BANK OF GREECE
EUROSYSTEM

Working Paper

Is COVID-19 reflected in AnaCredit dataset? A big data - machine learning approach for analysing behavioural patterns using loan level granular information

Anastasios Petropoulos
Evangelos Stavroulakis
Panagiotis Lazaris
Vasilis Siakoulis
Nikolaos Vlachogiannakis

315

MARCH 2023

BANK OF GREECE
Economic Analysis and Research Department – Special Studies Division
21, E. Venizelos Avenue
GR-102 50 Athens
Tel: +30210-320 3610
Fax: +30210-320 2432

www.bankofgreece.gr

Published by the Bank of Greece, Athens, Greece

ISSN: 2654-1912 (online)
DOI: https://doi.org/10.52903/wp2023315

IS COVID-19 REFLECTED IN ANACREDIT DATASET? A BIG DATA - MACHINE LEARNING APPROACH FOR ANALYSING BEHAVIOURAL PATTERNS USING LOAN LEVEL GRANULAR INFORMATION

Anastasios Petropoulos
Bank of Greece

Evangelos Stavroulakis
Bank of Greece

Panagiotis Lazaris
Bank of Greece

Vasilis Siakoulis
Bank of Greece

Nikolaos Vlachogiannakis
Bank of Greece

Abstract

In this study, we explore the impact of COVID-19 pandemic on the default risk of loan portfolios of the Greek banking system, using cutting edge machine learning technologies, like deep learning. Our analysis is based on loan level monthly data, spanning a 42-month period, collected through the ECB AnaCredit database. Our dataset contains more than three million records, including both the pre- and post-pandemic periods. We develop a series of credit rating models implementing state of the art machine learning algorithms. Through an extensive validation process, we explore the best machine learning technique to build a behavioral credit scoring model and subsequently we investigate the estimated sensitivities of various features on predicting default risk. To select the best candidate model, we perform comparisons of the classification accuracy of the proposed methods, in 2-months out-of-time period. Our empirical results indicate that the Deep Neural Networks (DNN) have a superior predictive performance, signalling better generalization capacity against Random Forests, Extreme Gradient Boosting (XGBoost), and logistic regression. The proposed DNN model can accurately simulate the non-linearities caused by the pandemic outbreak on the evolution of default rates for Greek corporate customers. Under this multivariate setup we apply interpretability algorithms to isolate the impact of COVID-19 on the probability of default, controlling for the rest of the features of the DNN. Our results indicate that the impact of the pandemic peaks in the first year, and then it slowly decreases, though without reaching yet the pre COVID-19 levels. Furthermore, our empirical results also suggest different behavioral patterns between Stage 1 and Stage 2 loans, and that default rate sensitivities vary significantly across sectors. The current empirical work can facilitate a more in-depth analysis of AnaCredit database, by providing robust statistical tools for a more effective and responsive micro and macro supervision of credit risk.

Keywords: Credit Risk, Deep Learning, AnaCredit, COVID-19

JEL-classification: G24, C38, C45, C55

Disclaimer: The views expressed on this paper are those of the authors and not of the Bank of Greece.

Correspondence:
Anastasios Petropoulos
Bank of Greece
Amerikis 3, Athens, 102 50
Email: APetropoulos@bankofgreece.gr

1. Introduction

The coronavirus (COVID-19) pandemic caused a structural break for all sectors of the global economy. Countermeasures, like countries' lockdowns, taken to contain the virus and save lives, hampered the economies from functioning properly and caused recession. From the perspective of financial institutions, credit portfolios were highly affected because of the volatility in the companies' balance sheets. The government support measures introduced to mitigate the impact of the pandemic had only short-term relief effects on the viability of Corporate and SMEs. Financial institutions had to increase their efforts for active management of their loan portfolios, via offering short term forbearances to increase short term liquidity of their customers. The unique features of the pandemic have led financial institutions and banking supervisors to move more quickly to enhance their data analytics into their credit-decision engines. From a central bank perspective, new approaches to credit-risk monitoring of the banking system, which attempt to combine the whole spectrum of information collected by banks like obligor and instrument level analysis, are necessary. The shift towards more robust data analytics is expected to come in the post-pandemic era, enabling a real-time monitoring and effective mining of supervisory data, as well as automating the feeding of results into the decision-making process. Hence, the need of advance statistical modelling techniques, able to capture the full nexus of default risk, is becoming even more crucial in the current economic environment for understanding the financial system dynamics and be able to make informed decisions.

In this new era, the utilization of big data that is available to central banks is the cornerstone for boosting their digital transformation. AnaCredit is an important initiative introduced in the aftermath of global financial crisis by European Central Bank to support the macroprudential and microprudential activities of the central banks in the Eurozone. This data source provides detailed loan-level information, on a monthly basis, for the corporate sector, so enabling a comprehensive analysis of the credit risk undertaken by the banking system. Particularly, the riskiness in the loan books of Eurozone Banks is fully captured through the numerous features it contains, enabling disaggregated analysis by various dimensions, like sector, bank, staging, geography, and obligor. On the flip side, big financial datasets usually pose significant statistical challenges because they are characterized by increased noise, heavy-tailed

distributions, nonlinear patterns, and temporal dependencies. These attributes in financial variables patterns became more intense after the COVID-19 shock.

To address the above-mentioned challenges, the enrichment of statistical techniques is important to accommodate for the increased availability of data, and to facilitate the extraction of any possible information they convey. Conventional econometric methods usually fail to efficiently capture the information contained in the full spectrum of these large datasets, as multicollinearity[1] is usually present in the independent variables and the interactions to be captured are of nonlinear nature. Machine learning algorithms are employed nowadays to tackle the issue of variable selection and of modelling the underlying complicated temporal dependencies. Deep learning algorithms have remarkably increased the capabilities of data analytics in performing pattern recognition and classification. Their structure offers the capacity to adapt in the dataset via continuous learning algorithms, and recognize new and evolving patterns, both in time series and cross-sectional datasets. In addition, deep learning effectively deals with high dimensional data that exhibit nonlinear behaviour. Thus, their complicated and non-parametric structure could lead to improved predictive performance in financial modelling.

Motivated by the impact of the pandemic in the loan portfolios and the high dimensionality and granularity of the information collected through the AnaCredit database, we analyse a corporate credit loans big dataset, using bleeding edge machine learning techniques and deep learning neural networks. The purpose of this study is to develop a behavioural scoring model that aims at evaluating the risk of existing customers based primarily on their recent transactional data, including repayment performance and delinquencies. Such models have (as business practice) no more than one 1 year of forecasting horizon, as they pursue to solve a classification problem in the short term to gauge banks'/lenders' decisions, but definitely not for longer. For the longer term, and for identifying economic structures or long-term economic relationships, conventional econometrics is fit for purpose. Hence, this study focuses on explaining the main behavioural determinants or triggers for an obligor to default,

[1] Techniques like regularization (e.g., LASSO) may be used to discard some features, though, this leads to the information they convey to be discarded as well.

and on assessing whether this relationship was affected by the COVID-19 pandemic. The novelty of our approach lies in the following areas:

i. We use advanced machine learning techniques, like Deep Neural Network (DNN) and Extreme Gradient Boosting, for the development of behavioral credit scoring, while benchmarking their performance with traditional econometric methods.
ii. We make use of the big and high dimensional AnaCredit dataset. To the best of our knowledge, this is the first study to apply machine learning to AnaCredit for modelling credit risk.
iii. We provide a thorough out-of-time evaluation of the proposed novel approaches and benchmark our results against multiple statistical methods to provide evidence of DNNs superior performance.
iv. We perform an extensive interpretability analysis, increasing transparency on the functionality of the proposed DNN, so removing its black box nature.
v. We disentangle the impact of COVID-19 in default risk across various dimensions.

The remainder of this study is structured as follows: In section 2, we focus on the related literature review on credit risk models and relevant studies regarding COVID - 19 impact on banks' loan portfolios. Section 3 describes the data collection and processing steps implemented. In section 4, we provide technical details regarding the estimation process of the developed models. In section 5, we outline the results of the validation process and compare the forecasting accuracy across the different models implemented. In section 6, we employ interpretability algorithms for Machine Learning, and provide significant insights regarding the COVID-19 impact and feature sensitivities captured by DNN. Finally, in the concluding section 7, we summarize the performance of the proposed methodologies and discuss future potential research.

2. Literature review

During the last decades a large number of approaches has been used to address the problem of modelling the credit quality of a company. However, it was only recently that more accurate and robust systems, which make use of novel statistical techniques from the field of machine and deep learning, have been employed to drive expert decisions.

Linear regression models (Avery, et al., 2004), Probit models (Mizen and Tsoukas, 2012) and Hazard Rate models (Chava and Jarrow, 2004 & Shumway, 2001) have been extensively employed in credit risk modelling, nevertheless, their core weakness stem from their inability to capture non-linear dynamics, which are prevalent in financial ratio data (Petr and Gurný, 2013). Furthermore, more advanced modelling techniques have also been performed to tackle the weaknesses of traditional credit scoring models, though, not fit for purpose for analysing big data. Galindo and Tamayo (2000) test CART decision-tree models on mortgage-loan data to detect defaults. 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. Finally, Huang et al. (2004) employ support vector machines (SVM – Vapnik 1998) to the credit risk estimation in an attempt to provide a model with better explanatory power.

A number of recent studies have employed Machine and Deep Learning methods in the field of credit risk evaluation models. Addo et al. (2018) focus on credit risk scoring by examining 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 subset of features selected by the models. They observe that the tree-based models are more stable than the models based on multilayer artificial neural networks. Petropoulos et al. (2019) combine dimension reduction algorithms along with different machine learning techniques and deep neural networks measuring credit risk on a 10-year loan dataset of Greek banks. Their results are benchmarked against other traditional methods, like logistic regression and discriminant analysis methods, yielding significantly superior performance.

Feng et al. (2021), developed an ensemble deep-learning model for credit risk evaluation to deal with imbalanced credit data, showing its relevant over-performance when compared to other models. Wang et al (2020), provide a comparative assessment of credit risk models of different Machine Learning techniques on bank loan data. Hamori et al (2018) analyse default payment data, and compare the prediction accuracy and classification ability of different machine learning methods and various neural-network methods, with a different activation function. Their results indicate that the

classification ability of boosting is superior to other machine-learning methods, including neural networks.

The issue of COVID-19 impact on the credit quality of loan portfolios has not yet been investigated in depth in literature. Most of the studies focus on sovereign risk assessment, as Augustin et al (2022), which found a positive and significant sensitivity of sovereign default risk to the intensity of the virus spread for fiscally constrained governments. Other studies, focus on bank risk directly, such as Aldasoro et al (2020), which detected differentiations in the virus impact (measured in CDS spreads) based on nationality and the level of risk that each bank had when entering the crisis.

In our study, we leverage on the recent application of Deep Learning methods on credit risk evaluation, trying to approach the effects of the COVID pandemic on the performance of bank loan portfolios assessed on a granular (loan by loan) level. The granularity of the dataset allows us to detect various behavioural impacts and isolate the impact from the COVID-19 outbreak on the corporate portfolios of Greek banks.

3. Data collection processing and variable selection

AnaCredit (Analytical Credit datasets) is a dataset with granular information on individual bank loans in the euro area, including Greece, which is provided by the banks via a standardised set of templates. The dataset used in our analysis comprises of loan level information on a monthly basis on Corporate and SME loans taken from the AnaCredit database for all Greek significant and less significant institutions.

The adopted definition of a default event in this dataset is in line with IFRS 9. Specifically, a default event is considered if the instrument becomes credit impaired in accordance with IFRS 9 (i.e., Stage 3), with a six-months observation period. At each monthly snapshot, all performing loans (i.e., Stage 1 or Stage 2) are considered, and at the end of the 6-month observation period they are classified as either performing or non-performing according to their IFRS 9 stage. The dependent variable in our dataset is a binary indicator (0 or 1), with the value of one flagging a default event (i.e., the instrument is categorized as non-performing at the end of the 6-month observation period).

The available dataset covers a period of 3.5 years, from September 2018 to February 2022. The final dataset contains more than 3 million of performing instruments, sufficiently covering both the period before and after the COVID-19 outbreak.

Figure 1 shows the number of performing instruments included in each snapshot and the corresponding Default Rate (DR). Before the outbreak of the COVID-19 pandemic the Default Rates had a continuously decreasing trend because of the improved financial conditions prevailing in Greece, reaching levels below 2%. COVID -19 pandemic impacted the credit quality of the loans and a new stock of defaults emerged. The Default Rates elevated again and approached the level of 4%. The latest observations shows that the COVID-19 impact in Default Rates is constantly fading, indicating that loan portfolios are shaking off the COVID-19 impact.

[Figure 1]

To perform the modelling and prediction methodology five different set of dependent variables were used.

The first set contains indicators related to basic macroeconomic variables. Specifically, the Gross Domestic Product (GDP) changes, the unemployment rate level, and the House Price Index were used.

The second set of independent variables used are related to obligors' information. For each obligor the Global NPE ratio is calculated as the ratio of stage 3 exposures over total exposure in the Greek banking system. In addition, for each obligors the level of the aggregated loans received from Greek institutions as well as the overall level of the company's turnover are determined. From these two metrics three independent variables are derived. Firstly, a four-cluster segmentation is used, and all companies are classified in one of the 4 turnover clusters. It has to be mentioned that the level of a company's turnover provides an estimation about its size. Secondly, with respect to the aggregated loan level information, the 6-month change is measured in order to capture whether a company increases significantly its exposure to the Greek banking system or not. Thirdly, the ratio of turnover over aggregated loan volume is computed to capture the level of the obligor's leverage.

The final dataset also contains a set of behaviour variables, including, the IFRS 9 stage status, whether the instrument is forborne or renegotiated, and if any legal actions

is taken by the bank. These variables were transformed to binary indicators (0 or 1) and coupled with their 6 months history (i.e., lags) were included in the final dataset. Furthermore, the economic sector (e.g., Constructions, Real Estate, Accommodation, etc.) is transformed to binary indicators (0 or 1) and is also considered in the analysis. The available granularity of the dataset permits a more thorough analysis of sectoral sensitivities during the COVID-19 pandemic.

