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

9th biennial IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 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

Novel Rating System for Greek banks business portfolios

Proactive monitoring

Economic assessment

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

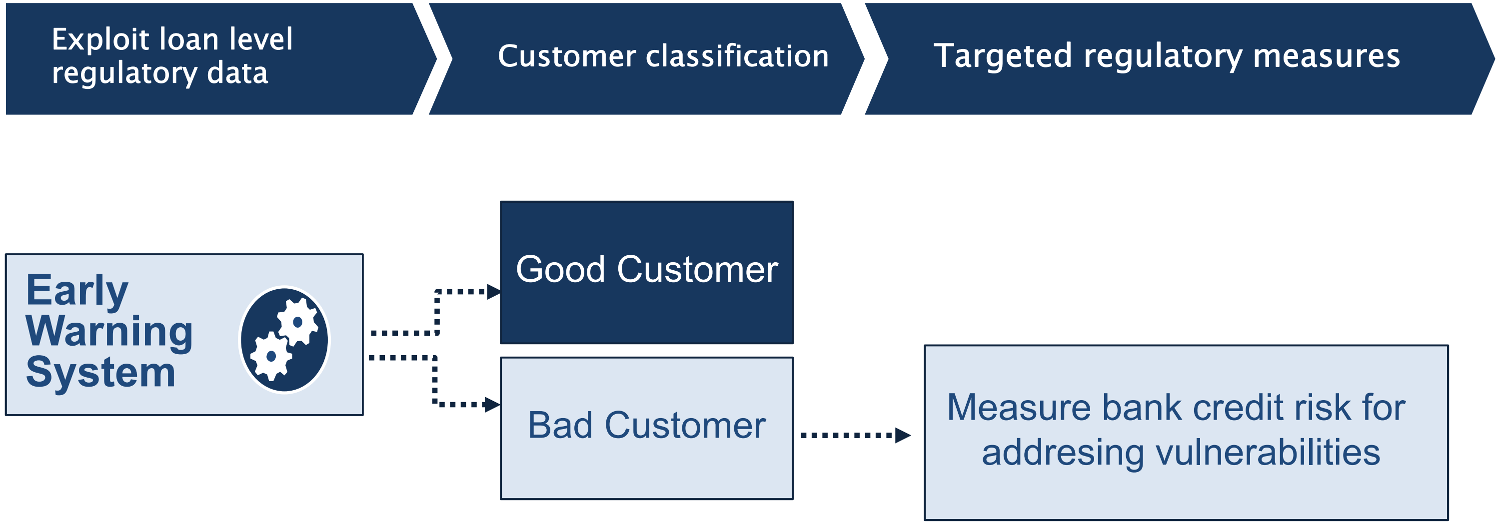
- "Learn" without being explicitly programmed
- Unveiling new determinants and unexpected forms of dependencies among variables.
- Tackling non linear relationships.

- 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.

BANK OF GREECE
EUROSYSTEM

Credit Risk Analysis

Bank of Greece – Regulatory Purpose



Credit Risk Analysis – Big Data

Anacredit project European Central bank

Reporting threshold 25.000 euro

AnaCredit data set

	Tabelle / Datencluster	Frequenz	# Attribute
1	Counterparty reference data	once[1]	23
2	Instrument data	once[1]	24
3	Financial data	monthly	14
4	Counterparty instrument data	once[1]	1
5	Joint liabilities data	monthly	1
6	Accounting data	quarterly	16
7	Protection received data	once[1]	10
8	Instrument-protection received data	monthly	2
9	Counterparty risk data	quarterly	1
10	Counterparty default data	monthly	2
	Identifier		7

Totals: 88 / 95

New attributes:
- Head office undertaking
- Immediate parent undertaking identifier

