates the companies' 5 year lagged historical data of properly selected set of financial ratios along with 10 quarters lagged historical macro variables relevant to the Greek economy (both shown analytically in the Appendix). This is based on the assumption that financial ratios carry all the information necessary to describe and predict the internal state of a company, providing adequate insights on how profitable an examined company is, what the trends are.

The combined dataset of lagged financial ratios and macro variables along with some data transformations, led to a set of 354 predictor variables (distinct time-series) as potential candidates for our modelling procedures. Fitting a machine learning model to such a huge number of independent variables (relative to the size of the dataset) is doomed to suffer from the so-called curse of dimensionality problem, whereby the fitted classifier may seem to yield very good performance in the training dataset, but it turns out to generalize very poorly, yielding a

catastrophically low performance outcome in the test data. Thus, to ensure a good performance outcome for our model, we need to implement a robust independent variable (feature) selection stage, so as to limit the number of used features to the absolutely necessary. Besides, apart from increasing the generalization capabilities of the fitted models, such a reduction is also important for increasing the computational efficiency of the explored machine learning algorithms.

We employ the Boruta algorithm to independently assign importance to the available features. The Boruta algorithm is based on a postulated Random Forest model. Based on the inferences of this Random Forest, features are removed from the training set, and model training is performed anew. Boruta infers the importance of each independent variable (feature) in the obtained predictive outcomes by creating shadow features. Specifically, the algorithm performs the following steps: First, it adds randomness to the given dataset by creating shuffled copies of all features (shadow features). Then, it fits a Random Forest on the extended dataset and evaluates the importance of each feature. In every iteration, it checks whether a real feature has a higher importance than the best of its shadow features, and constantly removes features which are deemed unimportant. The comparison is done based on Z score. The algorithm stops when all features are classified as important or are rejected as noise. In our study, we employ the Boruta Package, provided by the R programming language, to implement variable selection. In this way, all features relevant to both dependent variables are selected based on error minimization for the fitted Random Forest models, in each iterative step of the algorithm. From the Boruta variable selection process 65 variables out of 354 candidates were selected for the moment development alleviating dimensionality issues. The so-obtained dataset was split into three parts:

- An in-sample train dataset, comprising data pertaining to the 70% of the examined companies, obtained over the observation period 2005-2012, which was used for model development.
- An in-sample test dataset, comprising the data pertaining to the rest 30% of the companies for the period 2005-2012 which was employed for assessing the parameter calibration.
- An out-of-time dataset that comprises all the data pertaining to the observation period of year 2013-2015 (marked in green in Figure 1) which was employed for validation purpose and testing the generalization capacity of all candidate models.

4. Model Development

Given the extended number of employed predictors and the large scale dataset employed we resort to a methodology from the general domain of Machine Learning techniques called Extreme Gradient Boosting (henceforth XGBoost) and a Deep Learning Technique used to train, and deploy deep neural networks (MXNET). The supervisory motivation for employing such types of methodologies rests on the availability of large scale supervisory data, which are expected to further augment in the near future (e.g. ECB's AnaCredit project), upon which the capability of pattern

detection by traditional statistical methodologies is limited due to multicollinearity, dimensionality and convergence issues.

The XGBoost is a boosting tree algorithm that is an enhancement over tree bagging methodologies, such as Random Forests (Breiman 2000), which have gained significant ground and are frequently used in many machine learning applications across various fields of the academic community. The basic philosophy of bagging is based on combining three concepts: i) Creation of multiple datasets; ii) building of multiple trees and iii) bootstrap aggregation or bagging. It adopts a divide-and-conquer approach to capture non-linearities in the data and perform pattern recognition. Its core principle is that a group of "weak learners" combined, can form a "strong predictor" model.

For example in the case of Random Forests the algorithm is based on the random generation of a number of classification trees which is the so called Forest. Tree generation is randomly performed in an iterative mode so in each iteration, a random subsample of the included features is selected from the dataset by means of bootstrap. Then, a tree is generated from using the CART algorithm which contains a relatively limited number of features. After constructing the random trees, prediction is performed using Bagging. Each input is entered through each decision tree in the forest and produces a forecast. Then, all predictions of each tree are aggregated either as a (weighted) average or majority vote, depending whether the underlying problem is a regression or a classification, respectively.

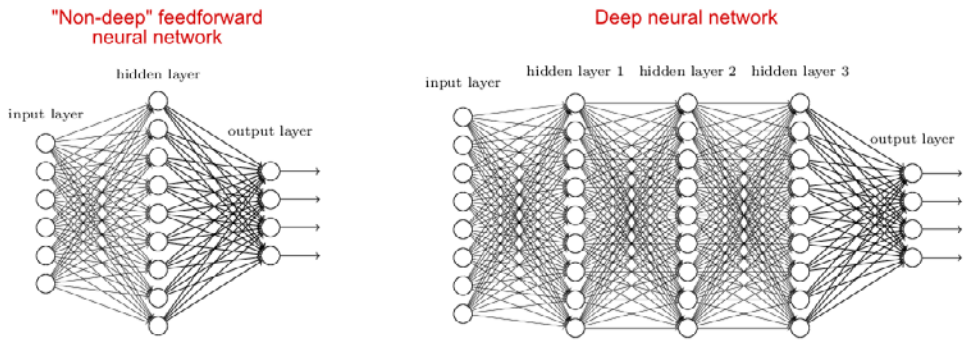
Gradient Boosting trees model is proposed by Friedman (1999) and has the advantage of reducing both variance and bias. It reduces variance because multiple models are used (bagging), whereas it additionally reduces bias in training the subsequent model by telling him what errors the previous models made (boosting). In gradient boosting each subsequent model is trained using the residuals (the difference between the predicted and true values) of previous models. XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting algorithm, offering increased efficiency, accuracy and scalability over simple bagging algorithms. It supports fitting various kinds of objective functions, including regression, classification and ranking. XGBoost offers increased flexibility, since optimization is performed on an extended set of hyperparameters, while it fully supports online training.

We developed XGBoost in the context of our study by utilizing the XGBoost R package. We performed an extensive cross-validation procedure to select a series of entailed hyper parameters, including the maximum depth of trees generated, the minimum leaf nodes size to perform a split, and the size of sub-sampling for building the classification trees and the variables considered in each split. The objective function used for the current problem was logistic due to the binary nature of the dependent variable while the area under the curve (AUROC) metric was used for model selection in the context of cross-validation. The AUROC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In practice, the value of AUROC varies between 0.5 and 1, with a value above 0.8 denoting a very good performance of the algorithm. To reduce overfitting tendencies, we tuned the γ hyper parameter, which controls model complexity by imposing the requirements that node splits should yield a minimum reduction in the loss function, as well as the α and λ hyper parameters, which perform regularization of model weights similar to shrinkage techniques such as LASSO.

Besides Extreme Gradient Boosting we implement also a Deep Neural Network (henceforth DNN) to address the issue of corporate default forecast. Deep learning has been an active field of research in the recent years, as it has achieved significant breakthroughs in the fields of computer vision and language understanding. In particular they have been extremely successful in as diverse time-series modelling tasks as machine translation (Cho et al., 2014, Tu et al., 2016.), machine summarization (See et al., 2017) and recommendation engines (Quadrana et al., 2017). However, their application in the field of finance is rather limited. Specifically, our paper constitutes one of the first works presented in the literature that considers application of deep learning to address this challenging financial modelling task.

Deep Neural Networks differ from Shallow Neural Networks (one layer) on the multiple internal layers employed between the input values and the predicted result (Figure 2). Constructing a DNN without nonlinear activation functions is impossible, as without these the deep architecture collapses to an equivalent shallow one. Typical choices are logistic sigmoid, hyperbolic tangent and rectified linear unit (ReLU). The logistic sigmoid and hyperbolic tangent activation functions are closely related; both belong to the sigmoid family. A disadvantage of the sigmoid activation function is that it must be kept small due to their tendency to saturate with large positive or negative values. To alleviate this problem, practitioners have derived piecewise linear units like the popular ReLU, which are now the standard choice in deep learning research ReLU, (Vinod & Hinton, 2010).

Figure 2: Shallow and Deep Neural Networks



On a different perspective, since DNNs comprise a huge number of trainable parameters, it is key that appropriate techniques be employed to prevent them from overfitting. Indeed, it is now widely understood that one of the main reasons behind the explosive success and popularity of DNNs consists in the availability of simple, effective, and efficient regularization techniques, developed in the last few years. Dropout has been the first, and, expectably enough, the most popular regularization technique for DNNs (Srivastava et al., 2014). In essence, it consists in randomly dropping different units of the network on each iteration of the training algorithm. This way, only the parameters related to a subset of the network units are trained on each iteration. This ameliorates the associated network overfitting tendency, and it does so in a way that ensures that all network parameters are effectively trained.