In addition, another set of indicators employed are related to instrument specific information, such as duration of the loan, interest rate level, institution granting the relevant instrument, payment frequency of the instrument, instrument type (e.g., overdraft), and purpose (e.g., working capital, commercial real estate purchase etc.).

Finally, two variables related to COVID-19 pandemic are calculated. In March 2020, the Greek Prime Minister announced that a nationwide lockdown will come into force for first time to restrict the spread of the coronavirus disease. This date is considered as the starting date of COVID-19 in Greece, from a financial perspective. For each snapshot we estimate the number of months elapsed since the COVID-19 outbreak. In addition, we assume that most of the financial support measures taken in Greece started concurrently with the outbreak and ended in December 2020. An additional variable that is considered in the final dataset indicates whether at the date of examination there are financial support measures in place or not.

The combination of the aforementioned sets of independent variables, after excluding the ones causing multicollinearity issues using variance inflation factor, led to a final set of 92 independent variables (shown analytically in the Appendix 1), with more than 3 million entries, so classifying the available dataset to the "big data" category. The so-obtained dataset was split into two parts:

- An in-sample train dataset that includes the data between March 2019 – June 2021, which was used for model development (Train Set).

- An out-of-time dataset that includes the data between July-August 2021 (marked in green in Figure 1), which was employed for validation purpose and testing of the generalization capacity across all candidate models (Test Set).

4. Model development

Due to the extended number of predictors and the large-scale dataset employed, we apply methodologies from the general domain of Machine Learning techniques. Big data often entail dimensionality issues, increased noise, convergence issues, and other significant statistical challenges, which cannot be addressed based on traditional statistical techniques. Thus, we investigate the predictive performance of a group of state-of-the-art machine learning techniques in the classification problem of future obligor default. These techniques are benchmarked against traditional statistical techniques employed in small scale probability of default modelling (i.e., Logistic regression (Logit)).

Five techniques are employed in this work, namely Random Forests, Extreme Gradient Boosting, Deep Neural Network, Shallow Neural Network, and Logistic Regression.

Random Forests (RFs) (Breiman, 2000) are frequently used in many machine learning applications across various fields of the academic community, and it is a popular method for modelling classification problems. RFs combine bootstrap aggregation and random features' subspace selection to generate a forest of trees. Their structure follows a divide-and-conquer approach used to capture non-linearity in the data and perform pattern recognition. Its core principle is that a combined group of "weak learners" can form a "strong predictor" model. In more detail, RFs combine many binary regression decision trees that are selected by bootstrapping samples of the modelled explanatory variables and the corresponding classifier variable. The final prediction is made using the majority voting in case of classification problem, or by averaging the predictions from all the individual trees in case of regression problems. The final set of random forest variables is selected using a variable importance index, which reflects the "importance" of a variable based on its contribution to classification accuracy. This is estimating by looking at how much the prediction error increases when omitting a considered variable. Our implementation of RFs was performed based on Python's sklearn package. We performed a grid search procedure over a 5-fold cross-validation to select a series of entailed hyper parameters, including: the number of decision trees in the forest, the maximum number of features considered to split a node, whether bootstrap samples should be used when building trees, the maximum depth of the tree, the minimum number of samples required to split an internal node,

the minimum number of observations required to be at a leaf node, and the function to measure the quality of a split. The final model includes 300 decision trees in the forest, up to 7 features considered to split a node, at least 50 observations at each leaf node, while the Gini metric was used to measure the quality of the splits.

The extreme Gradient Boosting (XGBoost) is a boosting tree algorithm (Chen et. al. 2016) that is an enhancement over tree bagging methodologies, such as Random Forests. 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 it 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. Extreme Gradient Boosting (XGBoost) 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 hyper-parameters, while it fully supports continuous training.

We implemented XGBoost by utilizing the XGBoost package for Python. We performed a 5-fold cross validation grid-search procedure to fine-tune and optimise our model. A series of entailed hyper parameters were included in the tuning procedure, such as the maximum number of trees generated, the maximum depth of trees, the learning rate, the L1 regularization (Lasso Regression) parameter on the weight (to avoid overfitting issues), 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 regression with squared loss, while the Mean Squared Error (MSE) was used to evaluate the performance of the cross-validated model. The final model has a max depth of 10 per each tree, a learning rate of 0.01, which determines a relatively small step size at each optimization iteration. There are 500 trees in our ensemble. The fraction of columns to be randomly sampled for each tree was set at 0.3, and the fraction of observations to be sampled for each tree was at 1. The alpha hyper-parameter (i.e., L1 regularization on the weights) was also set to 1 to reduce overfitting tendencies.

Furthermore, we implemented a Deep Neural Network (DNN) (LeCun et al., 2015; Heaton, 2018). 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. DNN have been extremely successful in tasks such as diverse time-series modelling, 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, so that our paper constitutes one of the first works.

Deep Neural Networks are characterised by the multiple internal layers employed between the input values and the predicted result, and differ from Shallow Neural Networks, which have one layer (Figure 2). Constructing a DNN without nonlinear activation functions is impossible, as without these functions the deep architecture collapses to an equivalent shallow one.

[Figure 2]

Since DNNs require 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 is 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 contains the associated network overfitting tendency, and it does so in a way that ensures that all network parameters are effectively trained.

We employ Dropout DNNs with Sigmoid activation function to train and deploy feed forward deep neural networks. More precisely, we employ TensorFlow package and the Keras API (Application Programming Interface) for Python. We postulated deep networks that are up to four hidden layers deep and comprise various numbers of neurons. Model selection using cross-validation was performed by maximizing the Accuracy metric. The trained model was compiled based on the Adam optimizer, which is considered the best among the adaptive optimizers. Binary Cross-Entropy loss was

used as the Loss Function of the model, comparing the predicted probabilities with the actual class output.

Dropout is implemented between layers, by randomly selecting the nodes to be dropped-out with a given probability (10%). Due to its size, the dataset cannot be processed all at once, so the data is split into smaller batches (we set the batch size to 500 observations). Training occurs over epochs, and it refers to the maximum number of passes through the entire training set. Epoch is set to 500, that is, the training data is used up to 500 times). The model configuration includes an early stop option, which stops the training when there is no increase in Accuracy for 10 sequential epochs.

As already mentioned, the final model was selected by maximizing the Accuracy metric. Figure 3 provides a visualization of the training process. They show that the loss is decreasing while the accuracy is increasing in each epoch, in both the train (blue lines) and test samples (green lines). It is noted that the training did not reach the maximum number of Epochs, due to early stopping.

[Figure 3]

The structure of the developed DNN model is shown in Figure 4. Specifically, there are 92 features that are fed into the model (input layer). The first hidden layer has 256 nodes (neurons/hidden units/outputs), the second layer has 128 nodes, the third layer has 64 nodes, and the forth hidden layer has 32 nodes. The output layer has one node. There is dropout (deactivation of nodes) between each layer, and the sigmoid activation function is used in each layer.

[Figure 4]

A Shallow Neural Networks was also developed, with the same specifications to the Deep Neural Network, described above. However, the Shallow Neural Network employees one hidden layer, compared to the four hidden layers of the Deep Neural Network.

We benchmark the abovementioned techniques versus traditional statistical techniques employed in probability of default modelling, i.e., Logistic regression (Logit). 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.

We developed a simple logistic regression model, without any AR or time-varying model parameters, and we implemented this approach in Python using the statsmodel module. The final Logit model include 21 features, which are presented in Appendix 4. Before estimating the logit model, we performed univariate feature selection (dropping 37 variables with low predictive power, i.e., low correlation with target variable), and then dropped collinear variables based on correlation cut-off threshold of 70% (18 variables). The correlation matrix before and after the filtering is shown in Figure 5.

[Figure 5]

5. Model performance evaluation

We performed a thorough validation procedure to assess the robustness of the above-mentioned models, in terms of in-sample (Train sample) as well as out-of-time (Test sample) performance. We employed a series of metrics that are broadly used, by both researchers and practitioners, to quantitatively measure the performance of a credit scoring model (Hossin and Sulaiman, 2015). We used various metrics to assess the discrimination power, the classification accuracy, and the predictive ability of the models. These measures are used to draw a full spectrum conclusion on the performance of each model relative to the others.

In more detail, we assess the discriminatory power of the evaluated models using the Area Under the Receiver Operating Characteristics (AUROC) metric (equivalent to the Gini metric), and the K-S metric. The results, for both Test and Train samples, are presented in table 1, where it is clearly noted that the Deep Neural Network (DNN) outperforms all other models.

[Table 1]

The ROC curves (and the corresponding AUROC metrics) are presented in Appendix 2, for all models and all samples. The ROC 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 area under the ROC curve is calculated and the AUROC metric is produced.

Table 2 presents a series of additional tests that were performed to measure the classification accuracy of the models, including the Accuracy, Precision, Recall, and ROC-AUC (Receiver Operating Characteristics - Area Under the Curve) metrics, in a binary set-up using appropriate cut-off scores for the predictions. The classification tables, for both Train and Test samples, are presented in Appendix 3.

[Table 2]

Based on the above results, DNN exhibit the best performance across all metrics, in both samples. With respect to the recall metric, it is intuitively the ability of the model to find all the positive observations (i.e., the defaulted obligors in our case). Since there is a much higher financial risk when insolvent customers are not identified, rather than when solvent customers are wrongly classified as insolvent, the proper identification of insolvent obligors is a desired characteristic of a scoring model. Thus, we consider that recall is an important metric in our case. It is also noted, that XGBoost performs second.

Finally, the predictive accuracy of the models was assessed based on the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and the Root Mean Squared Error (RMSE) metrics. The results are presented in Table 3 for both Test and Training samples, indicating again the superior performance of DNN.

[Table 3]

Deep Neural Network (DNN) outperforms all other models, along all measures, in both Test and Train samples. This performance consistency implies a much stronger generalization capacity compared to other state-of-the-art models, which renders our approach much more attractive to researchers and practitioners working in real-world financial institutions.

Another 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 credit risk monitoring purposes in the core banking operations. 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. For this purpose, we

replicate the credit rating system calibration process for all models in scope, by using the sklearn.calibration library in Python.

The calibration curves (also known as reliability diagrams) compare how well the probabilistic predictions of a binary classifier are calibrated. Calibration curves plot the true frequency of the positive label against its predicted probability, for binned predictions. The x axis represents the average predicted probability in each bin. The y axis is the fraction of positives, i.e., the proportion of samples whose class is the positive class (in each bin). A perfect model would be represented by a diagonal line, where the predicted probability would be the true frequency for each bin. Thus, the calibration curves plot the estimated probability of default, versus the actual default rate per bin. Each bin can be seen as a rating grade, which is defined based on each model output (i.e., range of score).

The calibration curves produced for each model, both for Train and Test samples are shown in Figure 6. The calibration of the DNN model is superior compared to the other models, as the DNN calibration curve (red line) is closer to a perfectly calibrated model (blue dotted line), in both samples. The rest of the models show significant weaknesses with respect to calibration. Particularly, SNN and LR models produce a poor calibration, as their calibration curves (purple and brown lines, respectively) are far from the perfectly calibrated model.

[Figure 6]

The predictive accuracy (per bin) of each calibration was also assessed based on the Mean Squared Error, in order to have a quantitative assessment of the calibration that can be produced by each model. Once again, DNN produce the lowest MSE in both samples as inferred from Table 4.

[Table 4]

6. Interpretability

The major criticism of Machine Learning algorithms is related to their lack of transparency. There was always present a trade-off between model complexity and model performance. Although those models have a wide range of applications and they are able to handle with big data, and recognize artificial trends and patterns, the lack of

interpretability renders them sometimes powerless as they are considered black boxes. As interpretability is of paramount importance, some of the model-agnostic approaches are presented in the following subsections. The goal is to apply both global and local interpretability algorithms to provide more clarity on the functionality of the DNN.

6.1 Local surrogate (LIME)

Local surrogate models (Ribeiro et al., 2016) are easily interpretable models (e.g., Linear Regression) that are used to provide transparency on individual predictions of a complex black-box model. To explain the prediction of a specific instance (point x), a new dataset consisting of perturbed samples (around the point x) is generated. Then, this new dataset is used in the "black-box" model and the relevant predictions are estimated. The same new dummy dataset is used but now using a simpler and easily interpretable model, such as a linear regression. The interpretable model for a specific instance is the model that produces predictions closer to the predictions estimated by the "black-box" model.

We use LIME library in python to perform the aforementioned local interpretation. For a series of instances, we produce the relevant instance explanation, some of which are presented in Figures 7-9.

[Figures 7-9]

Locally, the Global NPE ratio feature, which is defined as the ratio of stage 3 exposures over total exposure in the Greek banking system, is the most important feature in the selected cases. Other features that appeared to be locally important are the information of IFR9 stage, and the COVID-19 variable. In the abovementioned presented examples, information about the sector of the obligor and the creditor institution, appears to be an important feature.

Having obtained a first flavour of the hidden dynamics with respect to the features of the Deep Neural Network, using local interpretation of the model, we proceed by providing some extra clarity using Global Model Agnostic techniques.

6.2 Shapley Value (SHAP)

In the Shapley Value approach (Fryer et al., 2021), the contribution of each feature is measured by adding and removing the specific feature from all subsets of the rest of the features. To compute the contribution of a feature i, a model is trained with

the feature i present, and another model is trained with the feature excluded. Then, predictions from the two models are compared on the current input. The differences are computed for all possible features subsets. The Shapley value is the average marginal contribution of a feature value across all possible set of features.

To retrieve the Shapley values for models under investigation we have used SHAP package in python. Since Shapley value estimations are considered very time expensive, various random samples from the train dataset have been selected and the respective available graphs have been produced. To produce the SHAP feature importance graphs, the absolute Shapley values are estimated. Figure 10 provides a global interpretation, since values used in the feature importance graph are the result of the aggregation of the SHAP values for individual instances across the entire selected sample population. Then, the features are ordered from the highest to the lowest effect on the prediction, ignoring whether the feature affects the prediction in a positive or negative way.

