



**UNIVERSITA' DEGLI STUDI DI PADOVA**

**DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI  
"M. FANNO"**

**CORSO DI LAUREA MAGISTRALE IN ECONOMICS AND FINANCE**

**TESI DI LAUREA**

**"PREDICTIVE MODELLING OF FINANCIAL CRISES: MACHINE  
LEARNING ALGORITHMS IN A RECURSIVE REAL-TIME FRAMEWORK"**

**RELATORE:**

**CH.MO PROF. FORNI LORENZO**

**LAUREANDO: MORO FILIPPO**

**MATRICOLA N. 2023368**

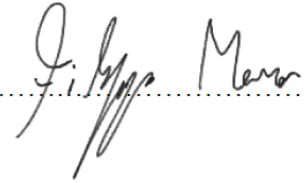
**ANNO ACCADEMICO 2023 – 2024**



Dichiaro di aver preso visione del “Regolamento antiplagio” approvato dal Consiglio del Dipartimento di Scienze Economiche e Aziendali e, consapevole delle conseguenze derivanti da dichiarazioni mendaci, dichiaro che il presente lavoro non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Dichiaro inoltre che tutte le fonti utilizzate per la realizzazione del presente lavoro, inclusi i materiali digitali, sono state correttamente citate nel corpo del testo e nella sezione ‘Riferimenti bibliografici’.

*I hereby declare that I have read and understood the “Anti-plagiarism rules and regulations” approved by the Council of the Department of Economics and Management and I am aware of the consequences of making false statements. I declare that this piece of work has not been previously submitted – either fully or partially – for fulfilling the requirements of an academic degree, whether in Italy or abroad. Furthermore, I declare that the references used for this work – including the digital materials – have been appropriately cited and acknowledged in the text and in the section ‘References’.*

Firma (signature) .....





# INDEX

INTRODUCTION .....	1
FINANCIAL CRISES .....	3
IMPLEMENTED MODELS .....	8
PERFORMANCE EVALUATION METRICS .....	26
LITERATURE SURVEY .....	33
PREDICTORS.....	46
BASELINE EXERCISE.....	55
VARIATIONS.....	70
LAGS ANALYSIS.....	79
2023 - 2025 PREDICTIONS.....	88
CONCLUSIONS .....	91
BIBLIOGRAPHY .....	93
APPENDIX .....	97



# INTRODUCTION

In recent decades, the global financial landscape has witnessed a surge in banking crises, causing substantial economic and social damage across developed and developing economies alike. The 1990s and early 2000s marked a period of heightened financial instability, with crises ranging from the collapse of banking systems in emerging markets to the profound disruptions experienced by developed western economies. The recurrence of banking crises is not a new phenomenon, with historical precedents such as the Great Depression serving as reminders of the potential catastrophic consequences. The post-World War II era saw a period of tight regulations on banks in response to past crises, temporarily mitigating the issue. However, a general trend toward financial liberalisation began in the 1970s, leading to a resurgence of banking problems. From the Mexican and Argentine crises in the 1990s and the Asian financial turmoil in 1997-1998, to the Great Financial Crisis (GFC) of 2008, these events have emerged across different geographical boundaries, posing significant challenges to policymakers.

The costs associated with resolving these crises have been enormous, often reaching double-digit percentages of GDP in affected countries (Laeven and Valencia, 2020). The global financial crisis of 2008 demonstrated the rapid and far-reaching consequences of a banking calamity. The interconnectedness of financial systems across continents led to prolonged economic and financial losses, with adverse effects persisting for years. The financial institutions, once considered guardians of depositors' savings, found themselves at the centre of a crisis that shook the foundations of economies worldwide. The aftermath of such crises includes severe unemployment, increased poverty, weakened exports, and pro-cyclical spending by governments, exacerbating the economic challenges faced by affected nations.

These crises, often triggered by a variety of factors, have highlighted the strong need for effective early warning systems and predictive models capable of identifying the conditions leading to financial turmoil. Identifying robust predictors of financial crises is challenging due to limited observed crises, the delayed signalling of crisis indicators, missing data, and the need for transparent models that facilitate timely intervention by macroprudential authorities. The development of such systems is crucial not only for mitigating the economic and social costs of crises but also for guiding policymakers in crafting effective pre-emptive measures.

This thesis has the goal to present to the reader the past evolution and the actual state of the art of the literature on the different models available to forecasters, their basic functioning, and the results obtained in terms of model effectiveness and variables' significance. Then, a baseline

analysis conducted by the author, followed by multiple variants of the same analysis, is evaluated, and compared to the literature's previous results. The author's models are trained targeting the pre-crisis periods, defined as the three years preceding a crisis event, and then tested recursively year-by-year on the testing data subset, trying to classify whether any observation belongs to the pre-crisis or tranquil-period category. A framework including multiple predictors and lags is used, with the objective of identifying the best algorithms and, when possible, the impact of each individual variable on the predicted outcome. An important condition applied to this objective is to evaluate each model in a framework as close as possible to a real-world implementation, so to obtain an honest and unbiased look at real performance. This implies a strict separation between the training dataset and the testing dataset during both training and testing.