Inspired from these merits, we employ Dropout DNNs with ReLU activations to train and deploy feed forward deep neural networks. More precisely we employ the Apache MXNET toolbox of R[2]. We postulated deep networks that are up to five hidden layers deep and comprise various numbers of neurons. Model selection using cross-validation was performed by maximizing the AUROC metric.

We benchmark the abovementioned techniques versus traditional statistical techniques employed in Probability of Default modelling, such as Logistic regression (Logit) and Linear Discriminant Analysis (LDA). Logistic regression is an approach broadly employed for building corporate rating systems and retail scorecards, due to its parsimonious structure. It was first used by Ohlson (1980) to predict corporate bankruptcy based on publicly available financial data. Logistic regression models determine the relative importance of each independent variable in the classification outcome using the fitting dataset. In order to account for non-linearities, and to relax the normality assumption, a sigmoid likelihood function is typically used (Kamstra et al. 2001).

Linear discriminant analysis (LDA) is a method to find a linear combination of features that characterizes or separates two or more classes of objects or events. The main assumptions are that the modelled independent variables are normally distributed and that the groups of modelled objects (e.g. good and bad obligors) exhibit homoscedasticity. LDA is broadly used for credit scoring. For instance, the popular Z-Score algorithm proposed by Altman (1968) is based on LDA to build a rating system for predicting corporate bankruptcies. The normality and homoscedasticity assumptions are hardly ever the case in real-world scenarios, thus, being the main drawbacks of this approach. As such, this method cannot effectively capture nonlinear relationships among the modelled variables, which is crucial for the performance of a credit rating system. We implemented this approach in R using the MASS R package. Before estimating both the logit and the LDA model we dropped collinear variables based on correlation cut-off threshold of 50%.

5. Model Evaluation

Classification accuracy, as measured by the discriminatory power of a rating system, is the main criterion to assess the efficacy of each method and to select the most robust one, in terms of discriminatory power and performance misinterpretation. We tested a series of metrics that are broadly used for quantitatively estimating the discriminatory power of each scoring model, such as the Area Under the ROC curve metric, as well as the Kolmogorov Smirnov (KS) statistic as performance measures.

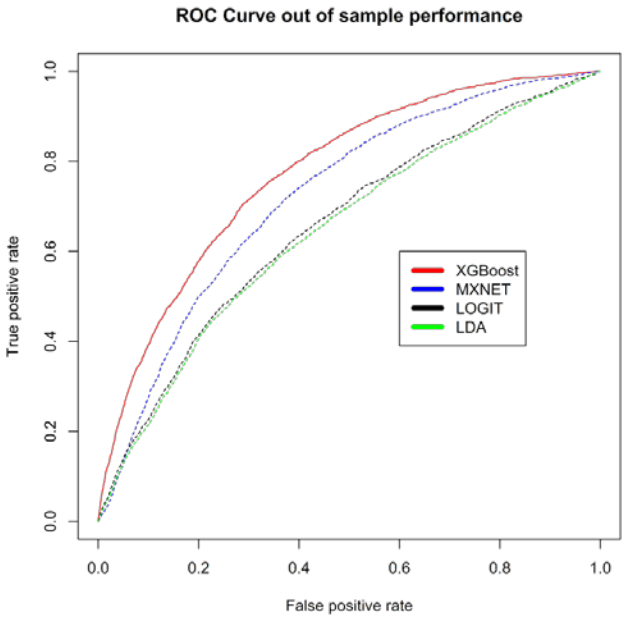
[2] https://mxnet.incubator.apache.org/api/r/index.htm

Classification Accuracy Model Comparison	KS	AUROC	Table 1
Logit	24%	66%	
LDA	23%	65%	
XGBoost	42%	78%	
MXNET	35%	72%	

Classification Accuracy Metrics: Kolmogorov - Smirnov (KS), Area Under ROC curve (AUROC).

Further, we present in Figure 3 the ROC curveS corresponding to the methodologies analysed. This curve is created by plotting the true positive rate against the false positive rate at various threshold settings. As such, they illustrate the obtained trade-offs between sensitivity and specificity, as any increase in sensitivity will be accompanied by a decrease in specificity. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the modelling approach. The corresponding ROC curve of extreme gradient boosting (XGBoost) is higher over all the considered competitors supporting the high degree of efficacy and generalization capacity of the proposed employed machine learning system.

Figure 3: ROC curve for forecasting a default event on 1 year horizon

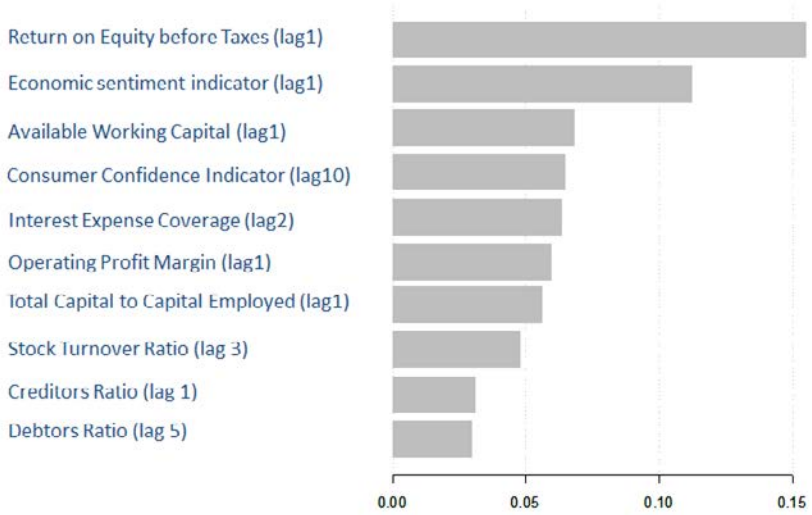


From Table 1 and Figure 3 we deduce that the XGBoost and MXNET algorithms provide better classification accuracy compared to traditional classification methods such as Logistic Regression and Linear Discriminant analysis. As a robustness check other widely employed classification techniques were employed namely the CART algorithm, Random Forests and One Layer Neural Networks (Shallow) but their performance did not surpass the XGBoost and MXNET which is logical if you

consider that the first two are subcases of XGBoost whereas Shallow Neural Network are subcases of MXNET algorithm.

Boosting and Bagging algorithms, even though they are computation intensive, have the relative advantage that they are not "black boxes" regarding the factors affecting the final result, since they provide a module for calculating variable importance measures through reshuffling. In other words after predicting with the benchmark model the reshuffling technique predicts hundreds of times for each variable in the model while randomizing that variable. If the variable being randomized hurts the model's benchmark score, then it is an important variable. If, on the other hand, nothing changes, then it is a useless variable. We run the variable importance algorithm and we show in Figure 4 the ranked list of first ten more important variables

Figure 4: XGBoost variable importance plot. The x-axis describes the percentage contribution of the predictor in the "real" model.



It appears that the most important financial ratio predictor for the default probability of a company is Return on Equity followed by the availability of working capital and Interest Expense Coverage. In essence the company return, the availability of financial resources and the prudent leverage policy may assure the viability of a business. In addition the economic climate, seem to play an additional important role in business viability since the Economic sentiment indicator and the Consumer confidence indicator are rendered important in the model whereas other widely employed factors such as GDP growth seem not to be predominant. What is important is that XGBoost includes both macro variables and financial ratios capturing both the systemic and idiosyncratic behaviour in obligor's credit quality, thus both discriminatory and calibration test exhibit stability and steady performance.

6. Rating System Calibration

An essential aspect of each classification system lies in the creation of a way to represent the classification results to a rating system which can be employed for supervisory purposes in the course central banking operations. For this purpose, we apply a credit rating system calibration process. Calibration of a credit rating system is a mapping process under which each score value is matched to rating grade, which is then associated with a probability of default. To perform the calibration of our systems, the development sample population of each scoring model was split into groups. Specifically, 50 groups (i.e. ranges of scores) were created of equal size, each one including 2% of the total population. Each group is associated with the default rate observed in the development sample.

When necessary, ranges of scores were grouped together, in order to ensure monotonicity of the obtained default rates, maximum intra-rate homogeneity of the observed default rates, and maximum inter-range heterogeneity. In order to overcome overfitting issues and create a reasonable system, each grade included at least 4% of the development population. Grouping optimization was performed based on the Information Value metric.

The following graphs visualize the calibration performed for each rating system. In specific, the first graph present the default rate associated to each of the 50 groups initially created, while the second graph show the default associated to the final selected grades.