[Figure 10]

An alternative and more information-enriched option to present the results of the Shapley values analysis is the Beeswarm plot (Figure 11). Along with the feature importance information (ranking) the x-axis value of the dot is determined by the Shapley value and shows whether the effect of that value is associated with a higher or lower prediction. Colour is used to display whether the original value of the feature is high (red dots) or low (blue dots). Combining the colour information with x-axis point, we can infer the relationship between the level of the feature and the level of the prediction. For instance, a Stage 2 loan is more prone to default comparing to a Stage 1 loan, according to the DNN. Stage 2 classification, which by itself signals increased credit risk, appears to be the most important factor (contribution of around 15%) in explaining default. Said that, the remaining features account for the remaining 85% of the total contribution, revealing the additional factors that do matter for predicting a firm’s default.

[Figure 11]

In addition, loans disbursed in the past carry lower probabilities to default comparing to new disbursements. We observe that from a credit risk perspective, portfolios of large banks (Banks 1-4, refer to SIs) behave better comparing to smaller

institutions. It is also evident that loans with higher level of interest rate across all institutions are more prone to a potential default. Furthermore, the Deep Neural Network model identifies the relationship of a potential credit event with respect to the level of the company's leverage, through the turnover over loan variable. Specifically, the greater the leverage (blue dots) the greater the possibility an instrument to migrate to stage 3 in the following 6 months.

COVID-19 variable is present to all models investigated, including the Deep Neural Network model. We proceed with performing a series of partial dependency plots in the following section to shed light on the Deep Neural Network "black-box" with respect to the COVID-19.

6.3 Partial Dependency Plot

Partial dependence plots (PDPs) are used to visualize the relationship between the dependent variable and a set of independent variables (features). Through a partial dependence plot the behaviour pattern between the target variable and a feature can be revealed regardless if this relationship is linear, monotonic or more complex. Assuming that X_s is the set of independent variables and X_c its compliment, then partial dependence plot (Goldstein et al., 2015) at a point x_s is defined as:

$$pdp_{X_s}(x_s) \stackrel{\text{def}}{=} E_{X_C}[f(x_s, X_C)] \qquad (1)$$

The partial function provides the average marginal effect on the prediction for a given value of features s. We present in Figure 12 the partial dependency plots for the feature related to COVID-19 across all models under investigation. The COVID-19 feature counts the number of months elapsed since the pandemic outbreak.

[Figure 12]

Apart from Logistic Regression model, which is not able to capture non-linear relationships, the other models recognise an inverse smile pattern between the average marginal effect on the probability of default and months since the start of the pandemic. Excluding the Logistic Regression from Figure 12, the previously mentioned nonlinear pattern is even more evident in Figure 13.

[Figure 13]

The increase of default events due to the COVID-19 outbreak is being captured by all models. All models, excluding shallow DNN, exhibit a less sensitive relationship in the beginning of the period, presumably due to the introduction of support measures taken in Greece (e.g., moratoria). In the long end of Figure 13, the impact of COVID-19 shock started fading, reflecting the reopening of economy and the end of lockdown. It is evident that any effects from COVID-19 are still present, retaining default probabilities to higher levels than before the pandemic.

As DNN exhibit superior performance against the rest of the candidate models, we continue the PDP analysis focusing only on DNN. The default probabilities estimated through the DNN model are presented along with the actual default rates with respect to the months elapsed since COVID-19 outbreak in Figure 14. It evident that DNN models can capture the actual default rate pattern.

[Figure 14]

European Commission highlighted in the technical note to the Eurogroup "Sectoral Impact of the COVID-19 crisis" that the pandemic impact varied dramatically across sectors of the economy. Using the following PDP that show the expected target response (i.e., prediction of default probabilities) broken down by sector, as a function of the number of months elapsed since the COVID-19 outbreak, we observe in Figure 15 that the average marginal effect on the default probabilities differentiates significantly across sectors. In Greece, accommodation and entertainment sectors were severely affected in the first year of pandemic, while energy sector remained less vulnerable in the whole period. Transportation and storage sector appears stickier comparing to other sectors in Greece, mainly due to the mix of companies included in our dataset (e.g., shipping, including seagoing, is reported under this sector).

[Figure 15]

As expected, stage 2 loans are more sensitive to the deterioration of the credit risk environment. This is illustrated in Figure 16, in which it can be observed that the increase in default rates is more pronounced. In tandem, the default rate of stage 2 loans remained rather stable in the first months, reaching their maximum level a few months later comparing to stage 1 loans. This presumable being explained by the fact that moratoria provided by Greek banks targeted to stage 2 loans and offered a short-term delay in the increase in the default rates.

[Figure 16]

Useful insights are provided by the fitted DNN also for other features of AnaCredit that are considered important from a credit risk perspective. As shown in Figures 17-19, it is evident that the proposed DNN model can capture the non-linear relationship between these variables and the default rate. Specifically, for the duration of the instruments there is abrupt increase in the default rates during the first 5 years (60 months on book). With respect to the feature turnover over loan, a leverage indicator, a convex relationship is observed that asymptotically flattens above 50%. Finally, for Global NPE ratio an increasing monotonic behavioural pattern is observed.

[Figures 17-19]

7. Conclusion

To tackle the issue of pattern detection in large loan level datasets we employ machine learning algorithms that reduce dimensionality in the data and increase accuracy in predicting the future behaviour of corporate loans. Our analysis is based on the Greek AnaCredit dataset, spanning a 42-month period of the Greek economy. The purpose of the analysis is to perform credit quality classification and quantification of Probability of Default during the COVID-19 pandemic. To achieve this, we develop five behavioural credit scoring models, for which we perform extensive comparisons of the classification and forecasting accuracy using an out-of-time sample of 2 months period.

Our empirical results indicate that DNNs provide better performance in terms of classification accuracy and credit rating system calibration across all metrics, compared to widely employed techniques in credit risk modelling such as Random Forests, XGBoost, and Logistic Regression. In addition, the inclusion of macro, financial, and transactional-behavioural variables captures both the systemic and idiosyncratic behaviour in obligors' credit quality, thus, both discriminatory and calibration testing exhibit stability in performance. Our analysis provides strong evidence for the model's increased stability, as the high-performance levels observed in the in-sample dataset are retained when evaluation is performed using out-of-sample datasets. The performance consistency identified implies a much stronger generalization capacity compared to the state-of-the-art models, which renders DNNs much more attractive to researchers and

practitioners working in real-world financial institutions. Deep Neural Networks appear to capture the whole nexus of information that lies in the AnaCredit dataset, and thus, outperform all candidate statistical models assessed in this paper.

To provide a more concise view of the impact of COVID-19 in the corporate and SME default risk, we make use of the fitted DNN. Using partial dependence plots on the proposed DNN we provide evidence on the sectoral sensitivities with respect to the probability of default. Accommodation and entertainment sectors suffer a more intense impact during the first year of pandemic, while energy sector remain less vulnerable throughout the sample period. With respect to credit quality, stage 2 loans exhibit higher sensitivity, with a sharp increase in default rates that peak a few months later than stage 1 loans. Furthermore, our empirical results indicate that the moratoria offered a short-term delay in the increase of the default rates due to the pandemic. The removal of lockdowns and the reopening of the economy have eased the COVID-19 effect, as probabilities of default have steadily started to decrease without yet reaching the pre-pandemic levels.

This study provides evidence that DNNs can be the base for building the next generation of supervisory tools for monitoring and modelling credit risk in the short run. Our proposed approach is a fully-fledged automated system that can be used by financial experts in central banks for quantifying credit risk, make projections in the short run, and potentially drive decision making. Said that, we recognize that one caveat of our analysis is that we have not assessed the out of time performance of our models for a longer period (i.e., done only for two snapshots), and this is something to be done in the future once data are available. Furthermore, our model could be further complemented by financial ratios as explanatory variables in predicting default, once these become available to us. Though, we would like to note that building a financial rating system, which makes full use of financial ratios, is different from a behavioural scoring model. Finally, as a future research step we intend to embed re-enforcement learning algorithms, which increase the capabilities of the DNN behavioural credit scoring to adapt more quickly to new patterns introduced in loan level data.

Tables and Figures

Figures

Figure 1: Greek banking system instrument level Default Rate evolution

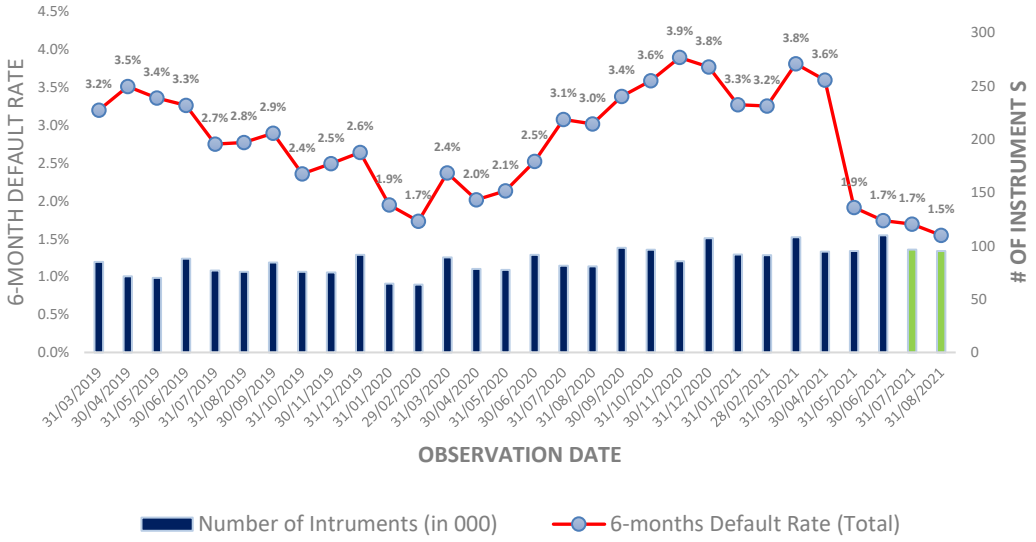


Figure 1 presents the evolution of the default rate along with the respective performing exposures through the period examined. DR is calculated as: DR_t = $DefaultFlows_{t,t+6}/PerformingExposures_t$, where t denotes the respective month-end.

Figure 2: Shallow and Deep Neural Networks

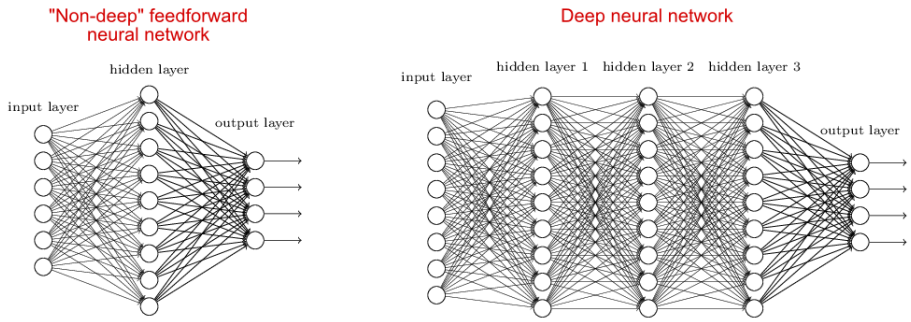


Figure 2 illustrates the structure of a non-deep/shallow vs. a deep neural network.

Figure 3: DNN - visualization of the training process

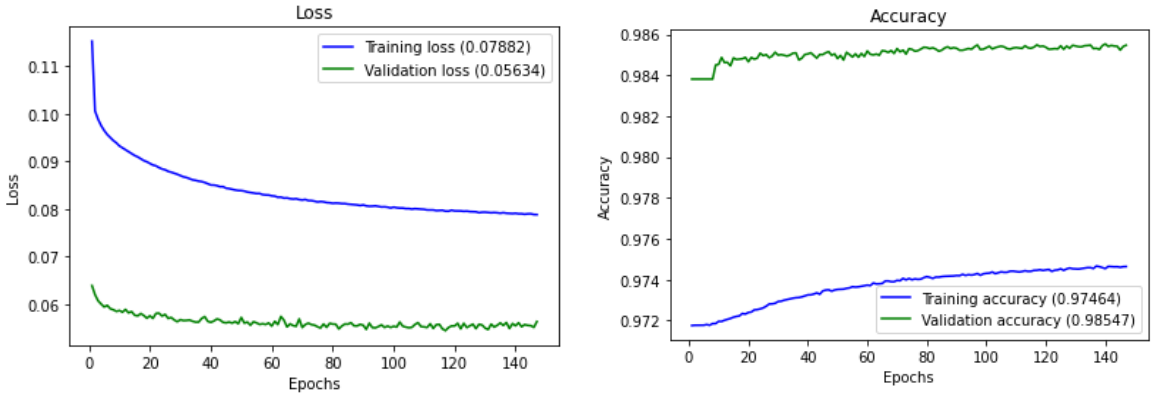


Figure 3 presents how loss and accuracy metrics, both in the training and the validation samples, evolve as the number of epochs increase. One Epoch is when the entire dataset is passed through the neural network once.