This dissertation is structured as follow: first, a definition of what constitutes a financial crisis is given, retrieving the criteria from the same authors who built the databases used in the analysis. After this, the most popular machine learning models in the literature are explained in their basic functioning, and an explanation of the different evaluation parameters is given. Following this, a review of the literature on the argument shows the methodologies adopted in different economic papers, achieving different, and sometimes opposing, interpretations. Then the author of this Thesis lists the twelve specific predictors implemented in his analysis, nine domestic and three international, explaining their involvement in the economic mechanisms leading to financial distress. These predictors are then included in a baseline exercise, followed by multiple experiment both inspired by the literature and designed by the author. The predictions are then analysed and assessed, with a subsequent examination of the accuracy of forecasts related to specific events. The results are finally used to observe what the best models forecast for the future in a small, selected panel of countries. This Thesis takes inspiration mostly from the work of Beutel et al. (2018), whose research will be presented in the following chapters.



# FINANCIAL CRISES

The objective of this chapter is to establish a precise definition of financial crisis. Such a definition is crucial for conducting an accurate comparison of this dissertation with the rest of the literature on the subject. Furthermore, extending this research to future real-world scenarios requires to update crisis databases with the most recent events. It is imperative that these new events align cohesively with past events' definition for consistency and relevance, so to reduce arbitrariness on what constitutes a financial crisis. This Thesis implements three main banking crises databases from three different authors. What follows is a report of crisis definition from each author and an overview of the databases' content.

## Definitions

The backbone of the crisis's dataset used in this work is retrieved from **Laeven and Valencia (2020)**, which is an updated version of previous datasets of the same authors (Laeven and Valencia, 2008, 2013, 2018). Along these many versions, the definition has remained the same. The authors state that a banking crisis is defined as an event that meets two baseline conditions:

- Significant signs of financial distress in the banking system (as indicated by significant bank runs, losses in the banking system, and/or bank liquidations).
- Significant banking policy intervention measures in response to significant losses in the banking system.

The first year when both criteria are met is considered as the year in which the crisis became systemic. In case of severe losses or liquidations, the authors treat the first criterion as sufficient if any of these two conditions are met:

- Non-performing loans (NPL) above 20% of total loans or at least 20% bank closures of banking system assets.
- Fiscal restructuring costs of the banking sector exceeds 5% of GDP.

The second baseline condition, policy interventions, is considered significant by the authors if at least three of the following six measures had been adopted:

- Deposit freezes and/or bank holidays: this aspect gauges government-imposed restrictions on deposit withdrawals or bank holidays.

- Significant bank nationalisations: involve government takeovers of systemically crucial financial institutions, encompassing cases where the government acquires a majority stake in their capital.
- Bank restructuring fiscal costs: refers to gross fiscal outlays dedicated to financial sector restructuring, notably recapitalisation costs. Considered significant if exceeding 3% of GDP, excluding direct treasury liquidity assistance.
- Extensive liquidity support: measured as central bank claims on other depository institutions and direct liquidity support from the Treasury. An extensive ratio exceeding 5%, more than doubling relative to pre-crisis levels, signifies significant liquidity support.
- Significant guarantees put in place: highlights substantial government guarantees on bank liabilities, encompassing full protection of liabilities or extensions of guarantees to non-deposit liabilities of banks. It excludes actions that solely elevate deposit insurance coverage.
- Significant asset purchases: denotes acquisitions of financial institutions' assets by the central bank, treasury, or government entities (like asset management companies). Significant asset purchases are those exceeding 5% of GDP.

These clear-cut thresholds allow for an objective distinction of events, which is lacking in the similar works of other authors who implemented criteria based on a narrative approach. Laeven and Valencia's database is constructed on the observations from 206 countries over the period 1970-2017. In total, the database covers 151 systemic banking crises meeting the already mentioned criteria. A great advantage of Laeven and Valencia's database is that they also report the duration and end period of each crisis. End dates are defined as the year before both real GDP growth and real credit growth are positive for at least two consecutive years. However, the authors truncate the maximum duration of a crisis at 5 years, because their metric may start picking up the impact of other shocks. This database is further extended by **Nguyen et al. (2022)** who used the same criteria as Laeven and Valencia to analyse banking crises of the years 2018 and 2019 but found that no new events have arisen.

The second database this dissertation draws from is that by **Reinhart and Rogoff (2008)**. To define a banking crisis, they do not rely on quantitative criteria, opting for an approach based on narrative. Stating the authors, banking crises may be of two types:

- Bank runs that lead to the closure, merging, or takeover by the public sector of one or more financial institutions.

- If there are no runs, the closure, merging, takeover, or large-scale government assistance of an important financial institution (or group of institutions) that marks the start of a string of similar outcomes for other financial institutions.

The authors also argue that they are aware that such approach could lead to the identification of a crisis date too early (because the worst of a crisis may come later) or too late (financial problems usually begin well before a bank is finally closed or merged). This database analyses the emergence of banking crises in history up to the Great Financial Crisis, for both advanced and emerging countries, also indicating the duration.