Figure 5: Estimated Default Rate of the Initial Grouping (50 Groups)

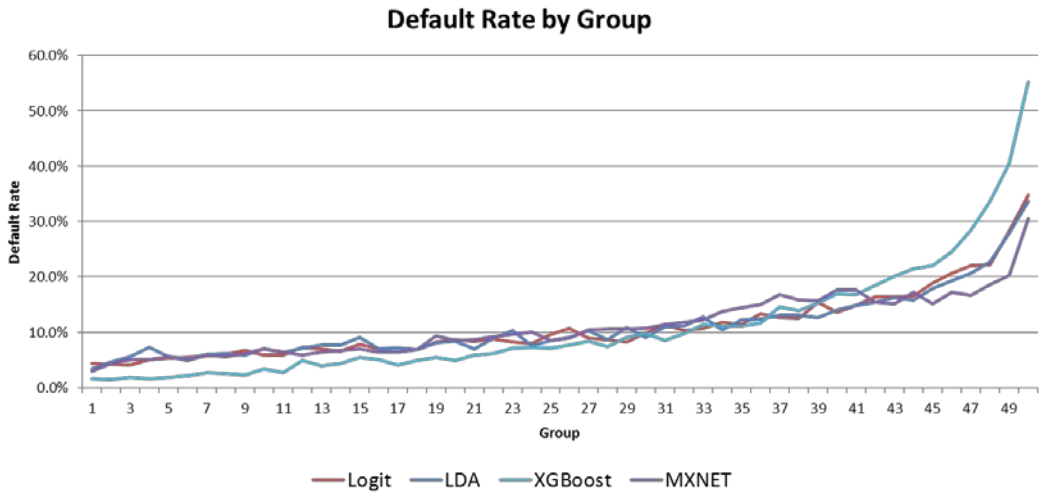
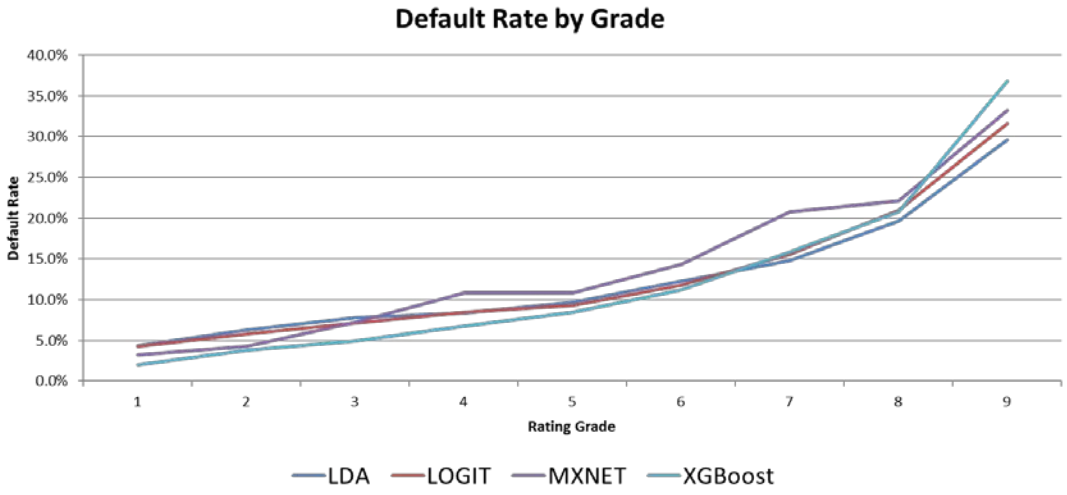


Figure 6: Estimated Default Rate of the Final Selected Grades (9 Grades)



Based on the Figures 5 and 6 it is clear that the XGBoost model is able to produce a more granular calibration. This means that the XGBoost calibration can assign lower default rates to the low grades, and vice versa higher default rates to higher grades, compared to the other models. In order to assess the calibration of the rating systems developed, the binomial test, the normal, the sum of square error, and the brier score validation metrics are utilized. The estimated probability of default was also compared with the out-of-sample observed default rate. The validation methodologies that are typically being applied in the industry include the most, if not all, of the metrics included in our analysis. Additional tests, such as the Bayesian error rate, Chi-square (or Hosmer-Lemeshow) test, that could also have been used, were omitted mainly due to the fact that they produce similar results and conclusions.

Performance Metrics — Table 2

Credit Rating System

	SSE	BRIER
Logit	4.3%	11.3%
LDA	4.8%	11.4%
XGBoost	0.2%	10.1%
MXNET	0.6%	11.0%

Rating System Calibration Metrics: Sum of Square Error (SSE), Brier's score (BRIER).

In the Appendix (Tables 4-7) are shown the calibration results for each evaluated model, i.e. the estimated probability of default per rating grade (based on the default rate on the development sample) and the observed default rate in the out-of-time period. We deduce that Binomial and Normal tests fail for the rating systems developed based on the LDA and Logit models. The estimated PDs are lower than the observed default rates for almost all grades. On the other hand, the MXNET and XGBoost rating systems perform better as the estimated PDs are not statistically different to the observed default rates.

Estimated and Actual default frequency metrics Table 3

	Estimated Probability of Default	Observed Default Rate (Out of sample)	Observed Default Rate (In sample)
Logit	8.20%	13.10%	11.00%
LDA	7.80%		
XGBoost	13.50%		
MXNET	15.00%		

Estimated Probability of Default vs observed Default Rate in out-of-sample and in-sample population

Based on Table 3 it is clear that the rating system developed based on the XGBoost model, is more accurate in terms of PD quantification, compared to the other candidate models due to the more granular calibration achieved by XGBoost. Analytically, XGBoost has marginally overestimated the observed default rate in the validation sample whereas MXNET overestimated the default rate which is good from regulatory perspective. On the other hand, LDA and Logit based systems significantly underestimated the observed default rate.

Deep neural networks provide promising results even though they do not outperform the XGBoost algorithm. The fact is that this methodology provides the opportunity of creating a large combination of different structures based on the number of layers, the selection of activation functions, the number of perceptrons and normalization layers which can be inserted in the optimization process. In the appendix (Figure 7) some illustrative alternative employed structures is shown. Therefore the potentials for Deep Neural Networks algorithms (such as MXNET) in pattern detection in the era of "big data" in which the central banking system is entering are enormous, given that the flexibility of structures is much greater than Boosting and Bagging mechanistic algorithms.

7. Conclusion

In order to tackle the issue of pattern detection in large loan level datasets for extracting information regarding credit risk and exposure credit quality, we employ a combination of data mining algorithms that aim to reduce dimensionality in the data and increase accuracy in predicting the future behaviour of corporate loans. Our analysis is based on a large dataset of loan level data, spanning in a 10 year period of the Greek economy with the purpose of performing obligor credit quality classification and quantification of Probability of Default under a through the cycle setup.

We perform extensive comparisons of the classification and forecasting accuracy of the proposed methods, using a 3-years' period out-of-time sample and we deduce that the Extreme Gradient Boosting technique along with Deep Neural Networks provide better performance in terms of classification accuracy and credit rating system calibration compared to widely employed techniques in credit risk

modelling such as Logistic Regression and Linear Discriminant Analysis. In addition the inclusion of both macro variables and financial ratios captures both the systemic and idiosyncratic behaviour in obligor's credit quality, thus both discriminatory and calibration test exhibit stability and steady performance.

Our findings provide significant oversight for regulatory purposes given that in the coming years, central banks will possess big databases increasing the need for robust data mining processes and financial statistical modelling to support more informed decision making. For example the proposed approaches could find fruitful ground on the European Central Bank's AnaCredit initiative for the collection of loan level data. "Big Data" as referred often entail dimensionality issues, increased noise and other significant statistical challenges which cannot be addressed from traditional statistical techniques.

Regarding the final model selection XGBoost seems to be the methodology marginally outperforming Deep Neural Networks (MXNET) but the latter methodology provides the opportunity of increased flexibility over boosting techniques through a large combination of different structures which may optimize the bias variance trade-off. As a prospect of future research it may be explored whether alternative Deep Neural Network structures, such as recurrent DNN or convolutional networks, may increase the classification accuracy or whether potential forecast combinations among machine and deep learning techniques may further allow boosting of the results.