Figure 4: DNN model structure

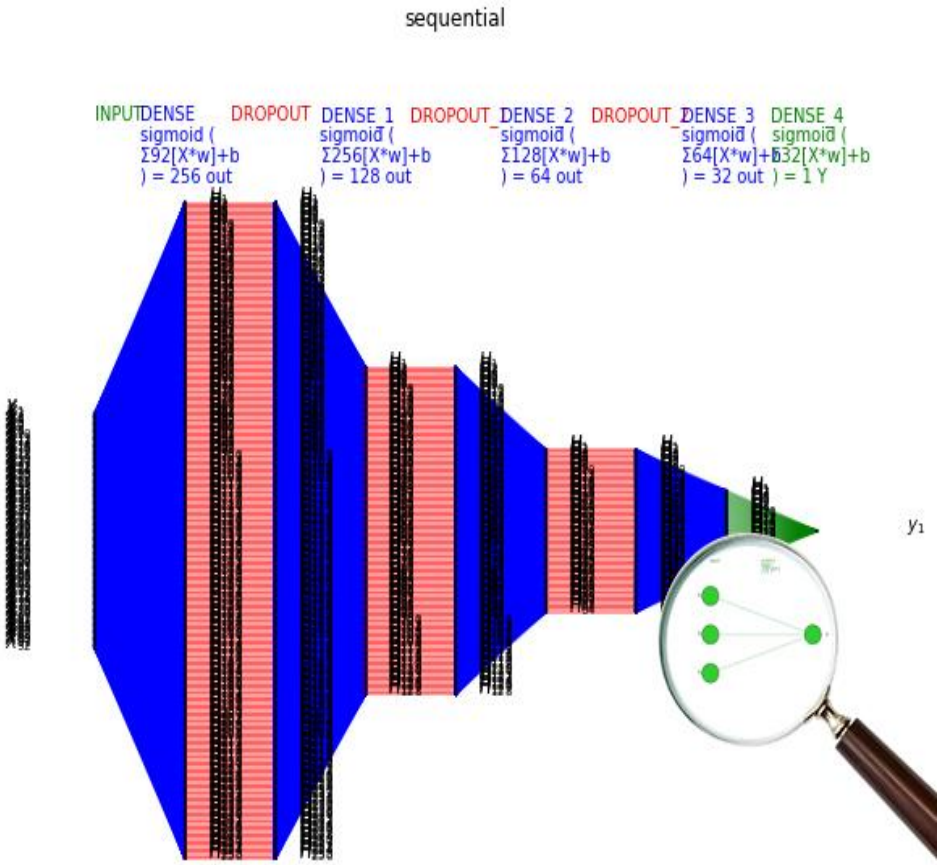


Figure 4 depicts the structure of the DNN which contain one input layer, one output layer and 4 hidden layers.

Figure 5: Correlation Matrix (for Logit development)

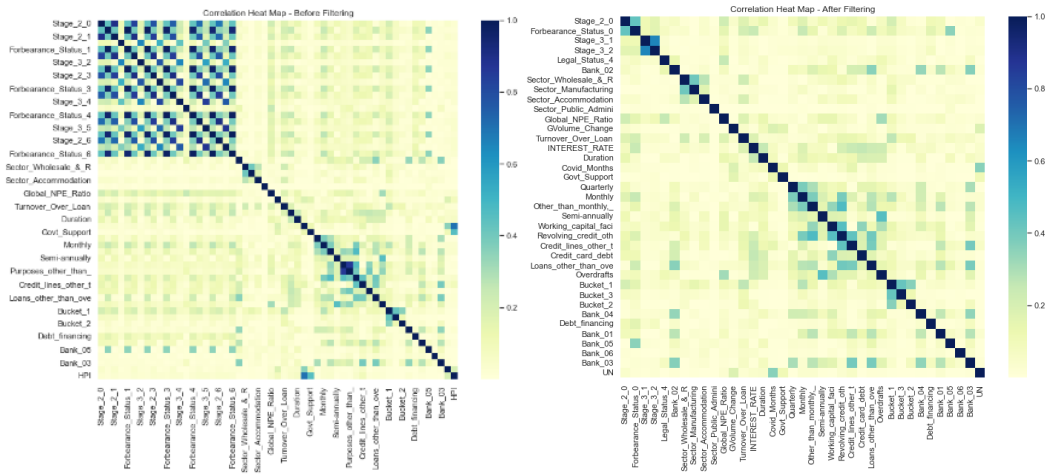


Figure 5 exhibits the correlation coefficients of the regressors to be used in the Logit model. The panel to the left includes all features assessed, while the panel to the right exclude collinear variables.

Figure 6: Calibration Curves

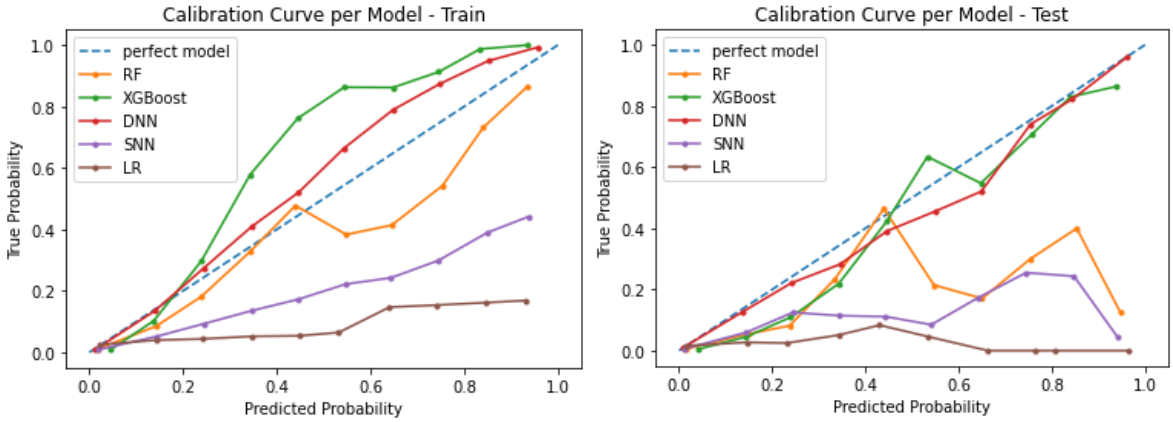


Figure 6 presents the calibration curves for the models developed, both for the test (out-of-time) and for the train samples. The close the line for each model to the diagonal line, the better the performance of the model.

Figure 7: Explanation Graph (Deep Neural Network) – Instance #1

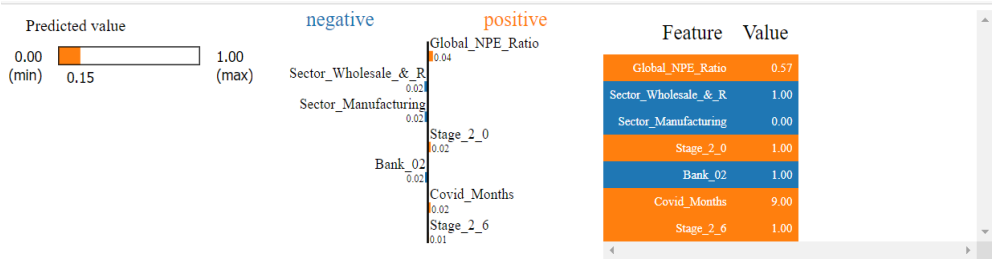


Figure 7 presents for an instance selected under LIME: the predicted value of the model (left part), the sensitivities of its features in deriving the respective predicted value (middle part), and the feature values (i.e., instance attributes) at the right part.

Figure 8: Explanation Graph (Deep Neural Network) – Instance #2

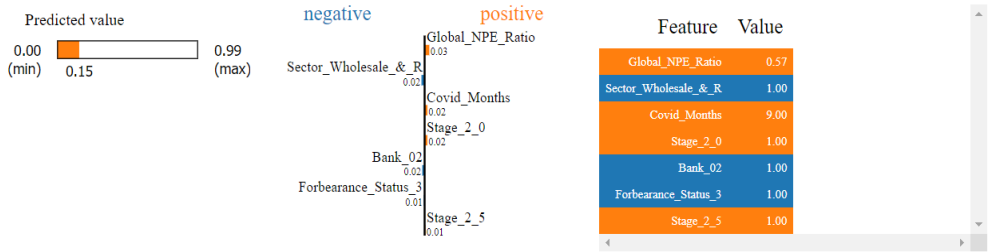


Figure 8 shows for an instance selected under LIME: the predicted value of the model (left part), the sensitivities of its features in deriving the respective predicted value (middle part), and the feature values (i.e., instance attributes) at the right part.

Figure 9: Explanation Graph (Deep Neural Network) – Instance #3

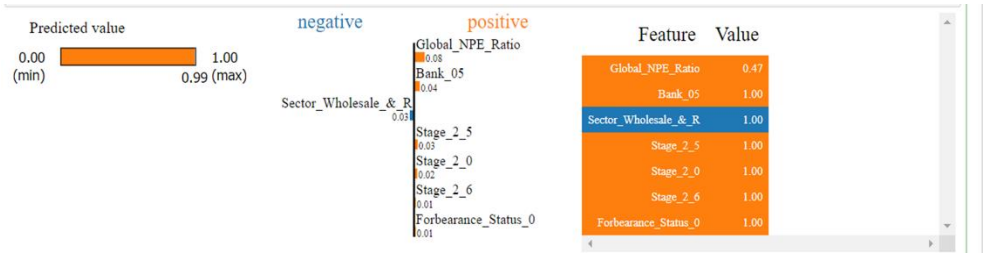


Figure 9 exhibits for an instance selected under LIME: the predicted value of the model (left part), the sensitivities of its features in deriving the respective predicted value (middle part), and the feature values (i.e., instance attributes) at the right part.

Figure 10: Feature Importance Graph (Deep Neural Network)

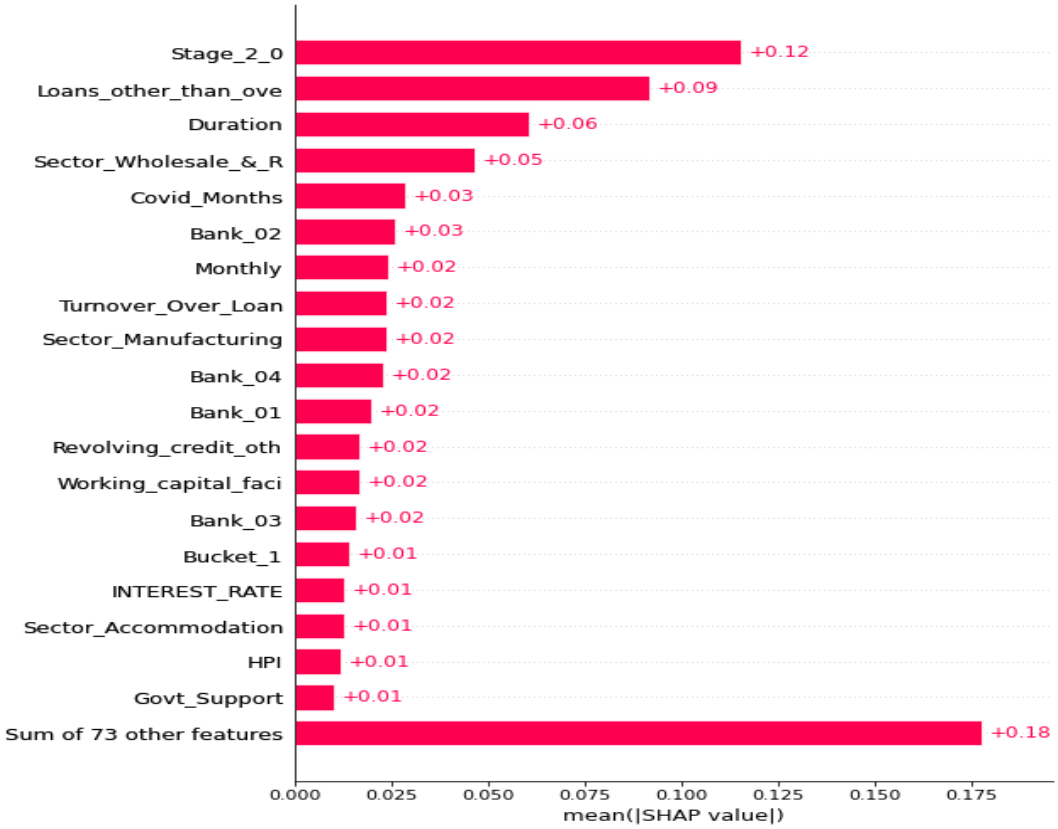


Figure 10 presents SHAP feature importance values, which reflect the magnitude (in absolute terms) of the most important explanatory variables in determining the target variable.

Figure 11: Beeswarm Graph (Deep Neural Network)

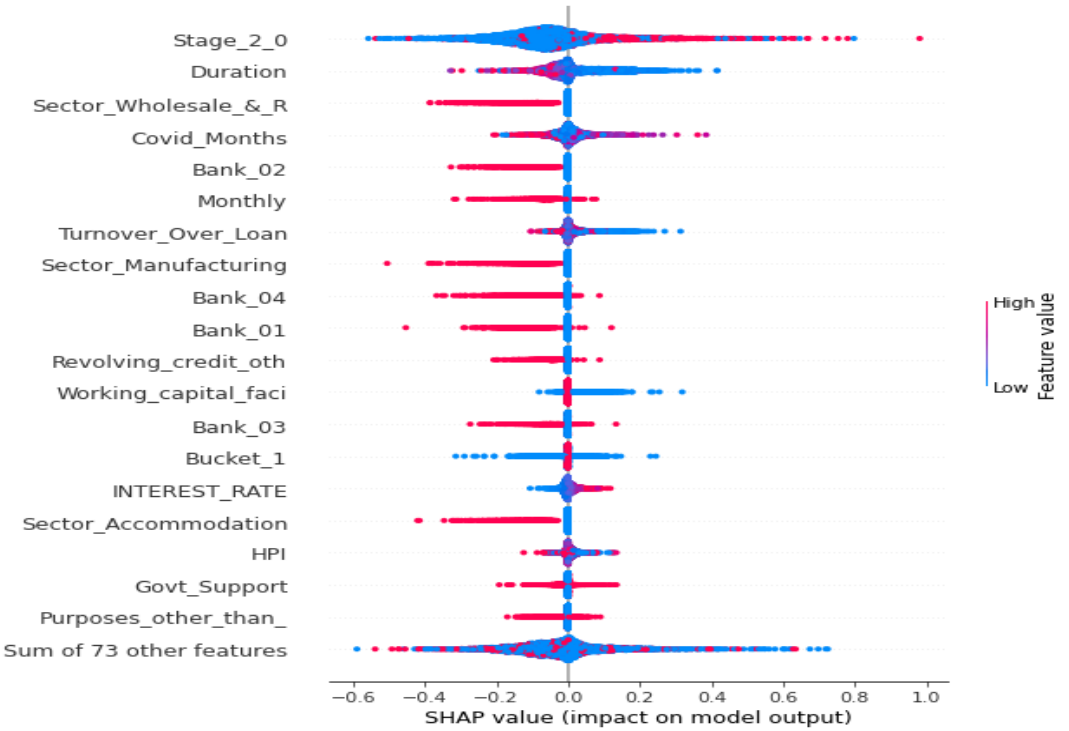


Figure 11 exhibits the Beeswarm plot, in which the x-axis value of the dot is determined by the Shapley value and shows whether the effect of that value is associated with a higher or lower prediction. Red dots indicate whether the original value of the feature is high, while blue dots indicate whether the original value of the feature is low.