One last addition to the complete dataset of this work is drawn from **Jordà et al. (2017)**. This is well-known research which encompass the main macroeconomic variables for 17 advanced economies since 1870 to 2013, as well as a dummy variable indicating the occurrence of a banking crisis in a particular year. Alas, they do not give a clear-cut quantitative definition of such events, nor they provide information about the duration of distress.

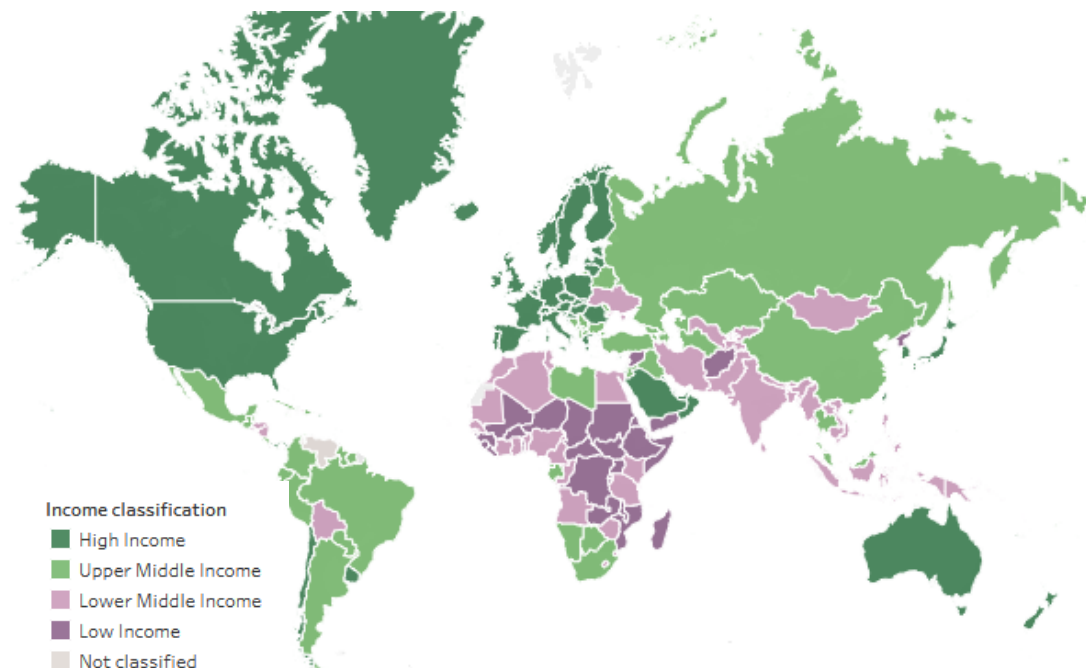
## **Database**

The works of these three research groups are implemented in the Thesis by aggregating the data in a single database. The author started from the database of Laeven and Valencia, being it the one including the most countries, and having a quantitative and replicable criterion for defining crises. This implies that the final database will cover the period 1970-2019. Then, the events drawn from the other two databases after the year 1970 are added. Of course, most of the events between the three databases are corresponding, but some differences can be point out. For example, a Brazilian crisis in 1985-1986 caused by raging inflation is showed only by Reinhart and Rogoff, the 1995 Argentinian crisis has a duration of one year for Laeven and Valencia, and two years for Reinhart and Rogoff, while Jordà et al. miss all observations from developing countries. Jordà et al. also suffer from the missing end year of each crisis, but is partially bypassed by the author of this Thesis by assuming a duration of three years for each event uniquely pointed out by Jordà et al., which is the average duration recorded by Laeven and Valencia. The final database is the sum of all observations, so to achieve the maximum number of crisis events adopting the widest definition as possible. Including as many events as possible is important to the scope of this work, since banking crises are relative rare events, and their low number in the database could affect later analyses and the correct functioning of machine learning models, or the interpretability of their results. A further step is taken by separating the events recorded in advanced and developing countries. This will be of some importance later,

when analysing the behaviour of predictors before, during and after a crisis, given the fundamentally different structures of these two categories of economies, their different weaknesses, and reactions to such events. The author of this paper separates these two categories by referencing the distinction elaborated by the World Bank, which uses as separators the average income in USD, based on 2022 gross national income (GNI) per capita, calculated using the World Bank Atlas method. This is done so to implement an objective criterion based exclusively on economic factors. The four classes are:

- Low income, \$1,135 or less
- Lower middle income, \$1,136 to \$4,465
- Upper middle income, \$4,466 to \$13,845
- High income, \$13,846 or more

For the purposes of this dissertation, only countries defined as high income will be accounted as advanced economies.



*Figure 1.1: Classification by income. Source: World Bank*

The overall database includes 204 separated episodes of banking crises in 204 countries over 50 years, distributed as shown in Figure 1.2.