Appendix

Financial Ratios Employed

- Working Capital
- Employed Capital (Assets minus Current Liabilities)
- Return on Equity before Taxes
- Return on Equity before Interest and Taxes
- Profit before taxes to Employed capital
- Gross Margin to Sales
- Operating Margin to Sales and other income
- Earnings before Interest and Taxes to Sales and other income
- Sales and other income to Employed Capital
- Sales and other income to Equity
- Equity and Long Term Loans to Net Fixed Assets
- Debt to Equity
- Interest Expense Coverage
- Equity to Employed Capital
- Working Capital to Short Term obligations
- Immediate Cash Ratio
- Debtors Ratio
- Creditors Ratio
- Stock turnover Ratio

Macro Variables Employed

- Gross Domestic Product yearly growth
- Investment yearly growth
- Export yearly growth
- Consumption yearly growth
- Economic sentiment indicator
- Consumer Confidence Indicator
- Unemployment Rate
- Inflation
- Stock Market Returns
- Stock Market Volatility
- Deposit Rates
- Loan Rates
- 10 year Government bond spread
- 5 year Government bond spread
- 1 year Government bond spread

Binomial and Normal Validation Tests

Validation Testing - Logistic Regression model — Table 4

Rating Grade	Estimated Probability of Default	Observed Default Rate (Out of sample)	Binomial Test	Normal Test
1	4.21%	6.46%	0.00%	0.00%
2	5.77%	8.83%	0.00%	0.00%
3	7.10%	10.84%	0.00%	0.00%
4	8.38%	13.97%	0.00%	0.00%
5	9.32%	17.46%	0.00%	0.00%
6	11.77%	21.46%	0.00%	0.00%
7	15.53%	23.45%	0.00%	0.00%
8	20.95%	29.64%	0.00%	0.00%
9	31.57%	40.00%	5.07%	4.86%

Binomial and Normal tests examine the null hypothesis that the actual default rate of a credit rating grade is not greater than the forecasted probability of default

Validation Testing - Linear Discriminant Analysis — Table 5

Rating Grade	Estimated Probability of Default	Observed Default Rate (Out of sample)	Binomial Test	Normal Test
1	4.36%	7.37%	0.00%	0.00%
2	6.25%	9.90%	0.00%	0.00%
3	7.76%	12.28%	0.00%	0.00%
4	8.38%	13.80%	0.00%	0.00%
5	9.63%	21.68%	0.00%	0.00%
6	12.19%	20.89%	0.00%	0.00%
7	14.75%	26.01%	0.00%	0.00%
8	19.64%	27.58%	0.00%	0.00%
9	29.65%	30.19%	51.73%	52.55%

Binomial and Normal tests examine the null hypothesis that the actual default rate of a credit rating grade is not greater than the forecasted probability of default

Validation Testing - Extreme Gradient Boosting (XGBoost) Table 6

Rating Grade	Estimated Probability of Default	Observed Default Rate (Out of sample)	Binomial Test	Normal Test
1	1.99%	1.52%	95.15%	94.66%
2	3.78%	2.34%	99.96%	99.92%
3	4.89%	3.88%	96.81%	96.45%
4	6.72%	5.39%	98.28%	98.03%
5	8.41%	6.96%	98.75%	98.58%
6	11.18%	9.75%	99.07%	98.97%
7	15.86%	15.09%	86.36%	86.33%
8	20.82%	22.57%	1.64%	1.57%
9	36.81%	38.43%	5.95%	5.92%

Binomial and Normal tests examine the null hypothesis that the actual default rate of a credit rating grade is not greater than the forecasted probability of default

Validation Testing - MXNET Table 7

Rating Grade	Estimated Probability of Default	Observed Default Rate (Out of sample)	Binomial Test	Normal Test
1	3.2%	0.9%	99.4%	98.5%
2	4.2%	3.5%	80.9%	81.1%
3	7.2%	2.6%	100.0%	100.0%
4	10.8%	6.4%	100.0%	100.0%
5	10.8%	15.6%	32.0%	32.1%
6	14.3%	23.1%	8.2%	8.1%
7	20.8%	28.9%	61.5%	61.9%
8	22.1%	30.2%	90.5%	90.5%
9	33.2%	32.9%	56.6%	56.8%

Binomial and Normal tests examine the null hypothesis that the actual default rate of a credit rating grade is not greater than the forecasted probability of default

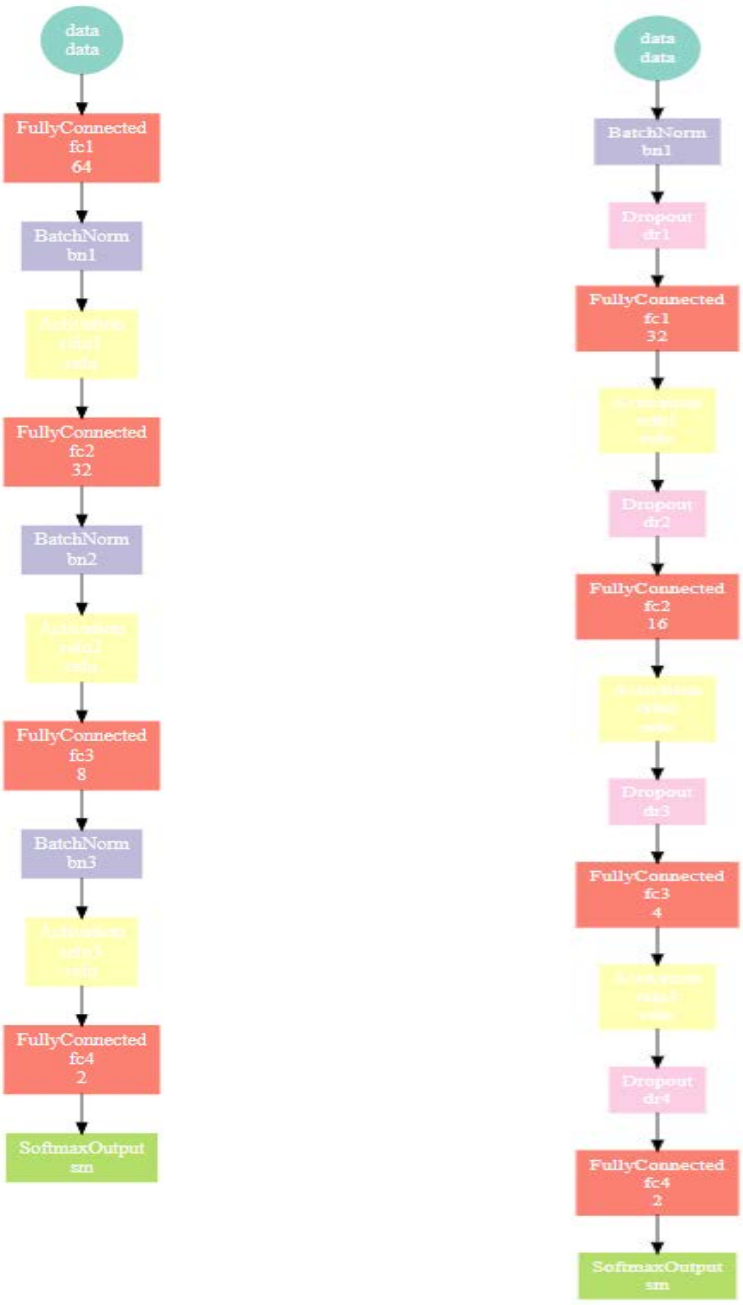


Figure 7: Illustrative structure of some Deep Neural Network structures employed in the optimization process

References

Addo P. M., Guegan D., and Hassani B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. Risks, 6, 2 (38): 2227-9091.

Altman, E.: "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." The journal of finance 23.4 (1968): 589-609.

Avery, R. B., Calem, P. S., & Canner, G. B. (2004). Consumer credit scoring: Do situational circumstances matter? Journal of Banking & Finance, 28 (4), 835–856.

Breiman, L. (2001). Random forest. Machine Learning, 45, 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). Classification and regression trees. CRC press.

Butaru, Florentin, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W. Lo, and Akhtar Siddique. 2016. Risk and risk management in the credit card industry. Journal of Banking and Finance 72: 218–39.

Chava, S., & Jarrow,R.A. (2004). Bankruptcy prediction with industry effects. Re-view of Finance, 8 (4), 537–569.

Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio. Y. (2014), "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," Proc. EMNLP.

Galindo, Jorge, and Pablo Tamayo. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. Computational Economics 15: 107–43.

Huang, S. C. (2011). Using Gaussian process based kernel classifiers for credit rating forecasting. Expert Systems with Applications, 38 (7), 8607–8611.

Huang, Zan, Hsinchun Chen, Chia-Jung Hsu,Wun-Hwa Chen, and SoushanWu. 2004. Credit rating analysis with support vector machines and neural networks: A market comparative study. Decision Support Systems 37: 543–58.

Kamstra, M., Kennedy, p. and Suan, TK. (2001): Combining bond rating forecasts using logit. Financial Review 36.2: 75-96.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking and Finance 34: 2767–87.

Mizen, P., & Tsoukas, S. (2012). Forecasting US bond default ratings allowing for previous and initial state dependence in an ordered probit model. International Journal of Forecasting, 28 (1), 273–287.

Ohlson, J.(1980): Financial ratios and the probabilistic prediction of bankruptcy." Journal of accounting research: 109-131.

Petr, G., & Gurný, M. (2013). Comparison of credit scoring models on probability of default estimation for US banks. Prague Economic Papers, 2, 163–181.

Petropoulos A., Chatzis S.P., Xanthopoulos S (2016). A novel corporate credit rating system based on Student's-t hidden Markov models. Expert Systems with Applications, 53, 87-105.

Quadrana, M., Hidasi, B., Karatzoglou, A. and Cremonesi, P. (2017), Personalizing Session-based Recommendations with Hierarchical Recurrent Neural Networks, Proc. ASM RecSys.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. The Journal of Business, 74 (1), 101–124.