Figure 12: PDP for COVID-19 feature (All models)

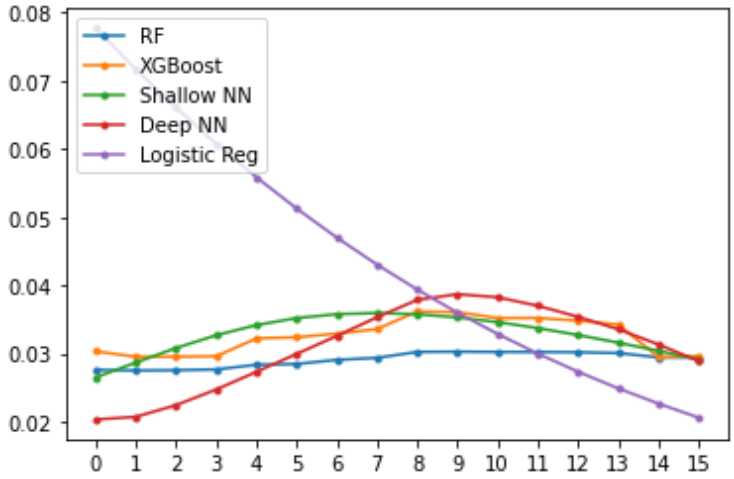


Figure 12 displays how the impact of the COVID-19 feature affect the probability of default across all models developed. Specifically, y-axis presents the marginal impact on the probability of default, while x-axis presents the number of months elapsed since the COVID-19 outbreak.

Figure 13: PDP for COVID-19 feature (excluding Logistic Regression)

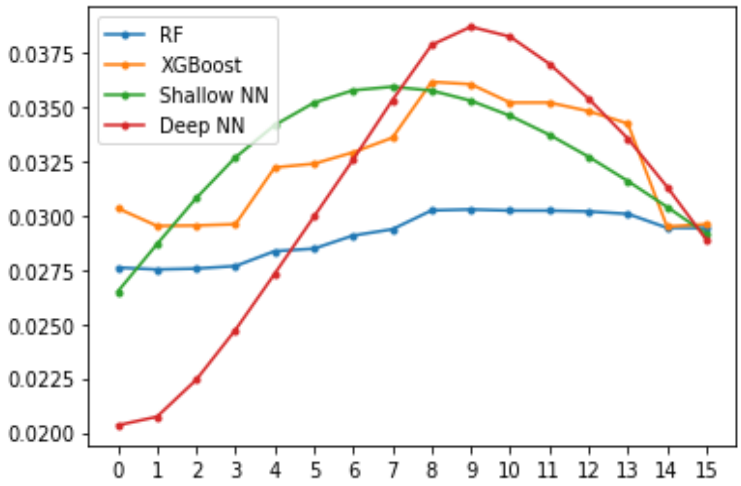


Figure 13 shows how the impact of the COVID-19 feature affect the probability of default across all models developed, apart from logistic regression. Specifically, y-axis presents the marginal impact on the probability of default, while x-axis present the number of months elapsed since the COVID-19 outbreak.

Figure 14: Estimated DNN PDs vs. actual Default Rates (by COVID-19 months)

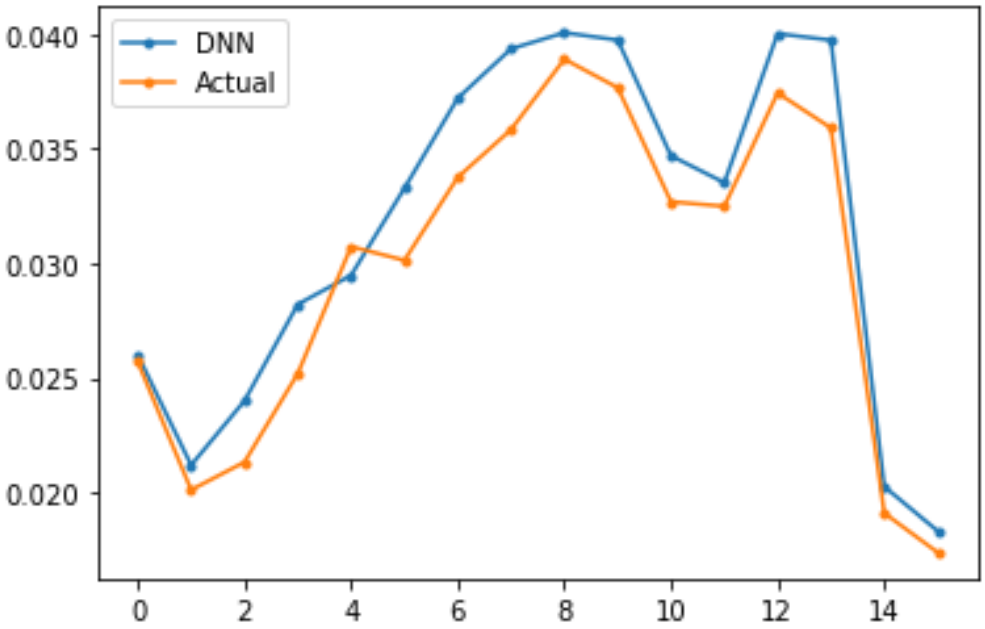


Figure 14 presents the evolution of the default probabilities estimated through the DNN model and of the actual default rates, with respect to the months elapsed since COVID-19 outbreak. Specifically, y-axis presents the probability of default, while x-axis present the number of months elapsed since the COVID-19 outbreak.

Figure 15: DNN – PDP – Sector Differentiation in Greece (by COVID-19 months)

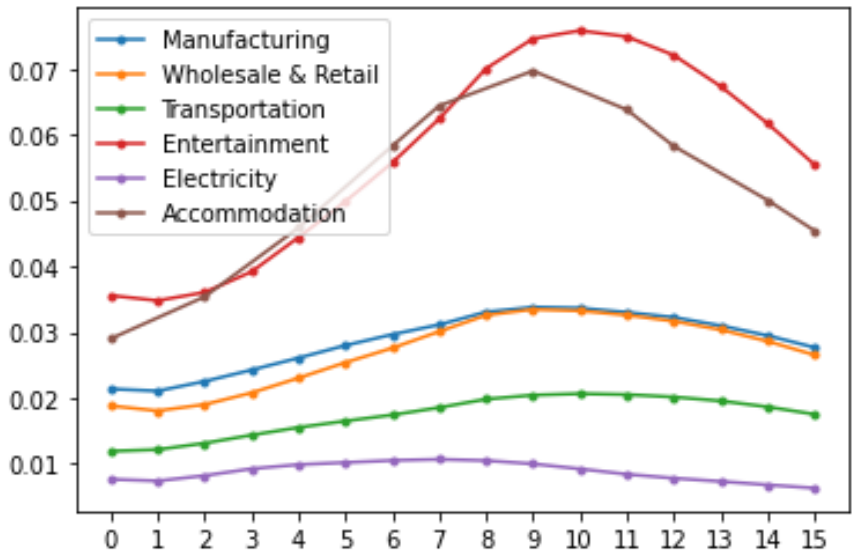


Figure 15 shows how the impact of the COVID-19 feature affects differently the probability of default for corporations that belong to different sectors, based on the DNN models. Specifically, y-axis presents the marginal impact on the probability of default, while x-axis present the number of months elapsed since the COVID-19 outbreak.

Figure 16: DNN – PDP – Stage 1 vs. Stage 2 (by COVID-19 months)

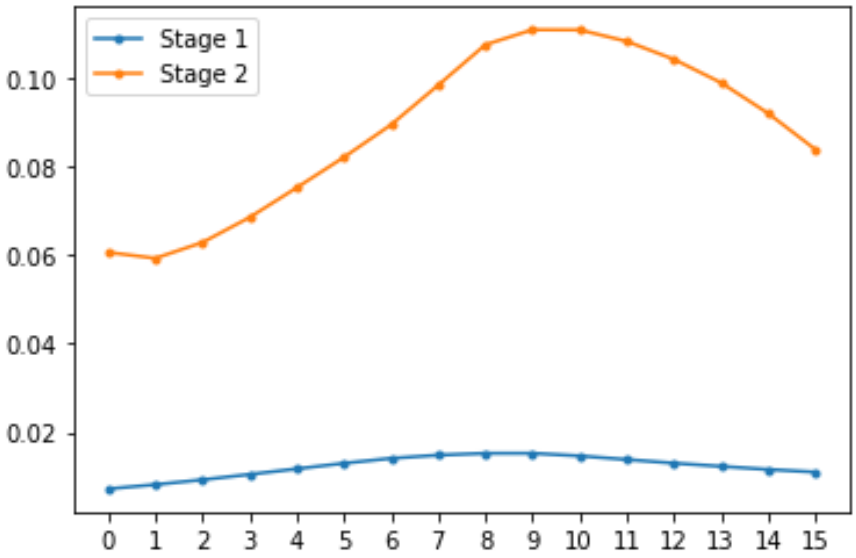


Figure 16 illustrates how the impact of the COVID-19 feature affects differently the probability of default for corporations that belong to Stage 1 vs Stage 2, based on the DNN models. Specifically, y-axis presents the marginal impact on the probability of default, while x-axis present the number of months elapsed since the COVID-19 outbreak.

Figure 17: DNN – PDP – Duration

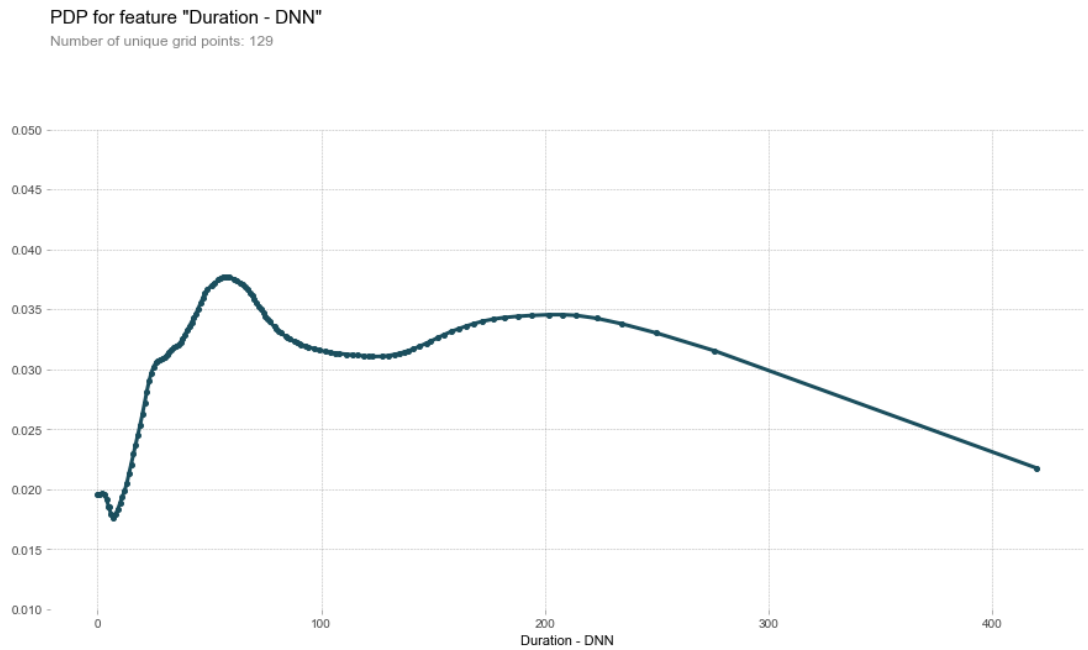


Figure 17 exhibits how the impact of the loan "Duration" affects the probability of default, based on the DNN models. Specifically, y-axis presents the marginal impact on the probability of default, while x-axis present the number of months elapsed since loan origination.

Figure 18: DNN – PDP – Turnover Over Loan

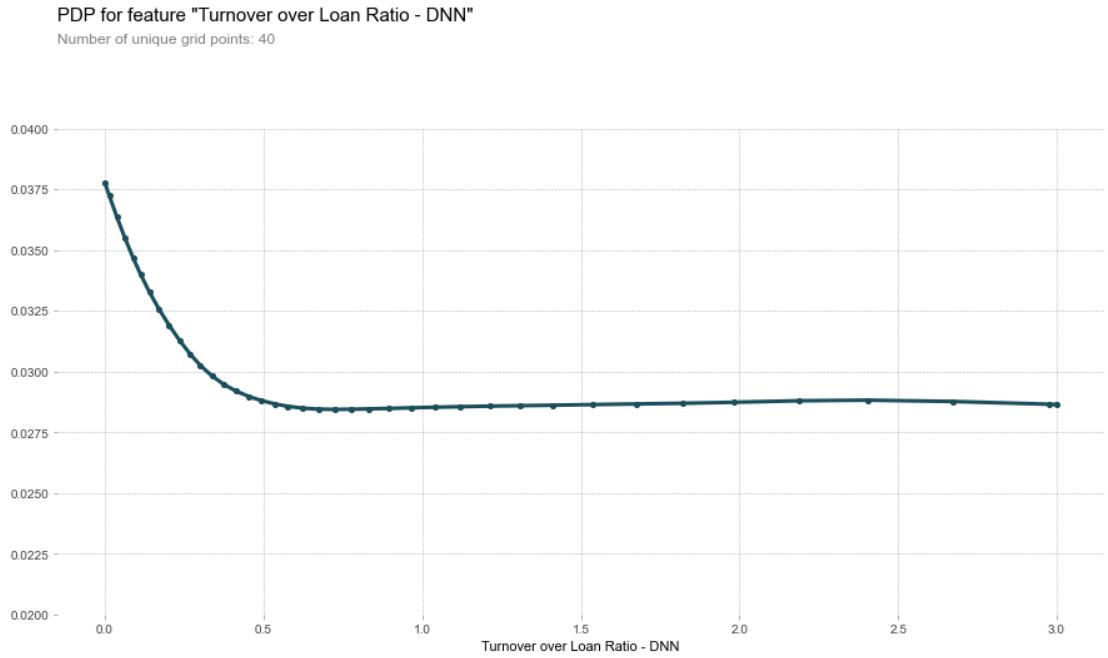


Figure 18 shows how the impact of the "Turnover over Loan Ratio" affects the probability of default, based on the DNN models. Specifically, y-axis presents the marginal impact on the probability of default, while x-axis present the values of the "Turnover over Loan Ratio".