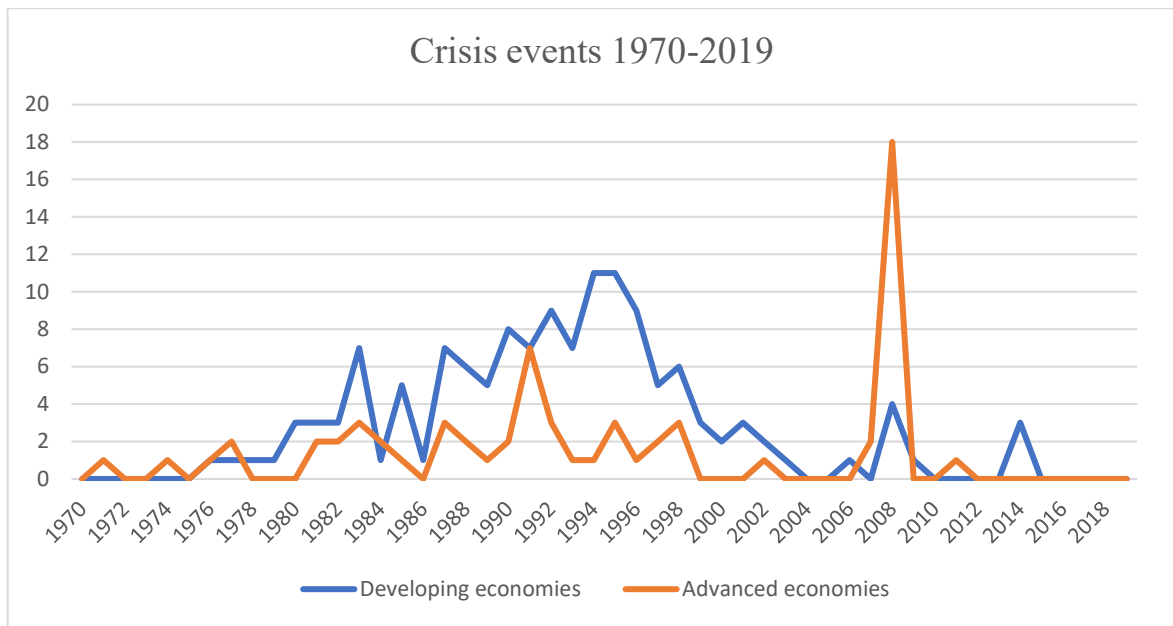


Figure 1.2: Number of crisis events. Author's own elaboration

Of all observations, 700 are recorded during a state of crisis by any of the three mentioned authors as explained above. Figure 1.3 shows the share of global GDP produced by countries which are facing a banking crisis year-by-year.

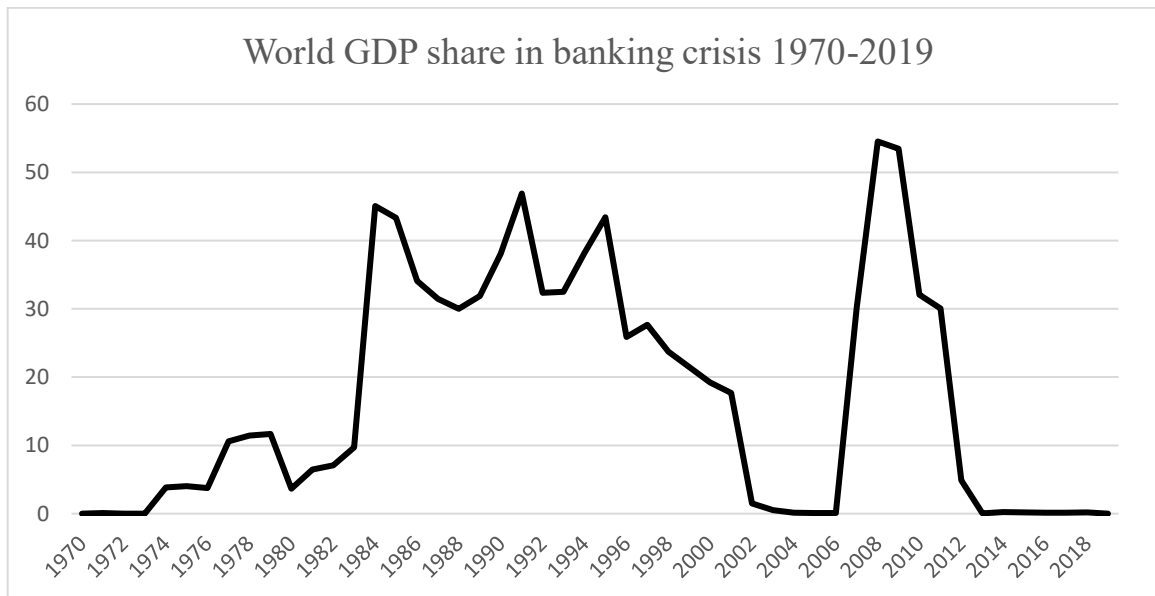


Figure 1.3: World GDP share of countries facing a banking crisis. Author's own elaboration