Srivastava, N, Hinton, J., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 15 (2014) 1929-1958.

Tu, Z., Lu, Z., Liu. Y., Liu, X., and Li, H. Modeling coverage for neural machine translation. Proc. ACL (2016).

Vapnik, V. N. (1998). Statistical learning theory. New York: Wiley.

Vinod, Nair & Hinton, Geoffrey (2010), Rectified Linear Units Improve Restricted Boltzmann Machines. Proc. ICML.

Yeh, C.-C., Lin, F., & Hsu, C.-Y. (2012). A hybrid KMV model, random forests and rough set theory approach for credit rating. Knowledge-Based Systems, 22, 166–172.

Zhao, Z, Xu, S, Kang, B. H, Kabir, M. M. J, Liu, Y, & Wasinger, R. (2015). Investigation and improvement of multi-layer perception neural networks for credit scoring. Expert Systems with Applications, 42 (7), 3508–3516.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC – Bank Indonesia International Workshop and Seminar on *"Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data"*

Bali, Indonesia, 23-26 July 2018

A robust machine learning approach for credit risk analysis of large loan-level datasets using deep learning and extreme gradient boosting[1]

Anastasios Petropoulos, Vasilis Siakoulis,
Evaggelos Stavroulakis and Aristotelis Klamargias,

Bank of Greece

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting

BIS International Workshop on Big Data for Central Bank Policies

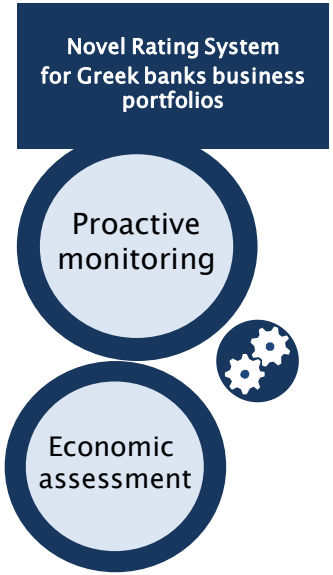
Indonesia, 25 July 2018

BANK OF GREECE
Anastasios Petropoulos
Vasilis Siakoulis
Evangelos Stavroulakis
Aristotelis Klamargias

The views expressed in this paper are those of the authors and not necessarily those of Bank of Greece

Credit Risk Analysis Tool

In a nutshell



Modeling technique

- Extreme Gradient Boosting
- Deep Neural Networks

Main Drivers

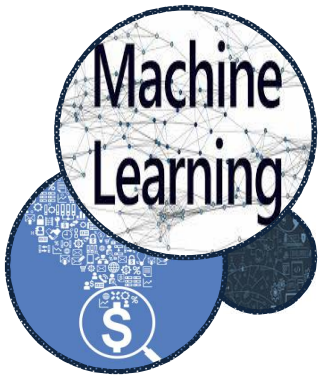
- Company Financial Ratios
- Macroeconomic factors

Implementation

- Corporate and SME loans of the Greek banking system, from the supervisory database of the Central Bank of Greece

Credit Risk Analysis

Machine and Deep learning techniques

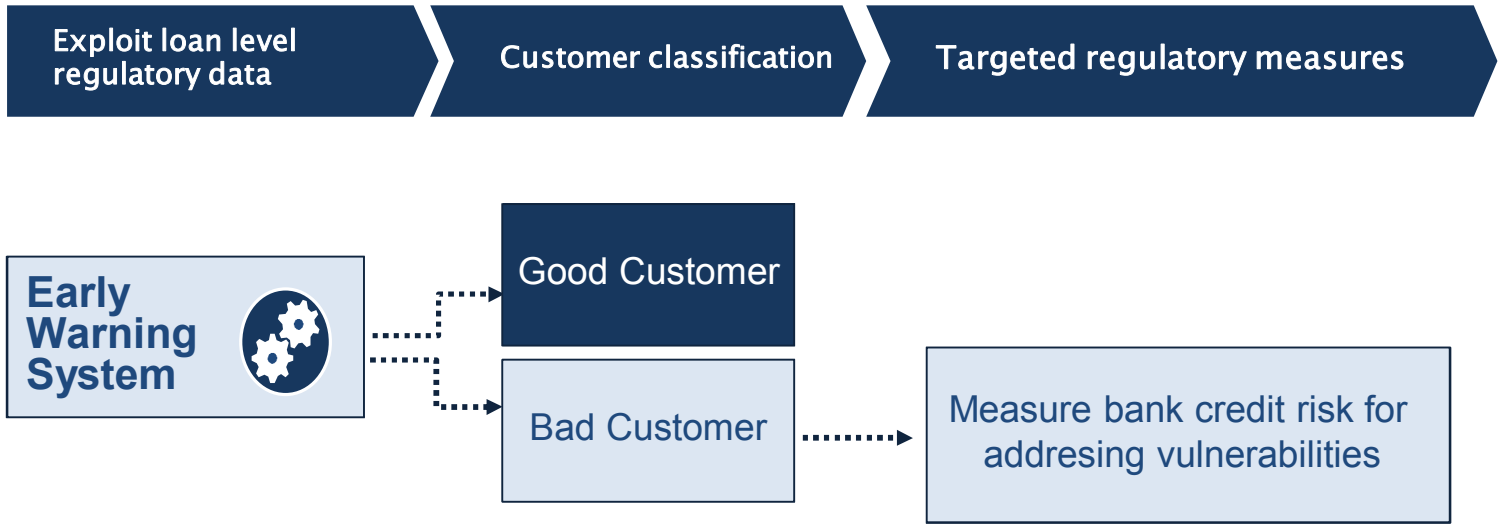


- "Learn" without being explicitly programmed
- Unveiling new determinants and unexpected forms of dependencies among variables.
- Tackling non linear relationships.

- Use of ML and Deep Learning are favored by the technological advances and the availability of financial sector data.
- Supervisory authorities should keep up with the current developments.