Figure 19: DNN – PDP – Global NPE

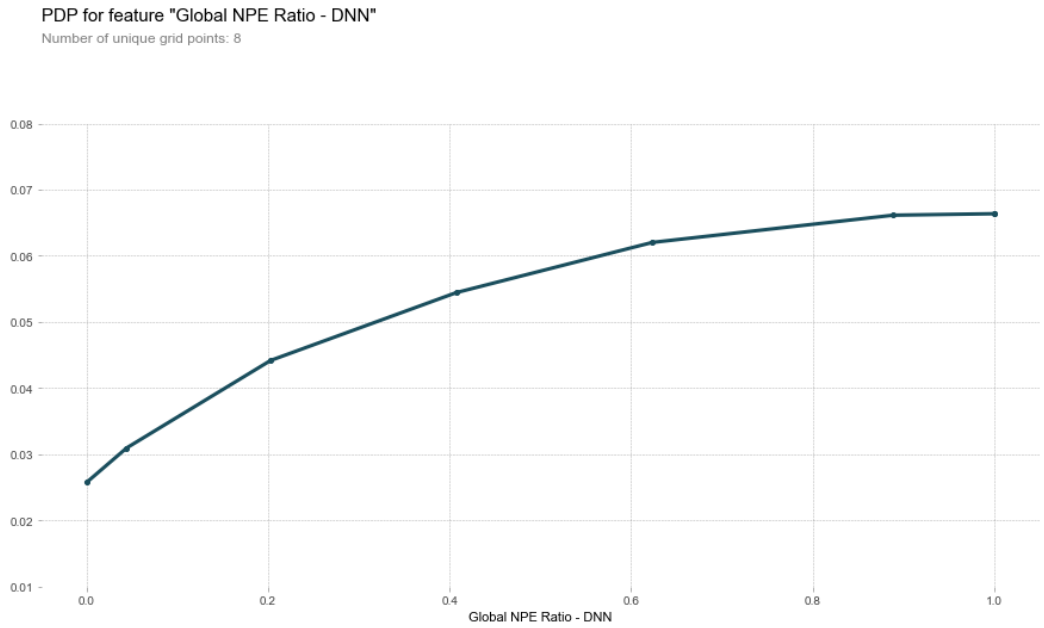


Figure 18 presents how the impact of the "Global NPE ratio" affects the probability of default, based on the DNN models. Specifically, y-axis presents the marginal impact on the probability of default, while x-axis present the values of the "Global NPE ratio".

Tables

Table 1: Model Discrimination (per sample)

Test Sample					
Model Discrimination Metric	Model				
	RF	XGB	DNN	SNN	LR
KS	0.620	0.673	**0.686**	0.618	0.265
AUROC:	0.885	0.914	**0.918**	0.875	0.652

Train Sample					
Model Discrimination Metric	Model				
	RF	XGB	DNN	SNN	LR
KS	0.582	0.680	**0.781**	0.576	0.150
AUROC:	0.870	0.918	**0.961**	0.865	0.600

Table 1 illustrates the model discrimination metrics along the 5 models developed, both for the test (out-of-time) and for the train samples.

Table 2: Model Performance (per sample)

Test Sample					
Model Performance Metric	Model				
	RF	XGB	DNN	SNN	LR
Accuracy:	0.926	0.946	**0.947**	0.941	0.939
Precision:	0.119	0.160	**0.168**	0.123	0.025
Recall:	0.560	0.605	**0.611**	0.425	0.074
ROC-AUC:	0.746	0.778	**0.781**	0.687	0.514

Train Sample					
Model Performance Metric	Model				
	RF	XGB	DNN	SNN	LR
Accuracy:	0.922	0.925	**0.926**	0.908	0.801
Precision:	0.184	0.225	**0.227**	0.163	0.044
Recall:	0.511	0.676	**0.852**	0.55	0.292
ROC-AUC:	0.723	0.804	**0.884**	0.734	0.554

Table 2 presents the model performance metrics, both for the test (out-of-time) and for the train samples, along the 5 models developed.

Table 3: Model Accuracy (per sample)

Test Sample					
Model Predictive Accuracy Metric	Model				
	RF	XGB	DNN	SNN	LR
Mean Absolute Error	3,58%	3,27%	**2,43%**	3,03%	3,13%
Mean Squared Error	1,47%	1,35%	**1,26%**	1,51%	1,65%
Root Mean Squared Error	12,12%	11,62%	**11,22%**	12,30%	12,86%

Train Sample					
Model Performance Metric	Model				
	RF	XGB	DNN	SNN	LR
Mean Absolute Error	5,00%	4,85%	**3,66%**	5,07%	7,82%
Mean Squared Error	2,44%	2,14%	**1,67%**	2,47%	3,44%
Root Mean Squared Error	15,61%	14,62%	**12,91%**	15,72%	18,54%

Table 3 shows the model accuracy metrics, when assessing their predicted default probabilities, for the test (out-of-time) and for the train samples along the 5 models developed.

Table 4: Model Calibration MSE

Mean Squared Error					
Sample	Model				
	RF	XGB	DNN	SNN	LR
Test	14,73%	0,72%	**0,34%**	20,20%	31,10%
Train	1,51%	3,64%	**0,71%**	10,75%	21,52%

Table 4 presents the MSE when assessing the calibration to a rating system for each one of the 5 models developed, both for the test (out-of-time) and for the train samples.

Appendix 1 - Variables Employed

The variables used in this study are presented below.

Set #1	Set #2	Set #5
GDPQ	Global_NPE_Ratio	Covid_Months
HPI	GVolume_Change	Govt_Support
UN	Turnover_Over_Loan	
	Bucket_1	
	Bucket_2	
	Bucket_3	

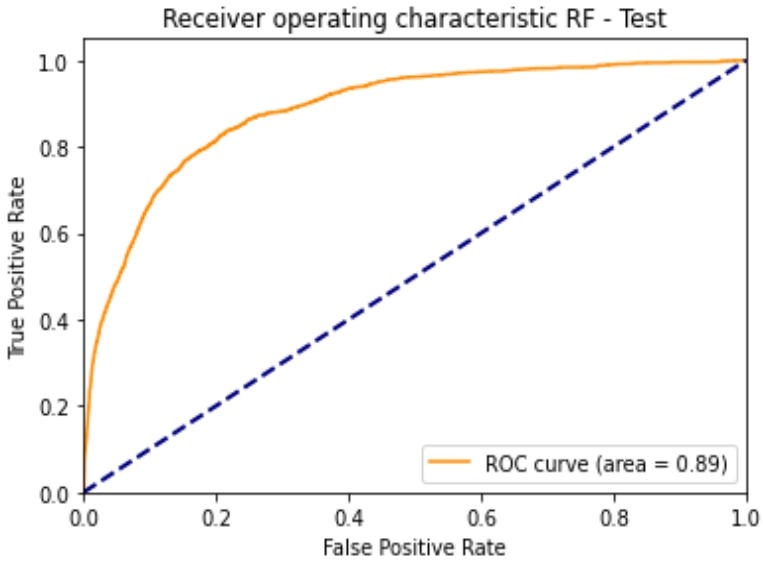
Set #3		
Stage_2_0	Legal_Status_3	Sector_Education
Stage_2_1	Legal_Status_4	Sector_Electricity,_
Stage_2_2	Legal_Status_5	Sector_Financial_&_I
Stage_2_3	Legal_Status_6	Sector_Human_Health_
Stage_2_4	Forbearance_Status_0	Sector_Information_&
Stage_2_5	Forbearance_Status_1	Sector_Manufacturing
Stage_2_6	Forbearance_Status_2	Sector_Mining_&_Quar
Stage_3_1	Forbearance_Status_3	Sector_Other_Service
Stage_3_2	Forbearance_Status_4	Sector_Professional,
Stage_3_3	Forbearance_Status_5	Sector_Public_Admini
Stage_3_4	Forbearance_Status_6	Sector_Real_Estate_A
Stage_3_5	Sector_Accommodation	Sector_Transportatio
Stage_3_6	Sector_Administrativ	Sector_Water_Supply
Legal_Status_0	Sector_Agriculture,_	Sector_Wholesale_&_R
Legal_Status_1	Sector_Arts,_Enterta	
Legal_Status_2	Sector_Construction	

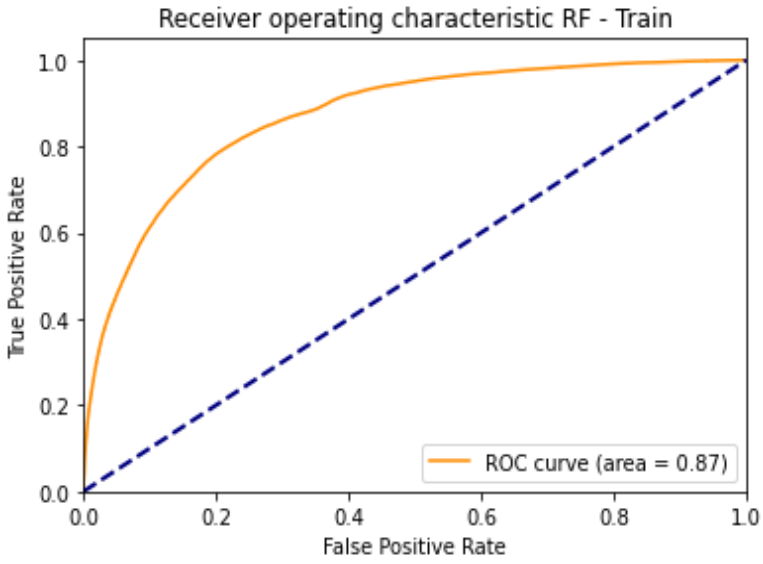
Set #5		
Bank_01	Credit_card_debt	Trade_receivables
Bank_02	Credit_lines_other_t	Working_capital_faci
Bank_03	Debt_financing	Zero-coupon_
Bank_04	Deposits_other_than_	
Bank_05	Exports	
Bank_06	Finance_leases	
Bank_07	Imports	
Bank_08	Loans_other_than_ove	
Bank_09	Monthly	
Bank_10	Other_than_monthly,_	
Bank_11	Overdrafts	
Duration	Purposes_other_than_	
INTEREST_RATE	Quarterly	
Annual	Residential_real_est	
Commercial_real_esta	Revolving_credit_oth	
Construction_investm	Semi-annually	

Appendix 2 - ROC Curves

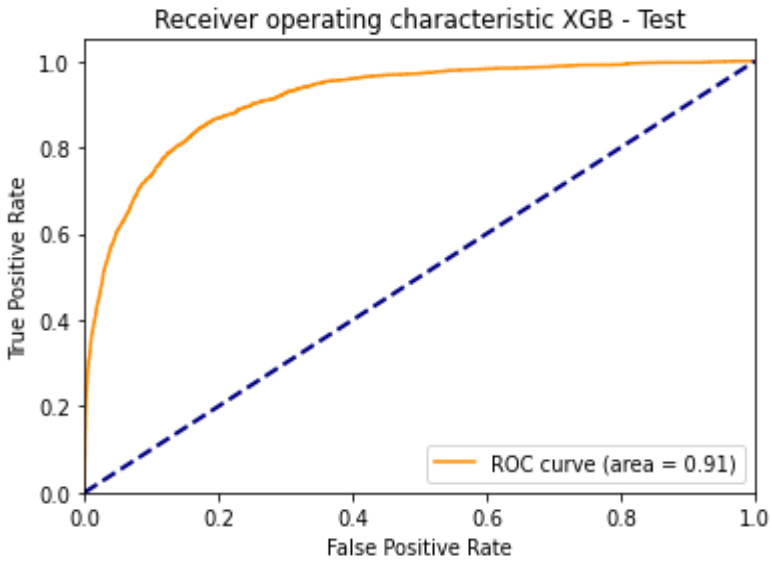
The performance of each model developed, both for the Test and for the Train sample, is shown below based on the ROC curve. Specifically, y-axis presents the True Positive Rate (i.e., predict correctly a defaulted obligor), while x-axis present the False Positive Rate (i.e., predict incorrectly a non-defaulted obligor as defaulted obligor). The more to the left and to the upper the ROC curve, the better the performance of the model.