# IMPLEMENTED MODELS

Before getting started with the review of the literature about the argument of this dissertation, the author proposes a summary of the forecasting methods and algorithms implemented later in the Thesis, which for the great part coincide with the main models generally used in this research field. Machine learning (ML) algorithms are a subset of artificial intelligence (AI) that allow systems to learn and improve from experience, finding hidden insights and complex patterns without being explicitly programmed (Janiesch, Zschech and Heinrich, 2021). In other words, these algorithms enable pattern recognition, as well as predictions and problem-solving by learning from data. The history of machine learning spans several decades and has evolved through various stages of development, driven by advances in mathematics, computing power, data availability, and theoretical concepts. Even though the start of this field research can be traced back to Alan Turing (1950), the relatively recent explosion of computing power and the rise of the internet, digitisation, and cloud computing led to an explosion of data, enabling the development of more sophisticated algorithms. One of these evolutions is represented by Artificial Neural Networks (ANN), which are a sub-class of ML models inspired by the structure and functioning of the human brain's neural networks, and consist of interconnected nodes, known as artificial neurons or units, organised into one or more layers. A further advancement is brought by Deep Neural Networks, characterised by multiple hidden layers between the input and output layers. This gives an important advantage in terms of capabilities and complexity, but at the cost of requiring substantial computational resources or specialised hardware. Here follows a Venn diagram of machine learning classes:

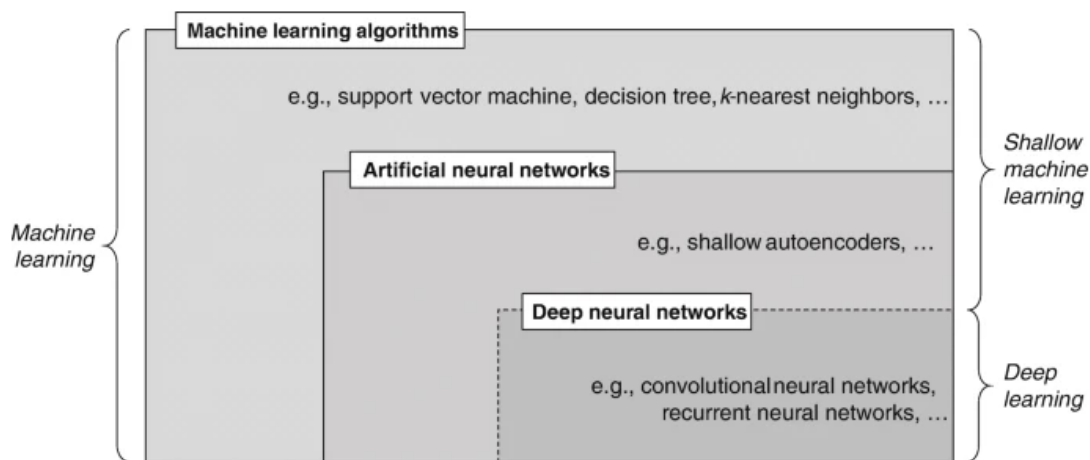


Figure 2.1: Machine learning diagram. Retrieved from Janiesch, Zschech and Heinrich (2021)

The author of the Thesis implements both shallow and deep machine learning models. The process of implementation of a ML model requires three fundamental steps:

- **Training:** during training, the algorithm adjusts its parameters or internal representations to minimise the difference between predicted outcomes and actual results. The goal is to optimise the model's performance on new, unseen data.
- **Validation:** after training, the model's performance needs validation on a separate dataset to assess how well it generalises to new data. Metrics such as accuracy, precision, AUROC score, or others specific to the problem domain are used to evaluate the model's performance and to fine-tune by adjusting hyperparameters.
- **Deployment:** once satisfied with the model's performance, it can be deployed to make predictions or classifications on new data.

ML models are quite heterogeneous, and can broadly be categorised into three main groups:

- **Supervised learning:** involves training a model on labelled data, where the algorithm learns from input-output pairs to make predictions or classifications when given new, unseen data. These models could be split further into Classification model in which the output variable has a finite number of categories, and Regression model in which the output value is a real or continuous value. Supervised learning Classification algorithms represent the core of this Thesis, as well as the main tool used by the cited researchers.
- **Unsupervised learning:** deals with unlabelled data and finds hidden patterns or intrinsic structures within it when there are no corresponding output variables. Common tasks include clustering, dimensionality reduction, and association rule learning.
- **Reinforcement learning:** focuses on making a sequence of decisions to achieve a cumulative reward. Agents learn through trial and error by interacting with an environment with the goal of maximising the reward.

Knowing the general functioning of these methods is also important for a later understanding of the results, their variability, and the differences with the results obtained by traditional econometric systems. For this regard, a description of the Logit Regression is also included since it will act as the main benchmark in this work and in great part of the papers in the literature.

## **Logit Regression**

Logit (or Logistic) Regression is a statistical method which origins trace back to the 18<sup>th</sup> century. It is used for binary classification tasks, where the goal is to predict the probability of an observation belonging to a certain class. The output of the regression is a score between 0

and 1, interpretable as the likelihood of an observation belonging to a particular class, and in this Thesis represents whether the analysed year constitutes a pre-crisis period or not. This model has been widely used in the past, most noticeably in the medic field to assess the likelihood of a disease given a set of symptoms. Logistic Regression takes advantage of the Logistic function to model the relationship between explanatory variables and the dependent variable, with the Logistic function being:

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}$$

with  $z$  being the linear combination of the independent variables and their coefficients:

$$z_a = \beta_1 x_{1a} + \beta_2 x_{2a} + \dots + \beta_n x_{na}$$

The Logit Regression then estimates the coefficients ( $\beta_0, \beta_1, \dots, \beta_n$ ) so to minimise the error between predicted probabilities and actual observations, by maximising the likelihood function. The Logit Regression is the preferred econometric tool on this field of research given its output, a probability comprised between 0 and 1 for any predictors' values, over the linear regression which output values may go from  $-\infty$  to  $+\infty$  as in this graph:

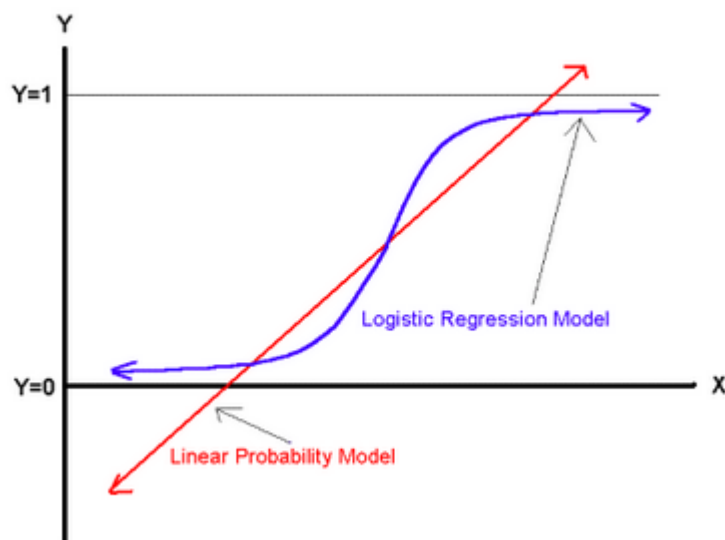


Figure 2.2: Logit and Linear Regression. Retrieved from [econometricstutorial.com](http://econometricstutorial.com)

Logit Regression has the advantage of easy interpretability since the resulting coefficients represent the effect of each independent variable in the log-odds of the dependent variable. However, Logit Regression assumes a linear relationship between all explanatory variables and the log-odds of the dependent variable, which is a weak assumption in the economic field this Thesis is studying. This is stressed by Lo Duca and Peltonen (2013) who argue that the probability of a crisis increases non-linearly as the number of fragilities increases.



## Binary Classification Trees

Binary Classification Trees (BCT), also known as Decision Trees, is a supervised machine learning model used for solving classification problems where the target variable has two possible outcomes (hence, binary classification). This technique originates from the field of statistics and operations research and can be dated back to the late 1950s, but one main precursor of this technique is the one proposed by Quinlan (1986). These models make predictions by recursively partitioning the feature space into regions based on feature values, and at each step, the data is split into “purer” sub-samples (also called child nodes) based on the feature that best separates the classes, that is, in this Thesis, whether the probability of a crisis either increases or declines significantly compared with the sample average (Dutttagupta, 2011). There may be many different measures quantifying the homogeneity of classes within each partition to be used for splitting the datasets, but in the financial crisis prediction literature the predominant parameter is the Gini Index, which corresponds to the following impurity function  $i(t)$  to be minimised:

$$i_{\text{gini}}(t) = \sum p_0(t)p_1(t)$$

Another, less used, candidate is the entropy (or information gain) index:

$$E = - \sum p_i * \log(p_i)$$

These indexes reach a minimum value of 0 when the child nodes contain only one class of observation, either pre-crisis or tranquil period. The process starts at the root node, which includes the entire dataset, and then after the first split the process continues recursively for each resulting partition until a stopping criterion is met (e.g., maximum tree depth or minimum number of samples in a leaf node). Once a stopping criterion is reached, the process creates terminal nodes or “tree leaves” that contain the predicted class for that region of the feature space. To make a prediction for a new sample, the forecaster traverses the tree from the root node down to a leaf node based on the feature values of the sample.

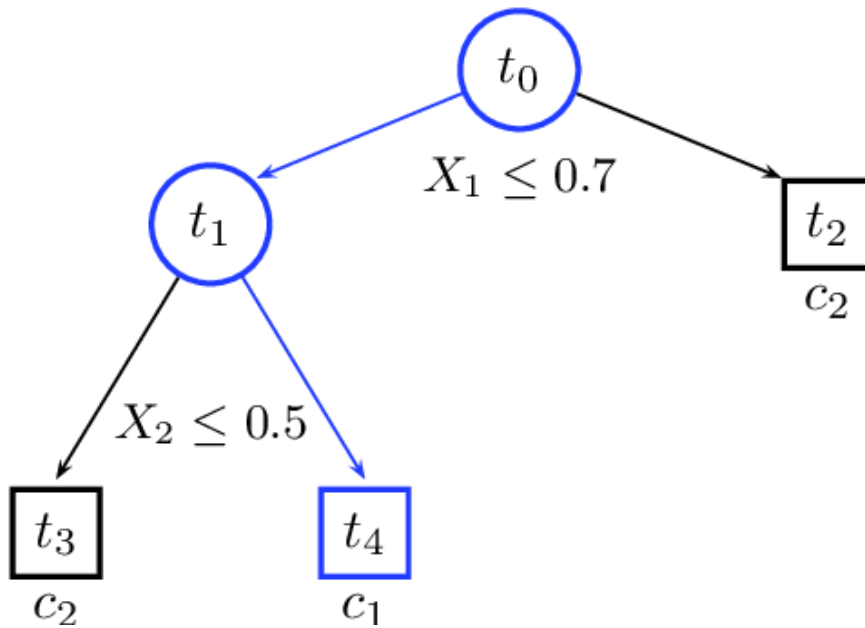


Figure 2.3: Simple representation of a BCT outcome

In this Thesis, BCT will be implemented using the Python scikit-learn library *DecisionTreeClassifier*, and the hyperparameters will be chosen among the following so to achieve the best results:

```

# Define the parameter grid for tuning
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [None],
    'min_samples_split': [5, 10, 15],
    'min_samples_leaf': [5, 10, 15]
}