Credit Risk Analysis

Bank of Greece – Regulatory Purpose

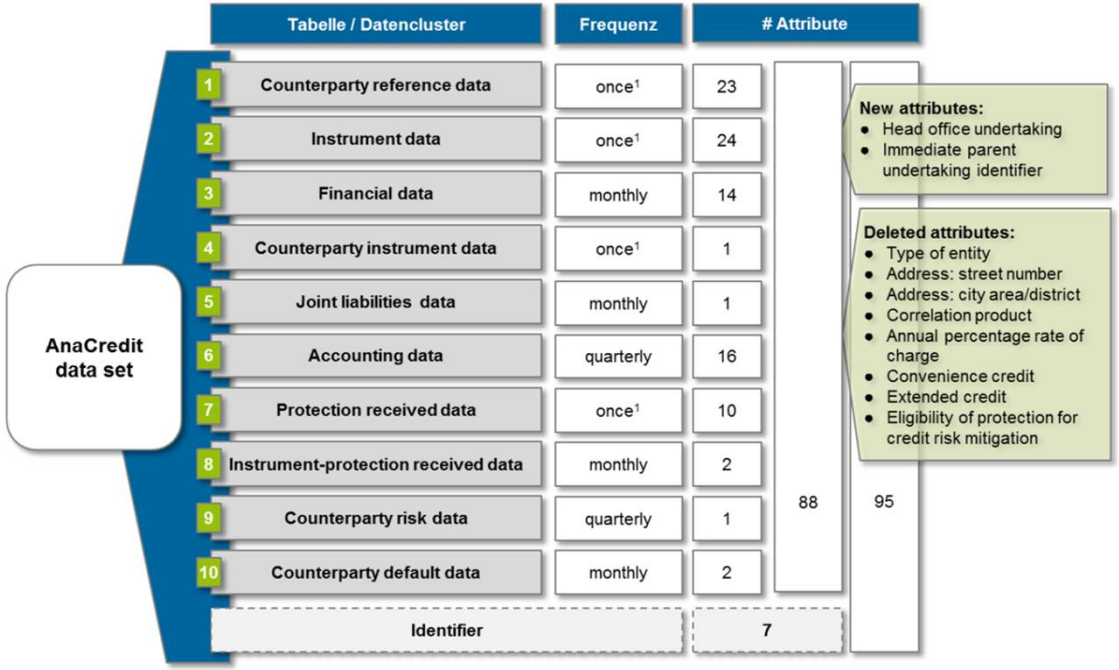


BANK OF GREECE
EUROSYSTEM

Credit Risk Analysis – Big Data

Anacredit project European Central bank

Reporting threshold 25.000 euro



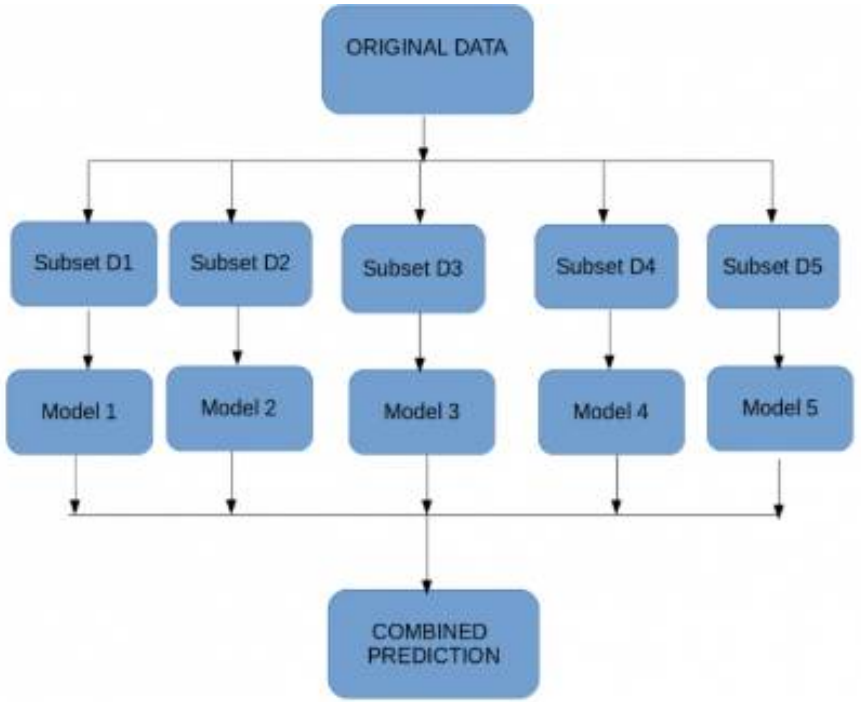
Source: ECB regulation on the collection of granular credit and credit risk data as of May 18th, 2016

- AnaCredit will be a new dataset with detailed information on individual bank loans in the euro area.
- The project was initiated in 2011 and data collection is scheduled to start in September 2018.

BANK OF GREECE
EUROSYSTEM

Credit Risk Analysis

Bagging – Different models vote for the result

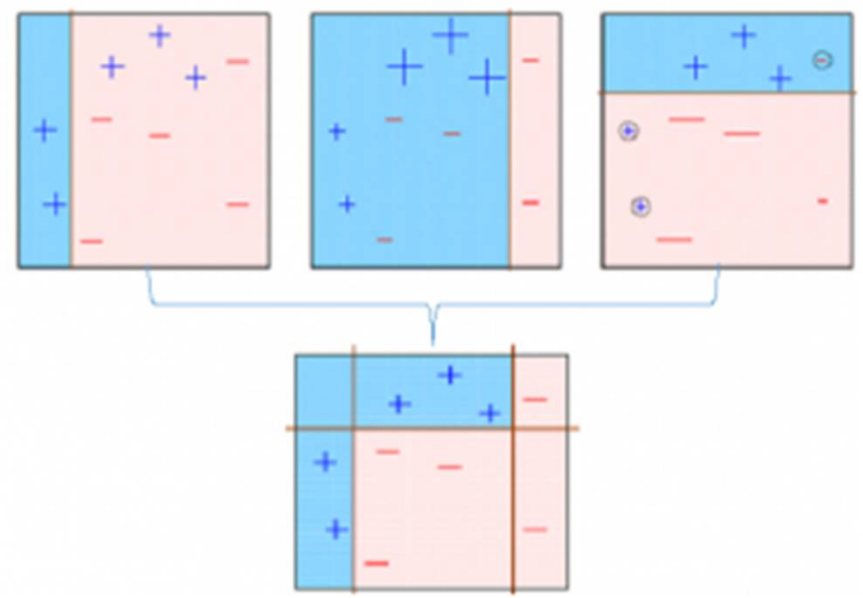


source: Analytics Vidhya

- Multiple subsets are created from the original dataset, selecting observations with replacement and a base model (weak model) is created on each of these subsets.
- The models run in parallel and are independent of each other.
- The final predictions are determined by combining the predictions from all the models.
- Random Forests are common employed bagging techniques.

Credit Risk Analysis

Boosting – Each model learns from the errors of the previous



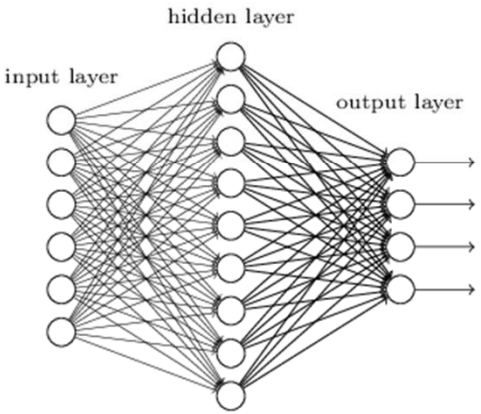
source:
Analytics
Vidhya

- A base model is created based on a subset of the original dataset which is used to make predictions on the whole dataset.
- Errors are calculated and observations which are incorrectly predicted, are given higher weights (large plus signs).
- Another model is created which tries to correct the errors from the previous model.
- Similarly, multiple models are created, each correcting the errors of the previous model.
- The final model (strong learner) is the weighted mean of all the models (weak learners).

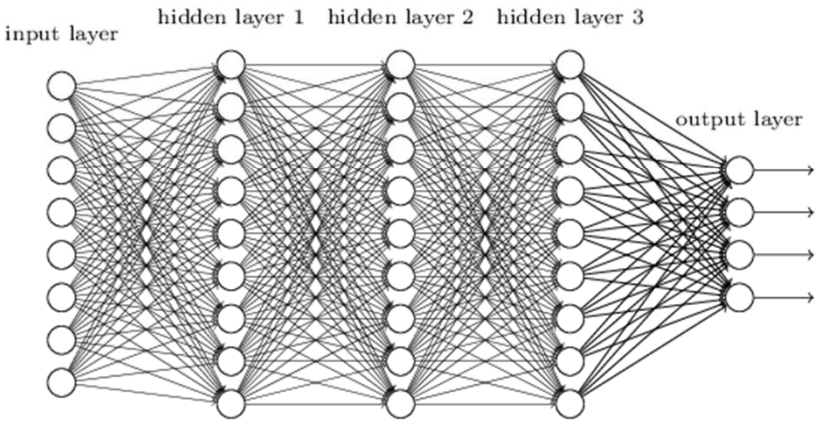
Credit Risk Analysis

Deep Neural Networks-Unlimited potential for Architectures

"Non-deep" feedforward neural network



Deep neural network

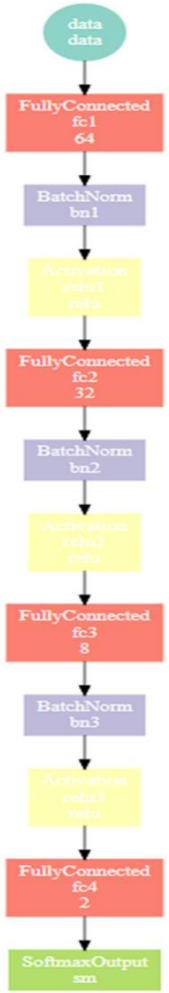


Deep neural network is simply a feedforward network with many hidden layers. It has the following advantages compared to one layer networks ("shallow")

- A deep network needs less neurons than a shallow one
- A shallow network is more difficult to train with our current algorithms (e.g. it has more nasty local minima, or the convergence rate is slower)

Credit Risk Analysis

Deep Neural Networks-Unlimited potential for Architectures



This methodology provides the opportunity of creating a large combination of different structures based on