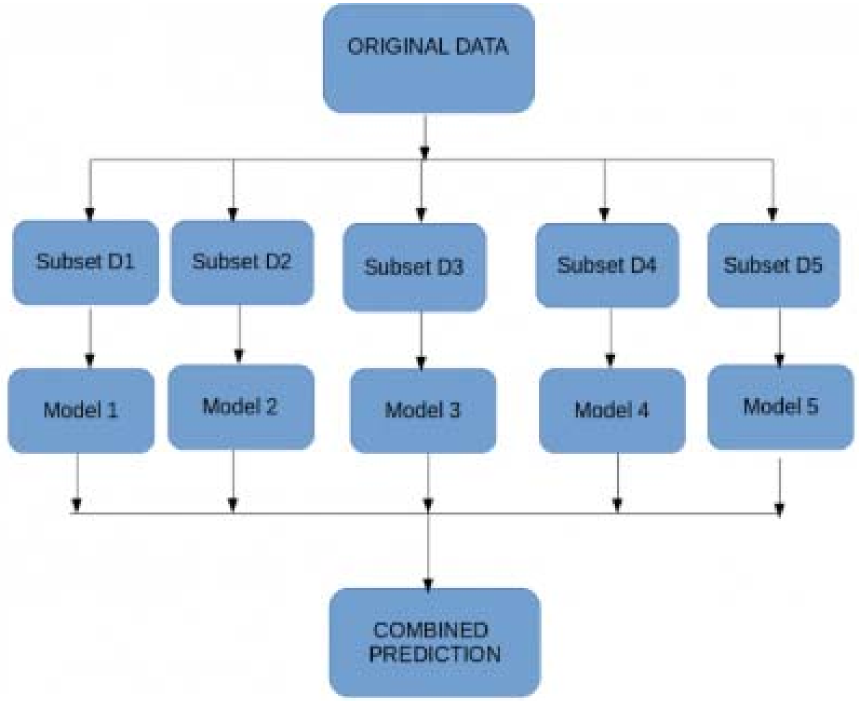
Deleted attributes:
- Type of entity
- Address: street number
- Address: city area/district
- Correlation product
- Annual percentage rate of charge
- Convenience credit
- Extended credit
- Eligibility of protection for credit risk mitigation

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.