```

Parameter ‘*criterion*’ determines the function used to measure the quality of a split. It can be set to “*gini*” for Gini Index or “*entropy*” for information gain, ‘*max\_depth*’ sets the maximum number of levels in the decision tree, ‘*min\_samples\_split*’ is the minimum number of samples required to split an internal node, and ‘*min\_samples\_leaf*’ is the minimum number of samples required to be at a leaf node. It helps control overfitting by stopping the splitting of nodes if the number of samples is below this threshold. BCT have the advantage of great interpretability since the decision rules are easy to visualise and understand, and the best thresholds to split the data are explicitly pointed out by the algorithm. On the other hand, BCT can be prone to overfitting reducing the real-world forecasting capabilities, it is unable to provide the contribution of a particular variable and, given that at each node the model identifies one variable that best discriminates between pre-crisis versus tranquil-period cases, it could incorrectly omit other variables that are possibly equally good splitters (Duttagupta, 2011). It is important to notice that bigger trees fit better specific data noise but performs substantially

worse on new sets of observations drawn from the same population, as it is less likely to generalise out-of-sample data (Bluwstein et al., 2020).

## Random Forest

The main issue with BCT, overfitting, can be partially solved with the use of Random Forest (RF) algorithms while preserving the positive characteristics of the simpler algorithm. Random Forest is an ensemble learning technique, resulting mainly from the contribution of Breiman et al. (1984), that combines multiple decision trees to create a more robust and accurate model. The term ‘ensemble’ points to the notion of a collection of multiple decision trees, where each tree is trained independently on different subsets of the training data and features. The core idea behind Random Forest is to build multiple independent decision trees using random selection of data, creating multiple diverse subsets in a process called bootstrap sampling. At each split, a random subset of data is chosen for further splitting, ensuring diversity in each tree. The multitude of trees constructed constitutes a ‘forest’, the algorithm then combines all BCT predictions (usually with averaging) to improve overall accuracy and reduce overfitting.

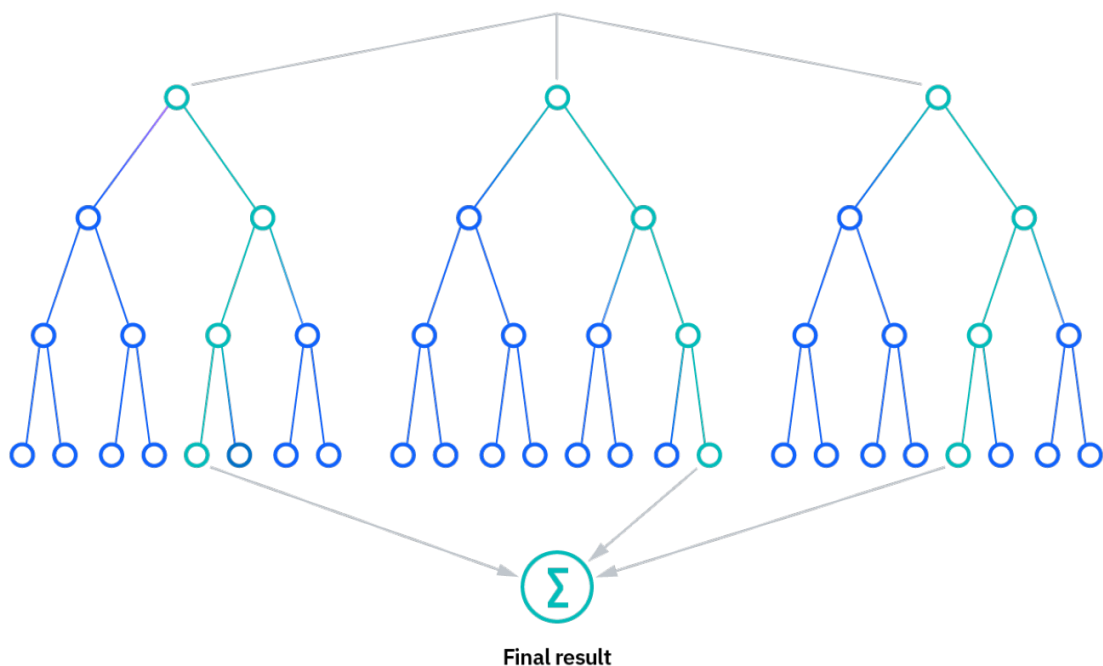


Figure 2.4: Simple Random Forest representation. Retrieved from [ibm.com](#)