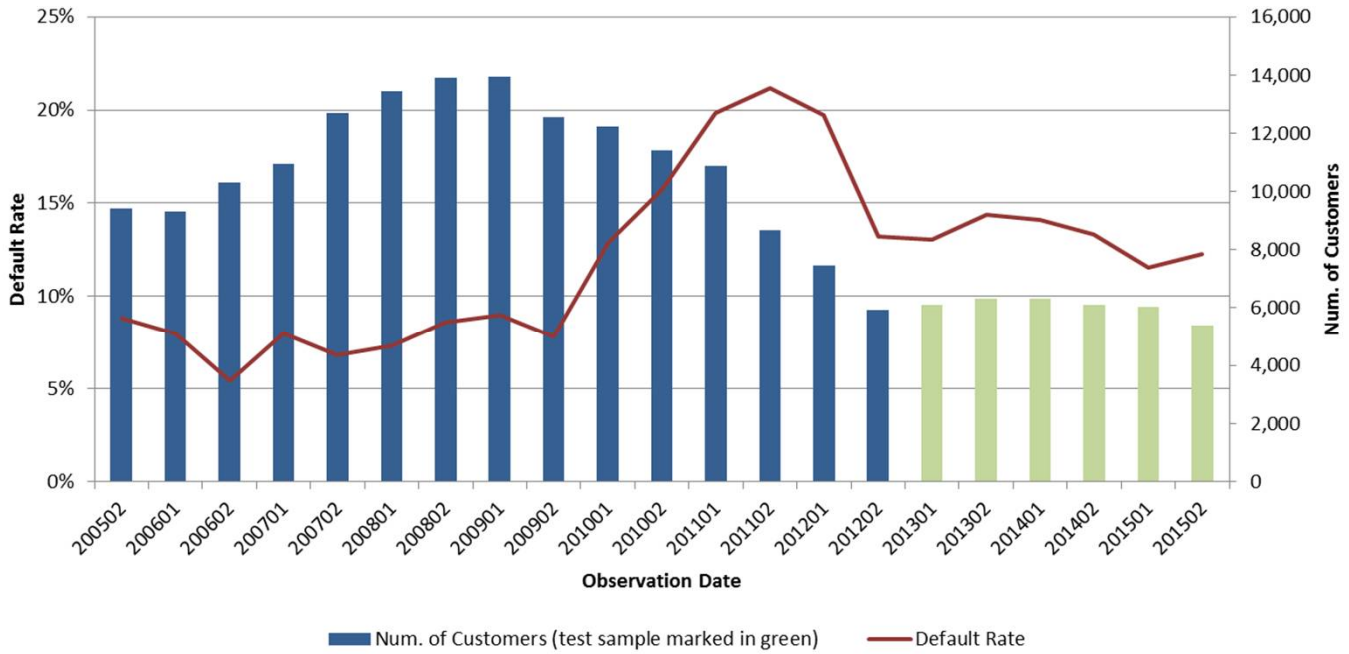
- Number of layers,
- Selection of activation function
- Number of perceptrons
- Normalization layers
- Dropout adjustments

Which can be employed in the optimization process

BANK OF GREECE
EUROSYSTEM

Credit Risk Analysis

Problem at hand



- We have collected loan level information on Corporate and SME loans of the Greek banking system, from the supervisory database of the Central Bank of Greece.
- A loan is flagged as delinquent if it is either 90 days past due or it gets rated as delinquent based on each bank's internal rating rules.
- The forecast horizon for a default event is 1 year whereas the variables employed include macro data and company specific financial ratios.

Credit Risk Analysis

Many Predictor Candidates - Curse of dimensionality

Boruta (aka Leshy): Slavik deity dueling in forests. 1906 illustration

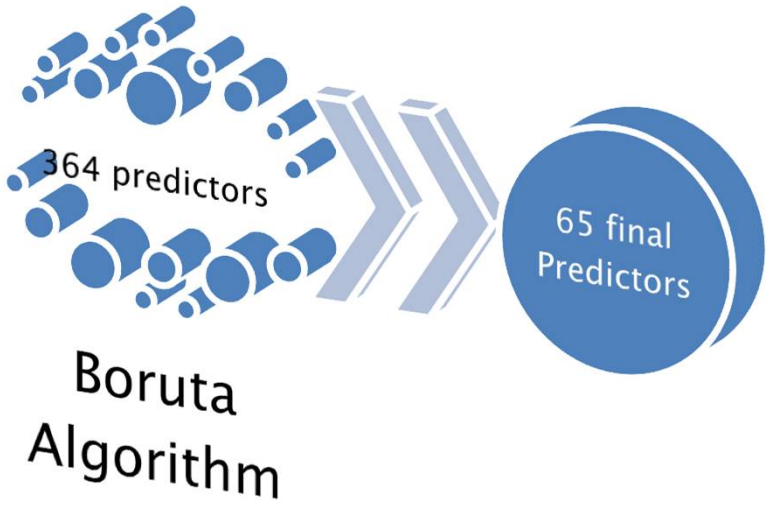
- We employ **Boruta algorithm** for tackling the dimensionality issue. This is sequential Random Forest based algorithm which removes non relevant variables decreasing the dimensionality space.

Credit Risk Analysis

Many Predictor Candidates - Curse of dimensionality

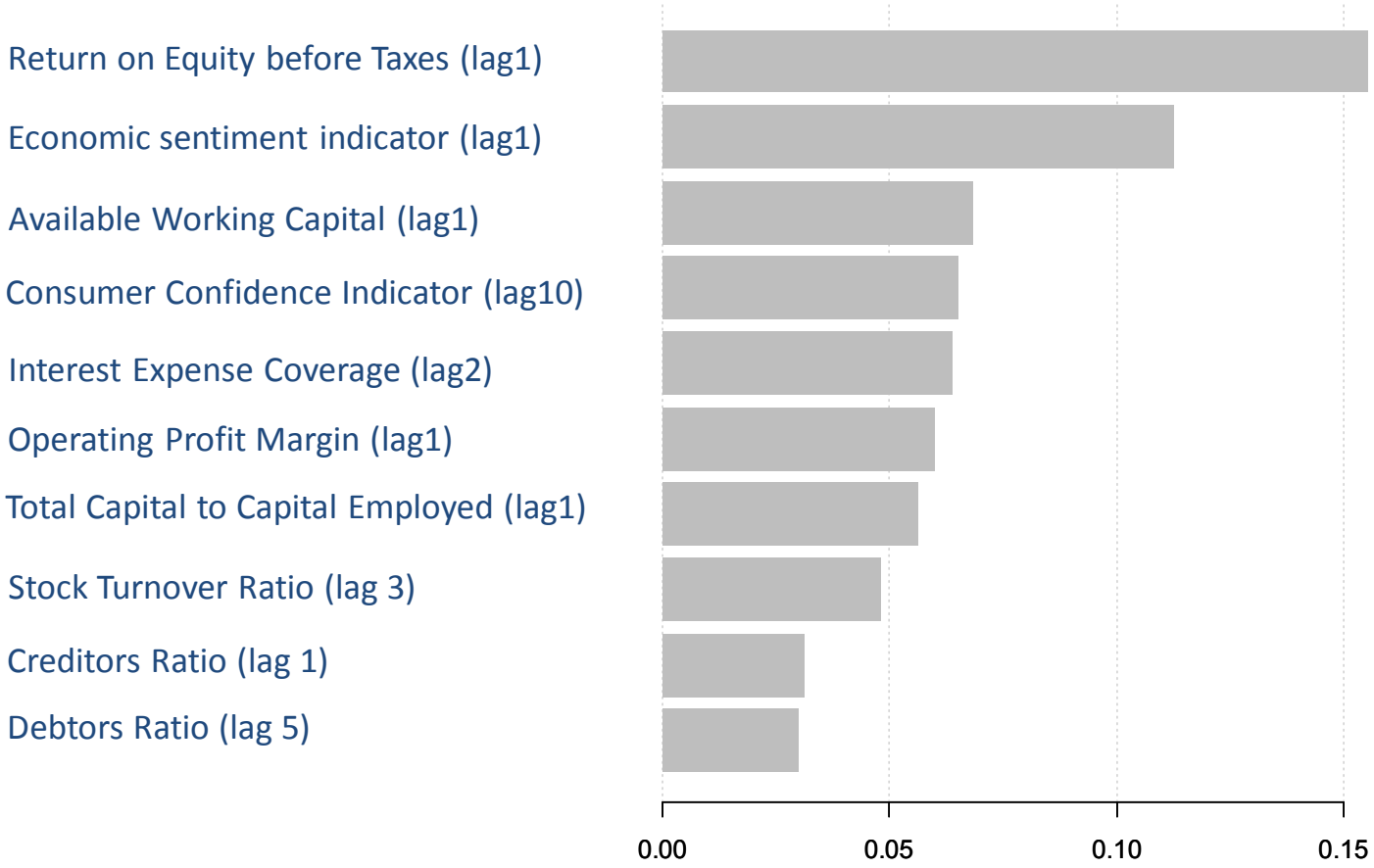
Boruta Algorithm – steps:

- First, it adds randomness to the given dataset by creating shuffled copies of all features (shadow features).
- Then, it fits a Random Forest model (bagging model) on the extended dataset and evaluates the importance of each feature based on Z score.
- In every iteration, it checks whether a real feature has a higher importance than the best of its shadow features, and constantly removes features which are deemed unimportant