Credit Risk Analysis

Bagging – Different models vote for the result

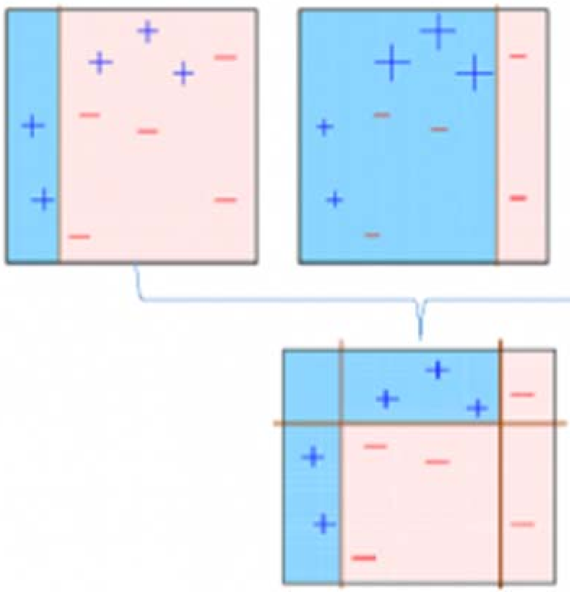


source: Analytics Vidhya

- 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

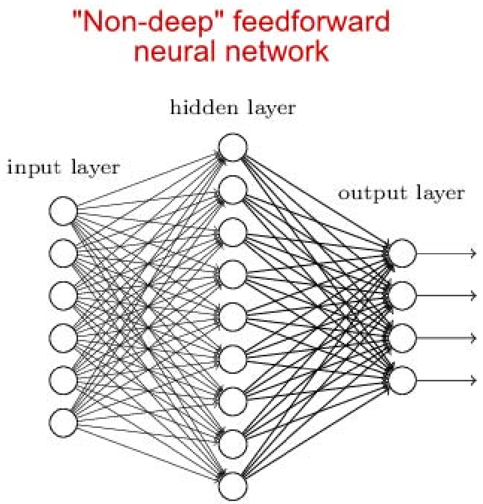


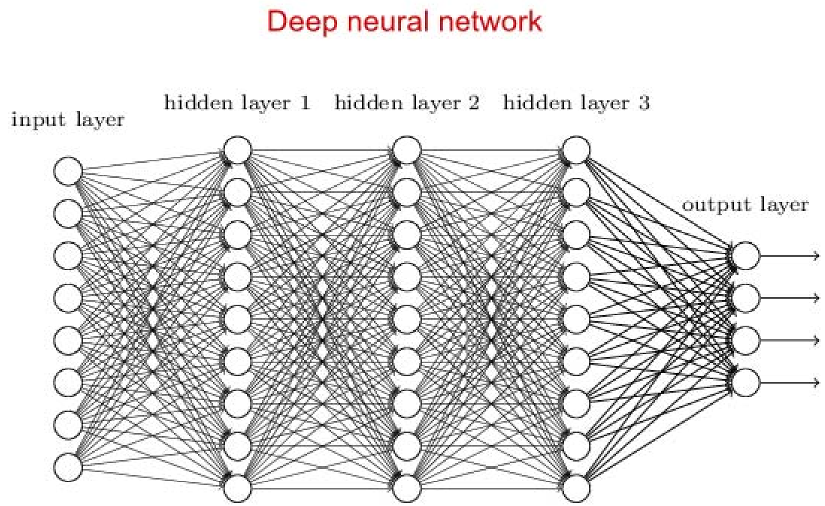
source: Analytics Vidhya

- 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





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

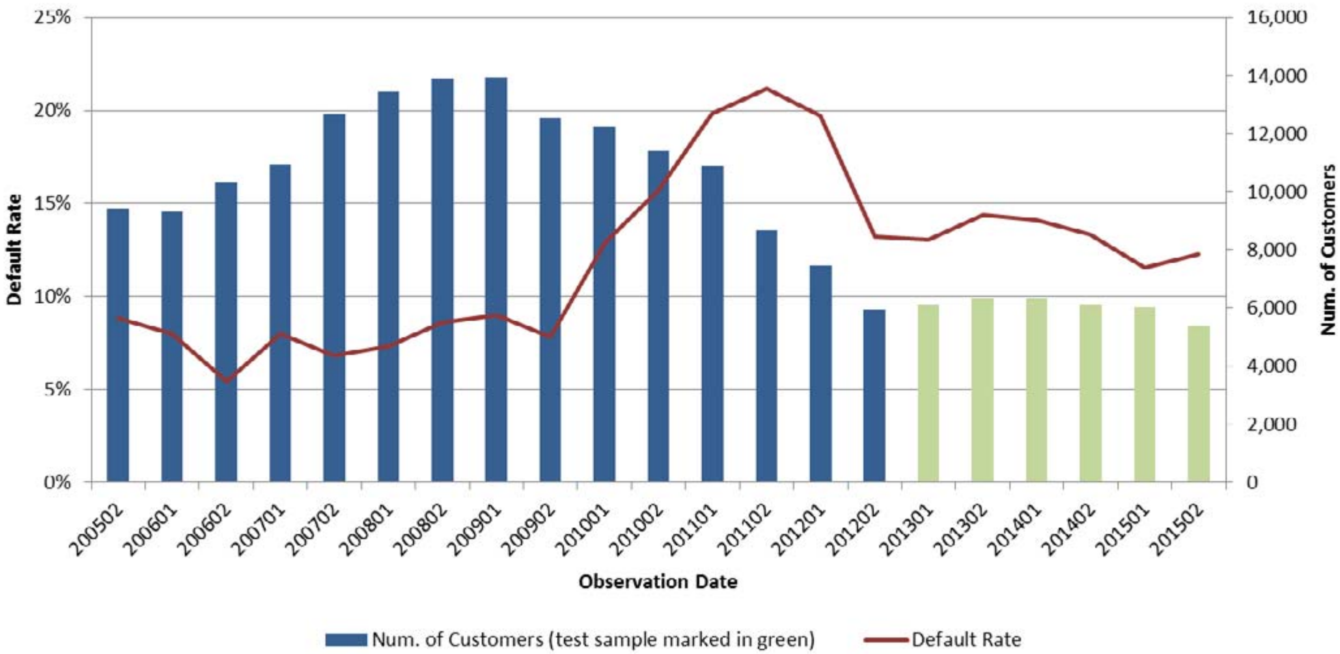
Deep Neural Networks-Unlimited potential for Architectures