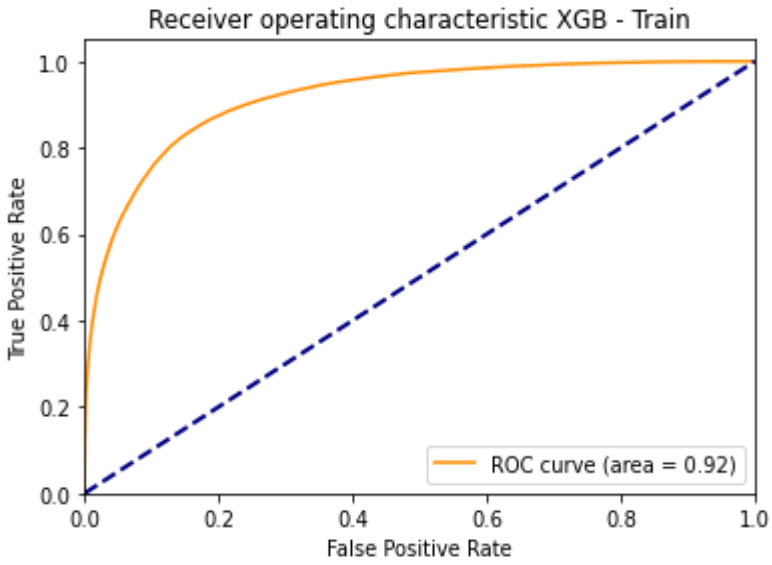
Random Forest



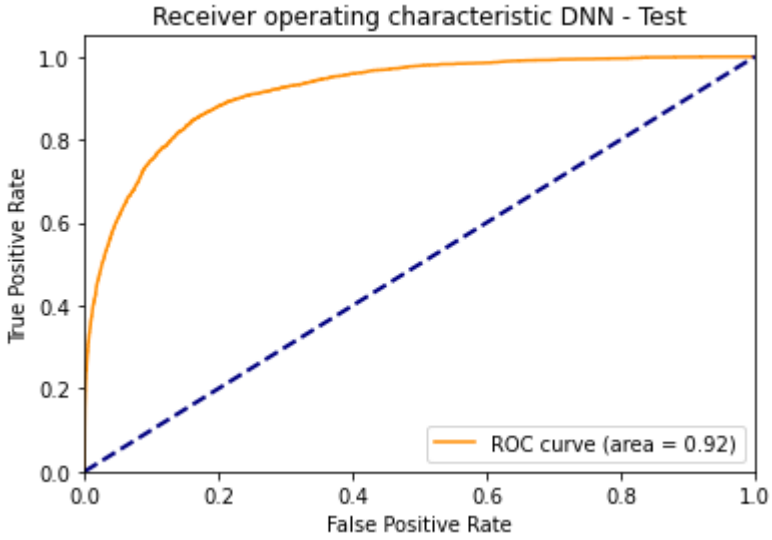


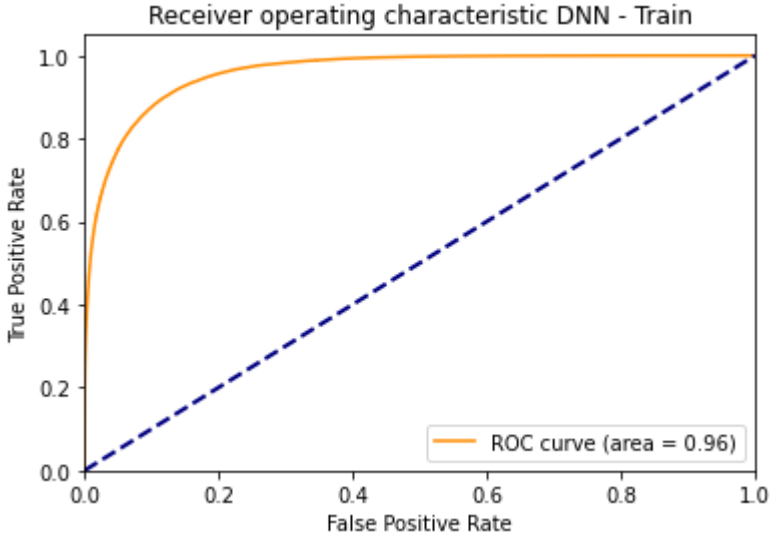
XGBoost


Receiver operating characteristic XGB - Test
True Positive Rate
False Positive Rate
ROC curve (area = 0.91)
0.0
0.2
0.4
0.6
0.8
1.0

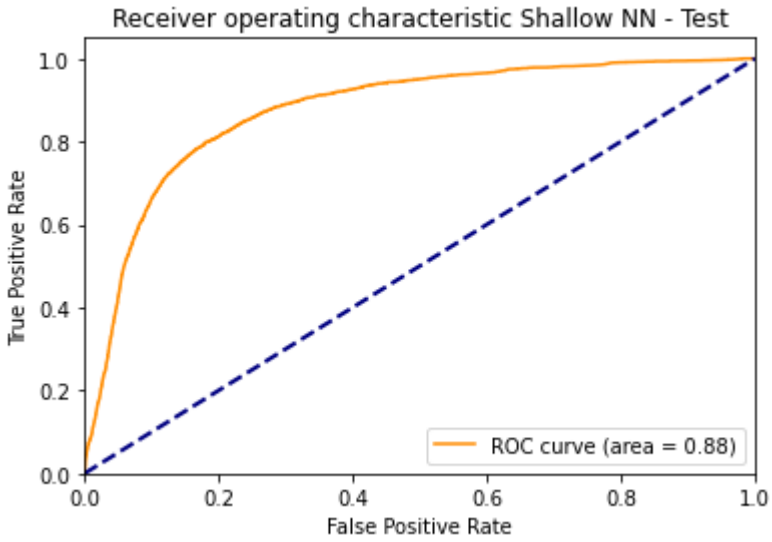

Receiver operating characteristic XGB - Train
True Positive Rate
False Positive Rate
ROC curve (area = 0.92)
0.0
0.2
0.4
0.6
0.8
1.0

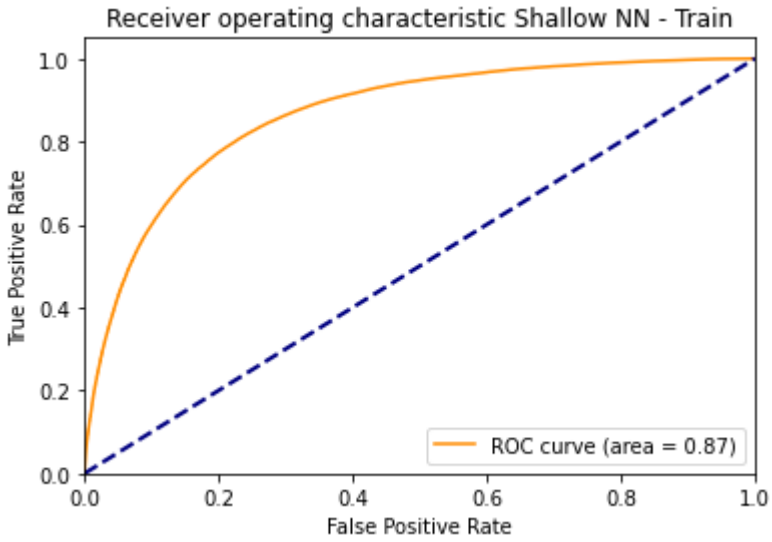
Deep Neural Network



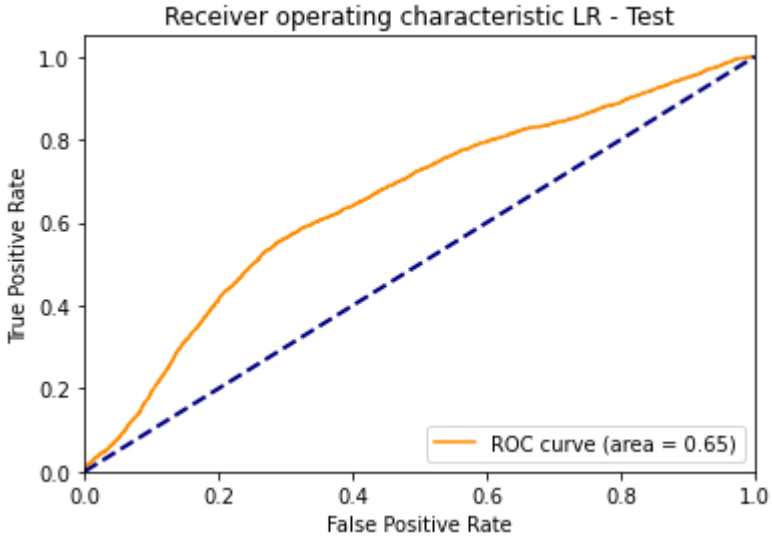


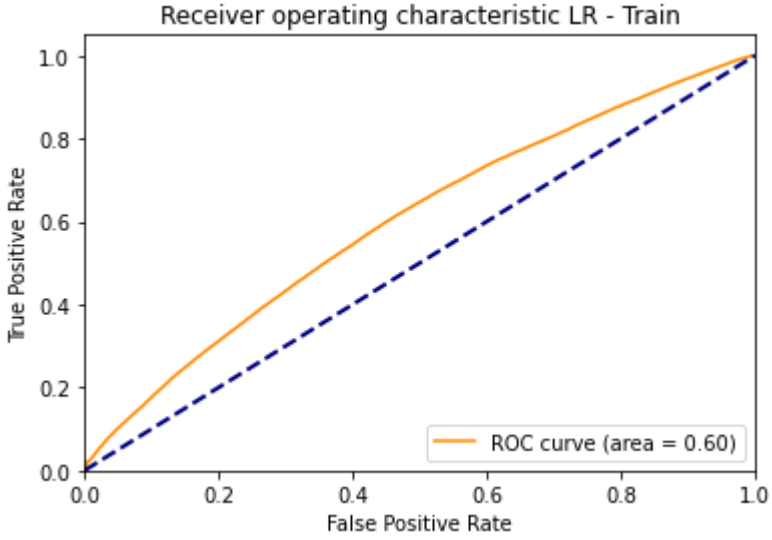
Shallow Neural Network





Logistic Regression

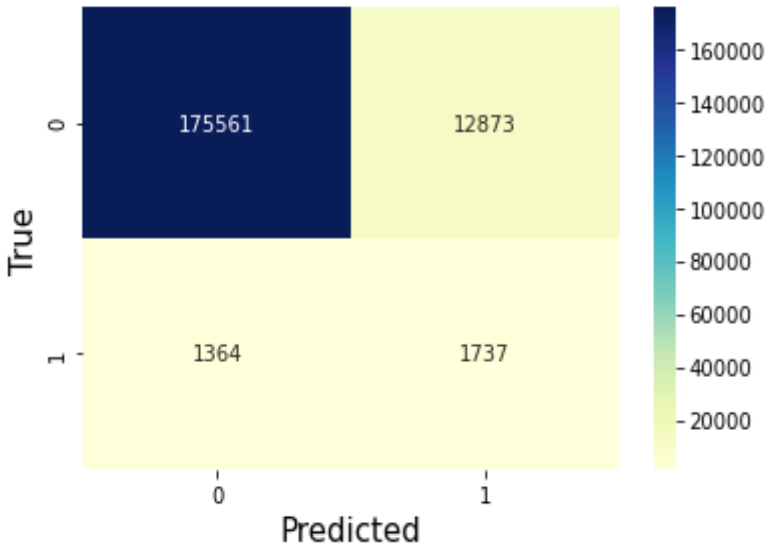




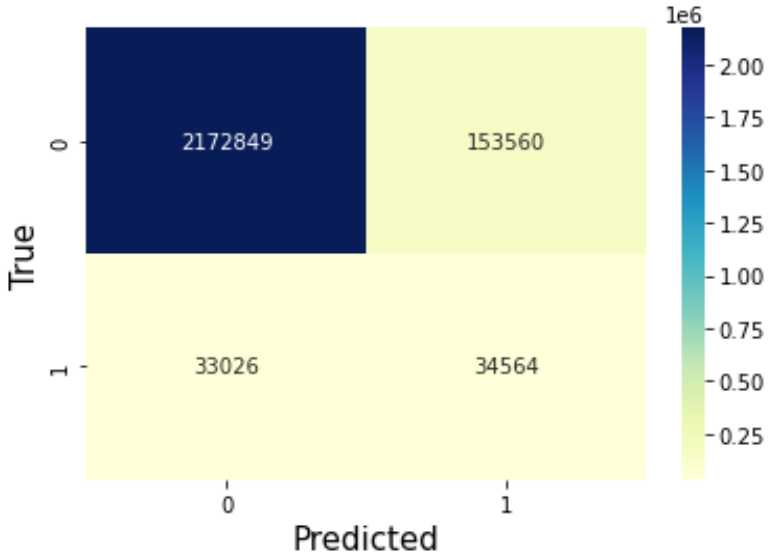
Appendix 3 - Classification tables

Confusion matrixes for each model developed, both for the Test and for the Train samples, are shown below. Specifically, y-axis presents the True cases (i.e. 0 indicates non-defaulted, 1 indicates defaulted), while x-axis present the respective predictions (i.e. 0 indicates non-defaulted, 1 indicates defaulted). In the diagonal of the matrix, we can identify all the cases correctly predicted. That is, in the [True=0, Predicted=0] case obligors classified correctly as non-defaulted are shown, and in the [True=1, Predicted=1] case obligors classified correctly as defaulted are shown. On the other hand, in the non-diagonal elements the wrong predictions are shown. That is, in the [True=0, Predicted=1] case non-defaulted obligors are incorrectly classified as defaulted, and in the [True=1, Predicted=0] case defaulted obligors are incorrectly classified as non-defaulted.