Extreme Gradient Boosting

Variable Importance



Extreme Gradient Boosting

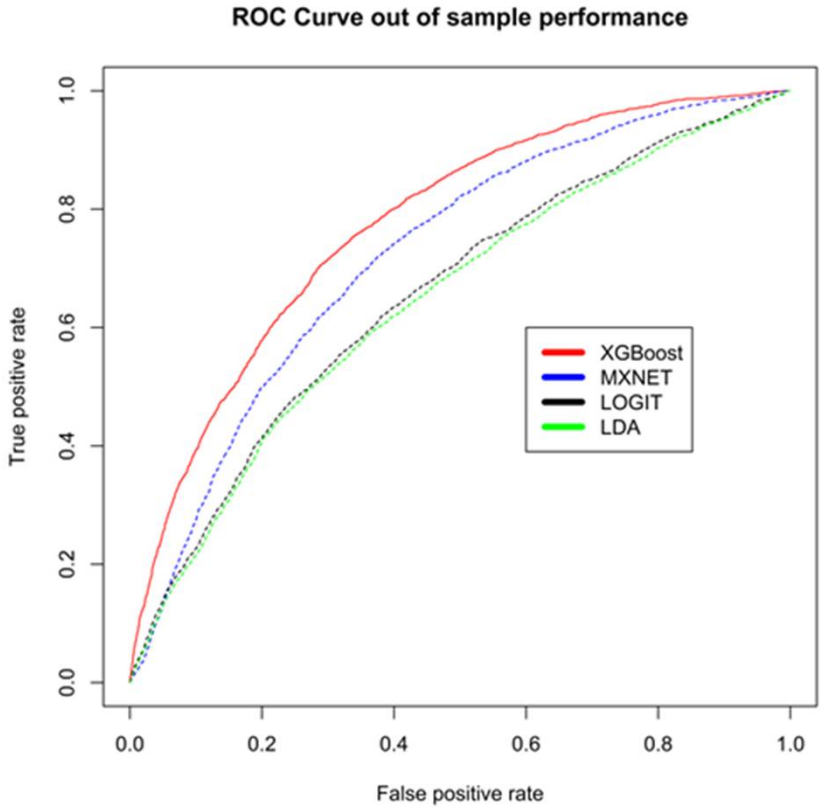
Classification Accuracy

Classification Accuracy Model Comparison	KS	AUROC
Logit	24%	66%
LDA	23%	65%
XGBoost	42%	78%
MXNET	35%	72%

Table 1

Classification Accuracy Metrics: Kolmogorov - Smirnov (KS), Area Under ROC curve (AUROC).

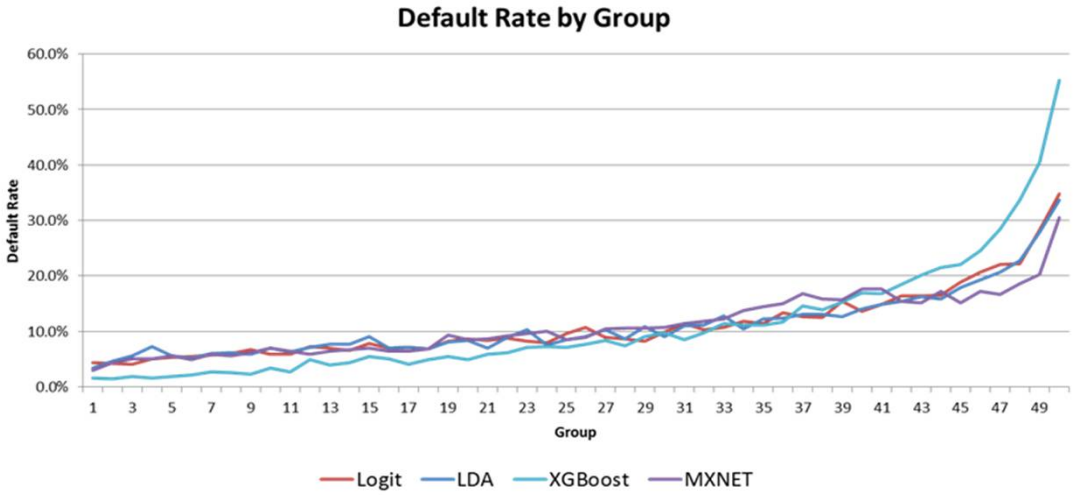
XGBoost and **MXNET** algorithms provide better classification accuracy compared to traditional classification methods such as Logistic Regression and Linear Discriminant analysis.



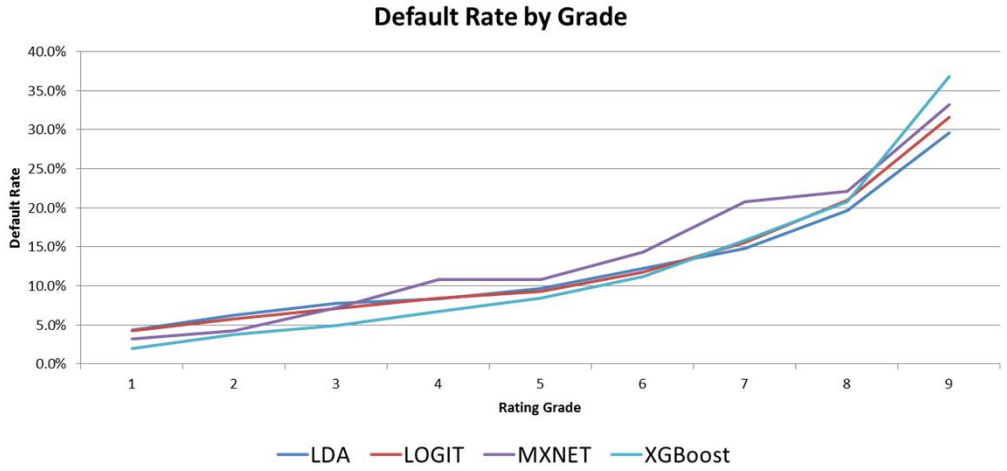
BANK OF GREECE
EUROSYSTEM

Credit Risk Analysis

Calibrating a Rating system



Initial credit rating segmentation in 50 grades



Final credit rating segmentation in 9 grades

BANK OF GREECE
EUROSYSTEM

Deep Neural Networks

Rating System Performance

Performance Metrics — Table 2

Credit Rating System

	SSE	BRIER
Logit	4.3%	11.3%
LDA	4.8%	11.4%
XGBoost	0.2%	10.1%
MXNET	0.6%	11.0%

Rating System Calibration Metrics: Sum of Square Error (SSE), Brier's score (BRIER).

Estimated and Actual default frequency metrics — Table 3

	Estimated Probability of Default	Observed Default Rate (Out of sample)	Observed Default Rate (In sample)
Logit	8.20%	13.10%	11.00%
LDA	7.80%		
XGBoost	13.50%		
MXNET	15.00%		

Estimated Probability of Default vs observed Default Rate in out-of-sample and in-sample population

- Based on SSE and Brier score the MXNET and XGBOOST rating systems perform better than Logistic Regression and Linear Discriminant analysis.
- The estimated PDs for MXNET and XGBOOST are closer to the observed default rates.

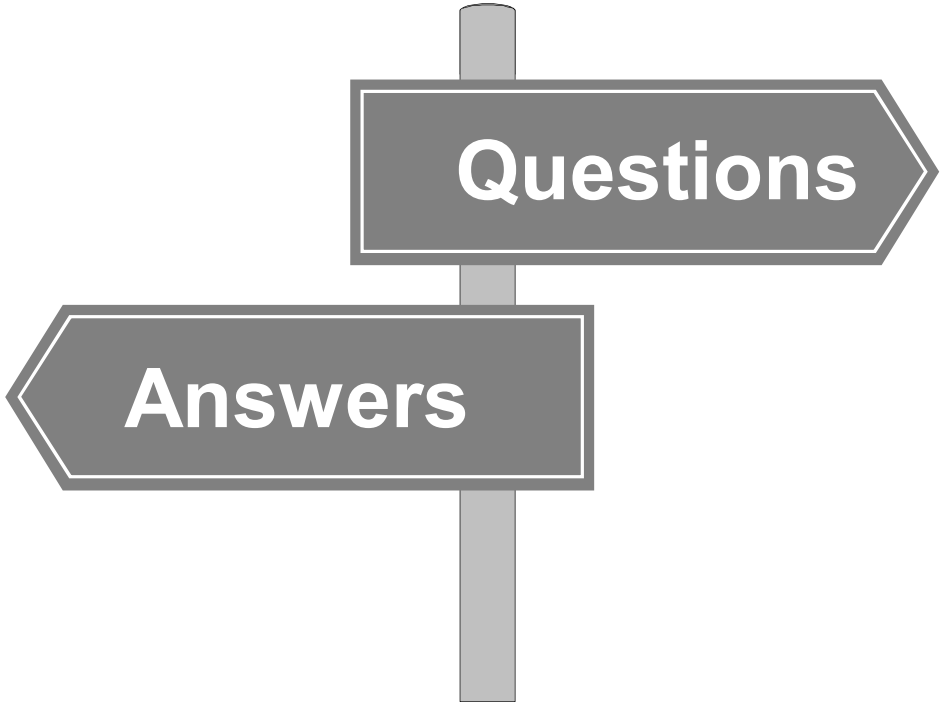
Credit Risk Analysis

Our Contribution

- Extensive exploration of advanced statistical techniques
- An automated algorithm for tackling dimensionality issues
- Application to a regulatory large size dataset
- Robust validation and Performance Measures
- Large potential for application in large datasets (Anacredit)

Credit Risk Analysis

Q&A



BANK OF GREECE
EUROSYSTEM

Thank you!