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

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

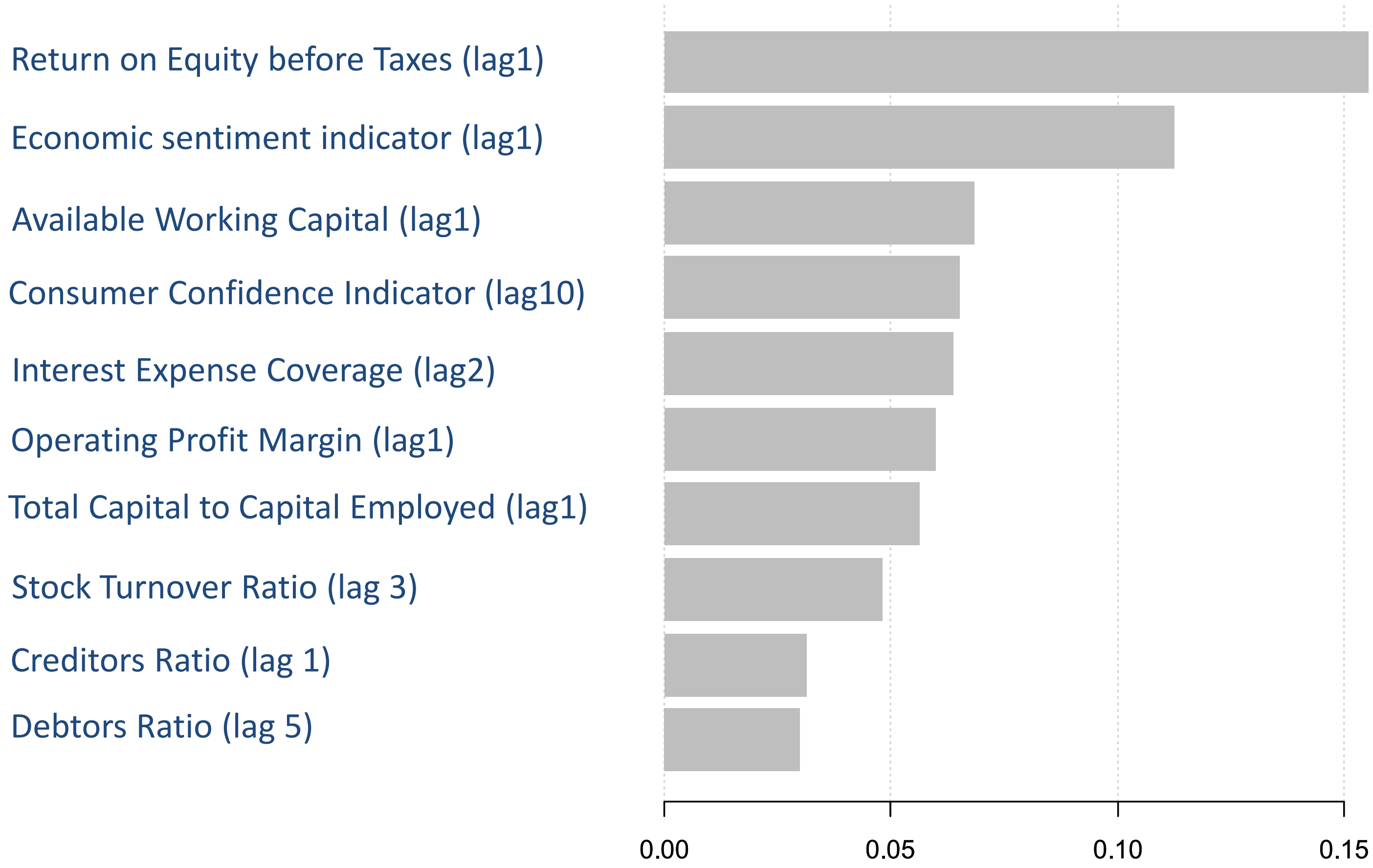
Many Predictor Candidates - Curse of dimensionality