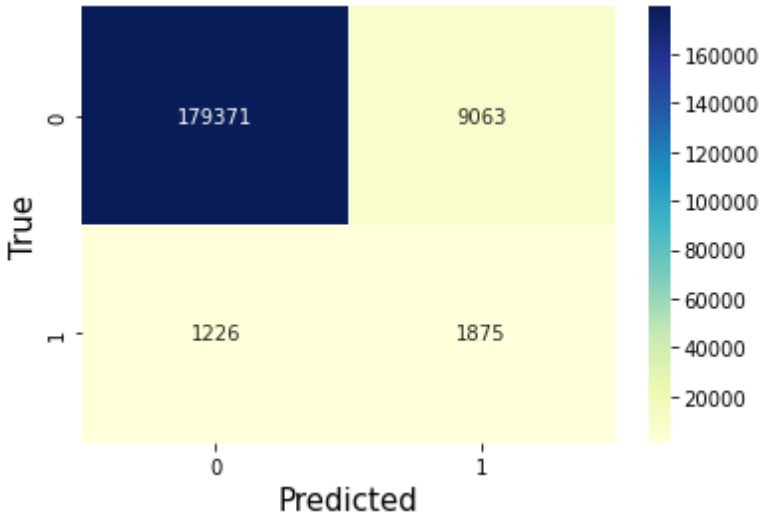
Random, Forest – Test Sample



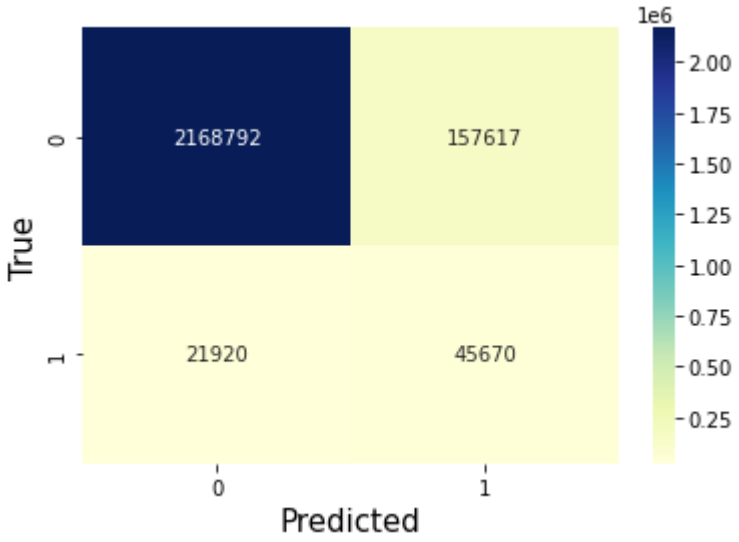
Random Forest – Train Sample



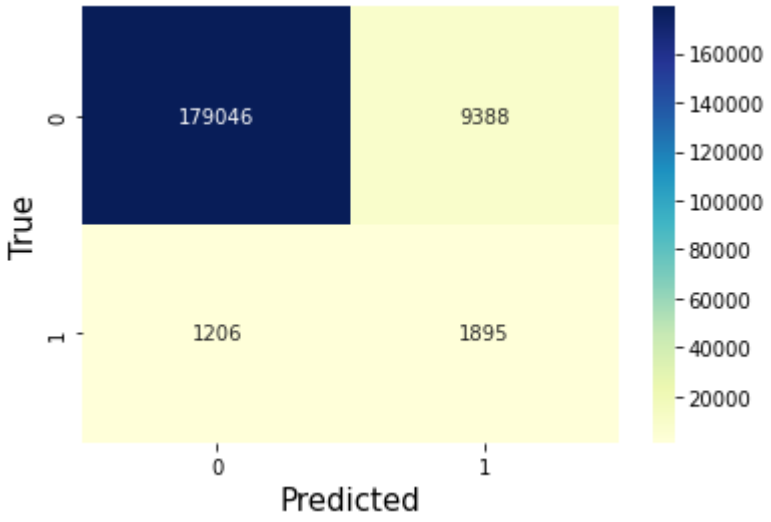
XGBoost – Test Sample



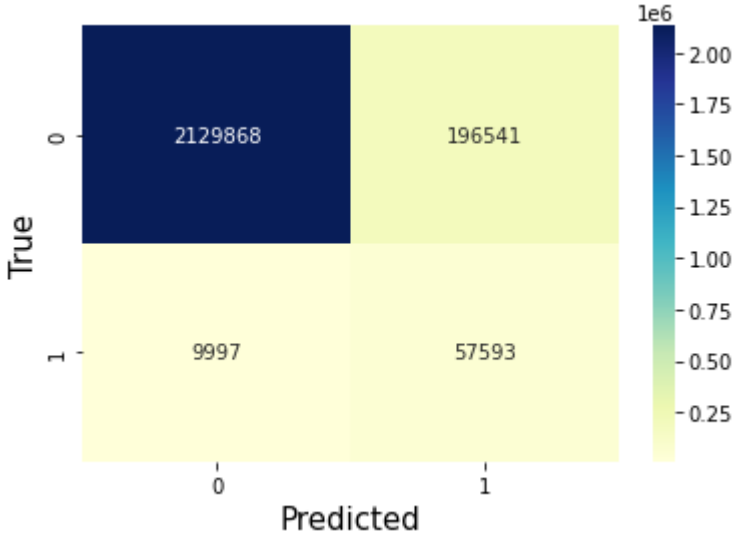
XGBoost – Train Sample



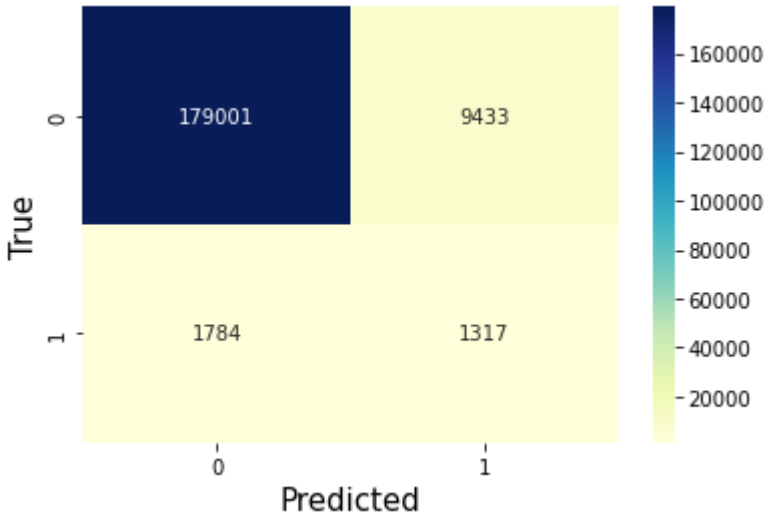
Deep Neural Network – Test Sample



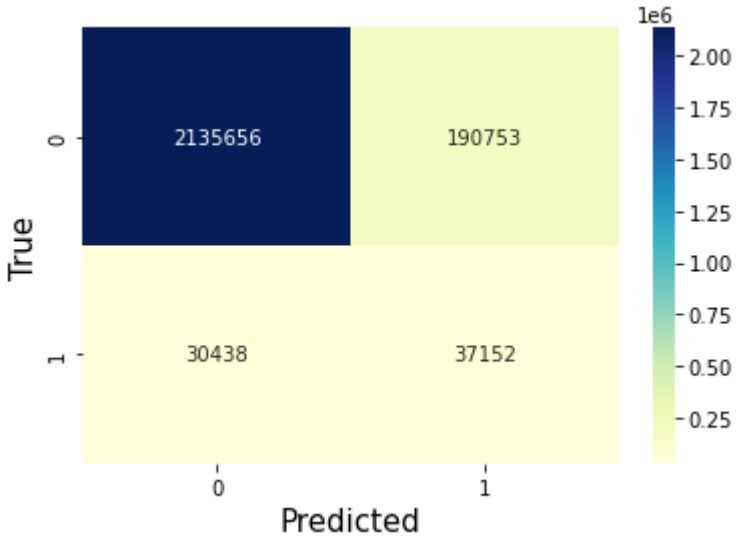
Deep Neural Network – Train Sample



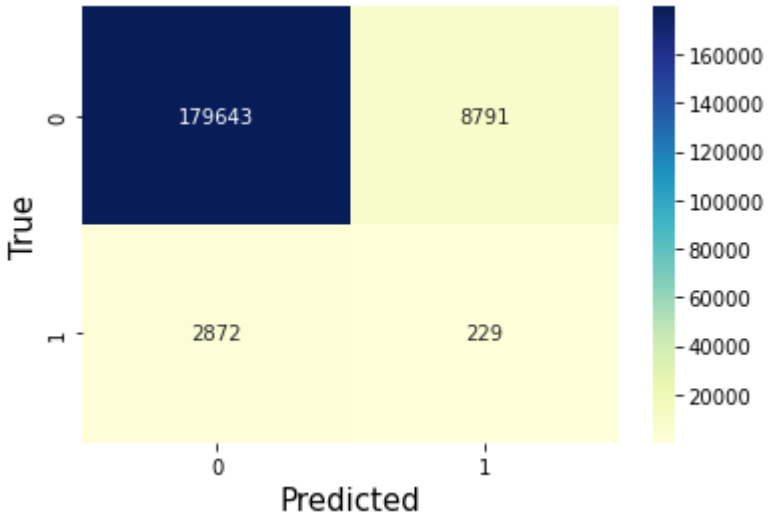
Shallow Neural Network – Test Sample



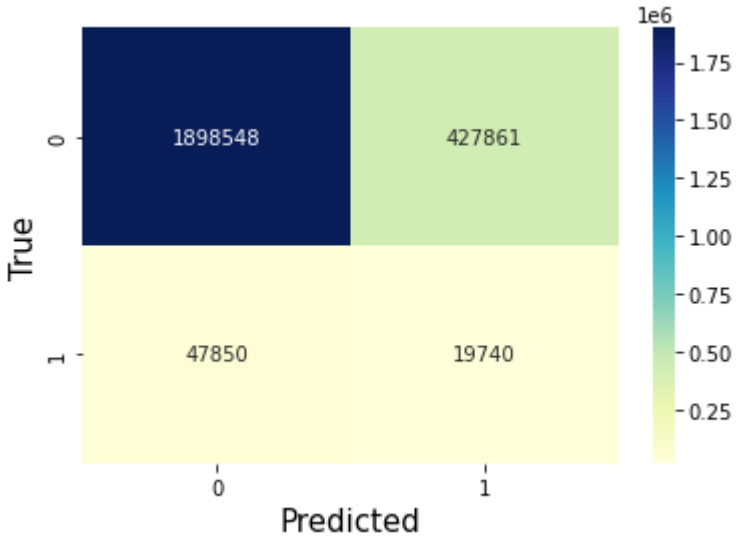
Shallow Neural Network – Train Sample



Logistic Regression – Test Sample



Logistic Regression – Train Sample



Appendix 4 - Logistic Regression Model

The variables used in this study to develop the logistic regression model are presented below.

LR – Selected Model Variables
Stage_2_0
Forbearance_Status_0
Stage_3_1
Stage_3_2
Legal_Status_4
Sector Wholesale & Retail Trade
Sector_Manufacturing
Sector_ Accommodation & Food Service
Sector_ Public Administration and Defence
Global_NPE_Ratio
GVolume_Change
Turnover_Over_Loan
Duration
Covid_Months
Govt_Support
Semi-annually
Bucket_3
Bucket_2
Debt_financing
Bank_05
Bank_06

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. (1968). 'Financial ratios, discriminant analysis and the prediction of corporate bankruptcy', The journal of finance 23.4: 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', Review of Finance, 8 (4), 537–569.

Chen, Tianqi, & Carlos Guestrin (2016). 'Xgboost: A scalable tree boosting system', Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

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.

Feng Shen, Xingchao Zha, Gang Kouc, Fawaz E. Alsaadi (2021). 'A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique', Applied Soft Computing Volume 98, 106852.

Fryer, D., Strümke, I., & Nguyen, H. (2021). 'Shapley values for feature selection: The good, the bad, and the axioms', IEEE Access, 9, 144352-144360.

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

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.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E.. (2015). 'Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation', Journal of Computational and Graphical Statistics, 24(1): 44-65.

Hamori, S.; Kawai, M.; Kume, T.; Murakami, Y.; Watanabe, C. (2018). 'Ensemble Learning or Deep Learning? Application to Default Risk Analysis.', J. Risk Financial Manag. 11, 12. https://doi.org/10.3390/jrfm11010012

Heaton, J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2018). 'Deep learning': The MIT Press, (2016), 800 pp, ISBN: 0262035618. Genetic Programming and Evolvable Machines, 19(1-2), 305-307.

Hossin, M., & Sulaiman, M. N. (2015). 'A review on evaluation metrics for data classification evaluations', International journal of data mining & knowledge management process, 5(2), 1.

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.

Iñaki Aldasoro & Ingo Fender & Bryan Hardy & Nikola Tarashev (2020). "Effects of Covid-19 on the banking sector: the market's assessment," BIS Bulletins 12, Bank for International Settlements.

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.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). 'Deep learning', Nature, 521(7553), 436-444.

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.

Patrick Augustin, Valeri Sokolovski, Marti G. Subrahmanyam, Davide Tomio (2022). 'In sickness and in debt: The COVID-19 impact on sovereign credit risk', Journal of Financial Economics, Volume 143, Issue 3, pp 1251-1274.

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.

Petropoulos A., Siakoulis V., Stavroulakis E., Klamargias A., (2019). 'A robust machine learning approach for credit risk analysis of large loan-level datasets using deep

learning and extreme gradient boosting', IFC Bulletins chapters, Volume 50, Bank for International Settlements.

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

Ribeiro, M., Singh, S., & Guestrin, C. (2016). 'Local interpretable model-agnostic explanations (LIME): an introduction', O'Reilly Media.

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. (2016). 'Modeling coverage for neural machine translation', Proc. ACL.

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.

BANK OF GREECE WORKING PAPERS

304. Kotidis, A., D. Malliaropulos and E. Papaioannou, "Public and private liquidity during crises times: evidence from emergency liquidity assistance (ELA) to Greek banks", September 2022.

305. Chrysanthakopoulos, C. and A. Tagkalakis, "The effects of fiscal institutions on fiscal adjustments", October 2022.

306. Mavrogiannis, C. and A. Tagkalakis, "The short term effects of structural reforms and institutional improvements in OECD economies", October 2022.

307. Tavlas, S. G., "Milton Friedman and the road to monetarism: a review essay", November 2022.

308. Georgantas, G., Kasselaki, M. and Tagkalakis A., "The short-run effects of fiscal adjustment in OECD countries", November 2022

309. Hall G. S., G. S. Tavlas and Y. Wang, "Drivers and spillover effects of inflation: the United States, the Euro Area, and the United Kingdom", December 2022.

310. Kyrkopoulou, E., A. Louka and K. Fabbe, "Money under the mattress: economic crisis and crime", December 2022.

311. Kyrtsou, C., "Mapping inflation dynamics", January 2023.

312. Dixon, Huw, T. Kosma and P. Petroulas, "Endogenous frequencies and large shocks: price setting in Greece during the crisis", January 2023.

313. Andreou P.C, S. Anyfantaki and A. Atkinson, "Financial literacy for financial resilience: evidence from Cyprus during the pandemic period", February 2023.

314. Hall S. G, G.S. Tavlas and Y. Wang, "Forecasting inflation: the use of dynamic factor analysis and nonlinear combinations", February 2023.