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

BANK OF GREECE
EUROSYSTEM

Extreme Gradient Boosting
Variable Importance

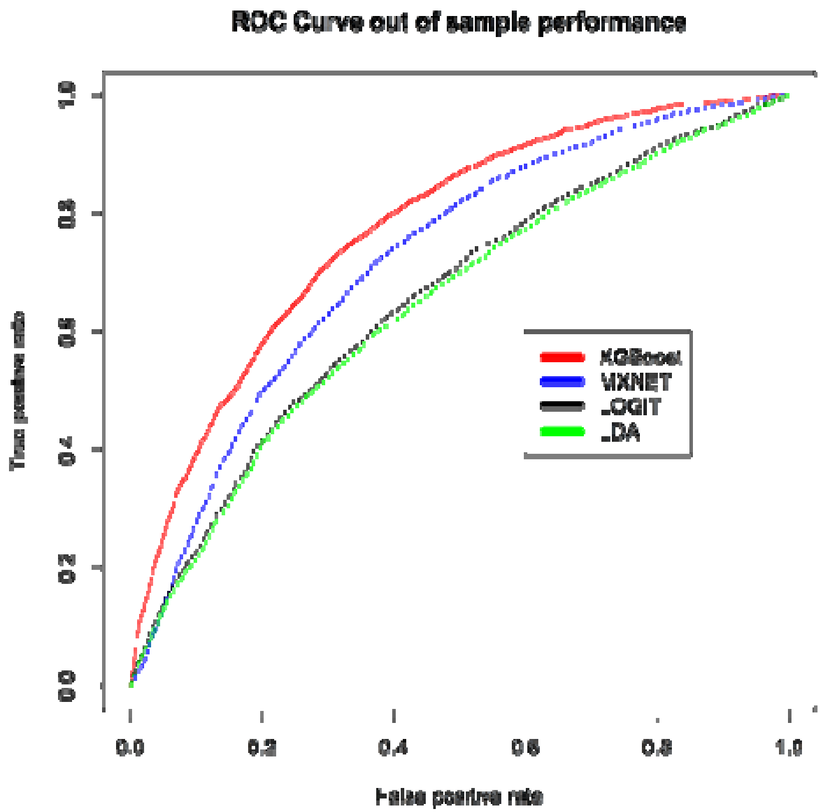


Extreme Gradient Boosting
Classification Accuracy

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

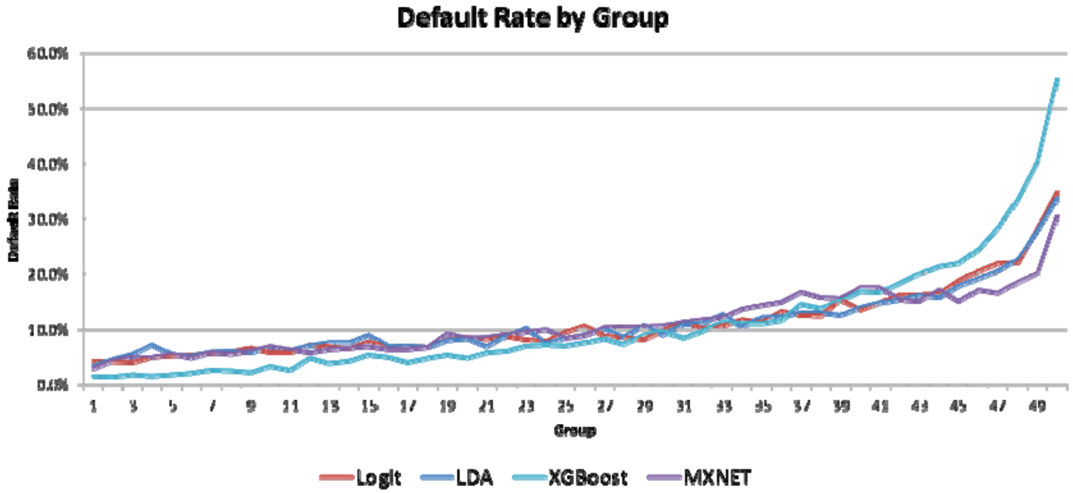
Classification Accuracy Metrics: Kolmogorov - Smirnov (KS), Area Under ROC curve (AUROC).

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

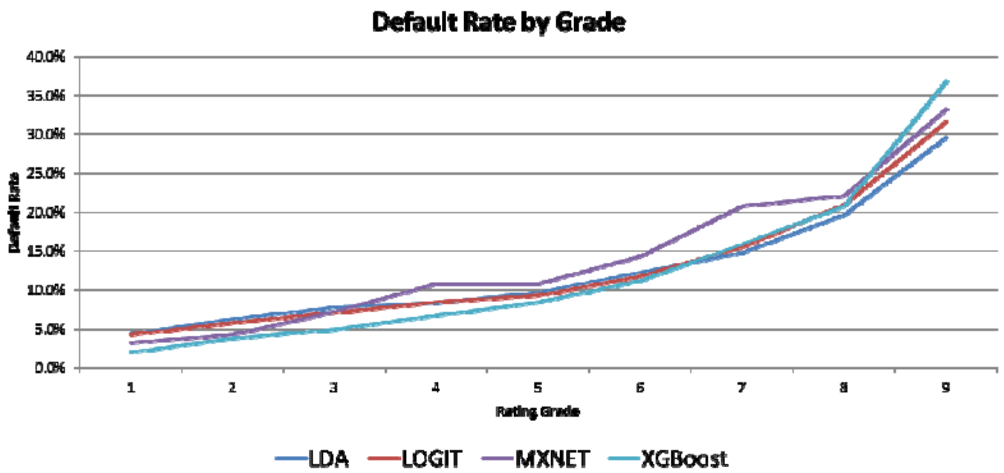


Credit Risk Analysis

Calibrating a Rating system



Initial credit rating segmentation in 50 grades



Final credit rating segmentation in 9 grades

Deep Neural Networks

Rating System Performance

Performance Metrics Credit Rating System		Table 2
	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.

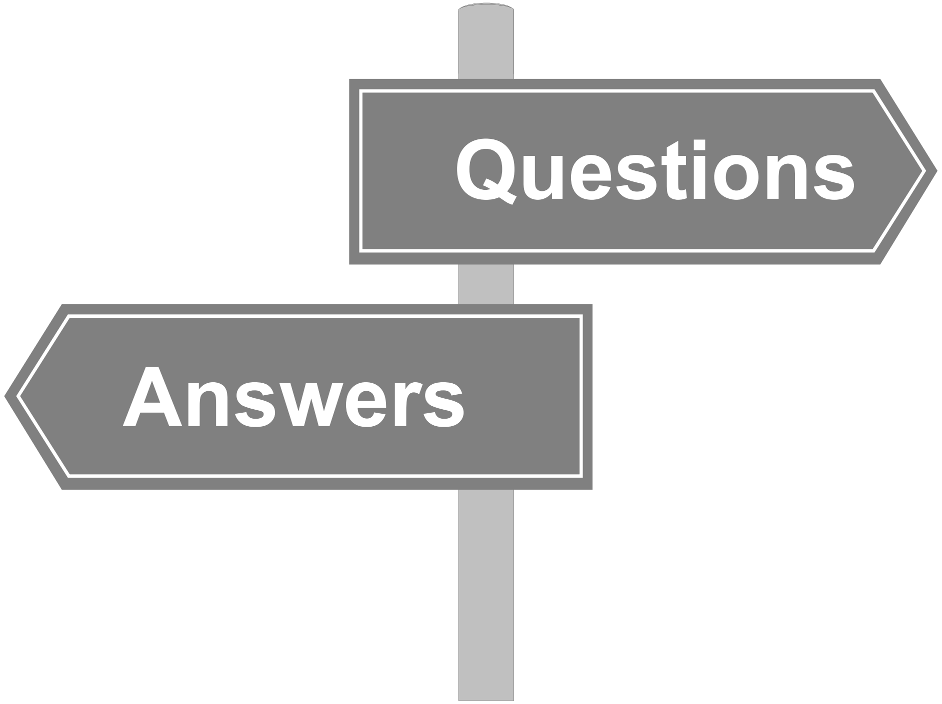
Credit Risk Analysis

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!