The author implemented Random Forest using the Python scikit-learn library `RandomForestClassifier`. Hyperparameters used to grow the individual trees can be fine-tuned, such as maximum depth and minimum leaves, as well as the number of total trees in the forest. In the author’s work, they are drawn from the following:

```
# Define the parameter grid for tuning
param_grid = {
    'n_estimators': [50, 100, 150, 200],
    'criterion' : ['gini', 'entropy'],
    'max_depth': [None, 5, 10, 15],
    'min_samples_split' : [5, 10, 15],
    'min_samples_leaf' : [5, 10, 15]
}
```

The parameter '*n\_estimators*' set the number of individual trees to be grown, while the other four parameters assume the same role as in `DecisionTreeClassifier`. Compared to the single BCT, Random Forest reduces overfitting by introducing randomness so to generalise better to new unseen data. It is also less sensitive to outliers and can provide a measure of feature importance based on how much each feature contributes to the model's accuracy. However, single Classification Trees may be highly correlated, therefore decreasing the general accuracy of the bootstrapped classifier.

## AdaBoost

AdaBoost (Adaptive Boosting) is an ensemble learning method firstly proposed by Freund and Schapire (1996). It combines multiple weak learners or classifiers to create a strong learner, improving the accuracy of weak models by sequentially training them on different subsets of the data, emphasising the misclassified samples from previous iterations so to give larger weights to the observations that are more difficult to predict. The weaker classifiers are simple models such as single level decision trees with relatively low predicting power when considered individually, often performing only slightly better than random guessing. AdaBoost works in iterations, starting by assigning equal weight to each observation and then continuing in further iterations by assigning higher weights to the samples that were misclassified in the previous one. After training each weak learner, AdaBoost assigns a weight (or importance) to it based on its performance in classifying the training data. Models that perform better receive higher weights. The final prediction is obtained through a weighted sum of the predictions made by all weak learners, so to create the strong learner.

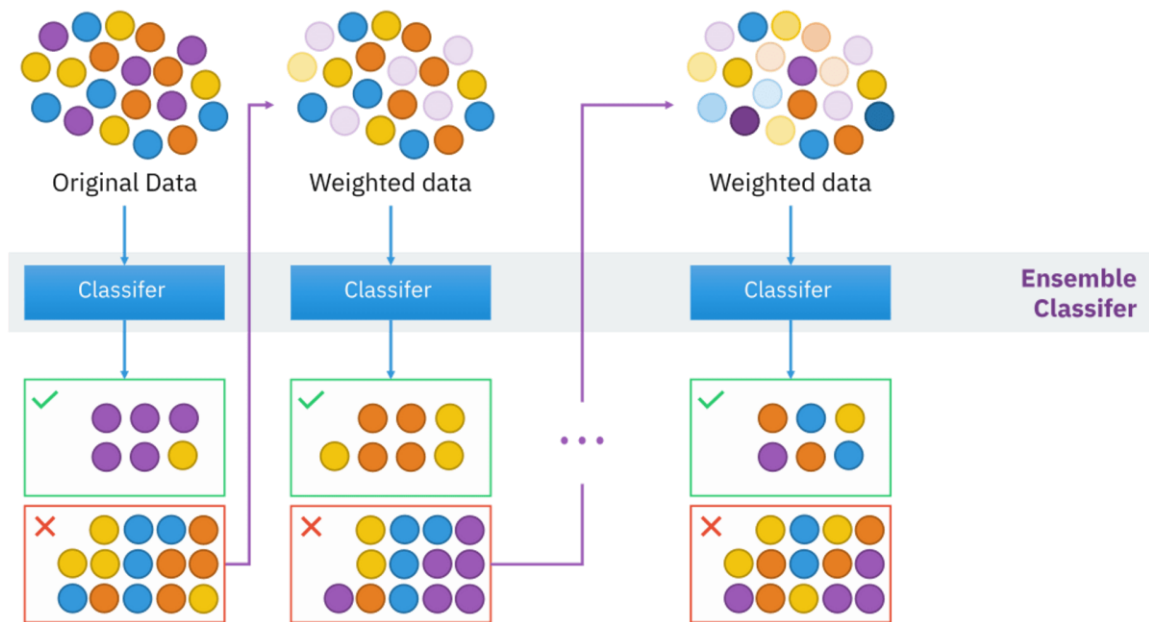


Figure 2.5: AdaBoost representation. Retrieved from [almabetter.com](http://almabetter.com)

The author of this dissertation implemented AdaBoost algorithm using the dedicated scikit-learn library *AdaBoostClassifier*. AdaBoost allows for a variety of hyperparameters selection to fine-tune the algorithm process, and in this dissertation, they are chosen among these:

```
# Define the parameter grid for tuning
param_grid = {
    'n_estimators': [50, 100, 200],
    'learning_rate': [0.1, 0.5, 1.0, 2.0]
}
```

Parameter ‘*n\_estimators*’ represents the number of weak learners or iterations, while ‘*learning\_rate*’ controls the magnitude of the update applied to the weights of misclassified samples from the previous iteration to emphasise their importance in subsequent rounds. The term ‘adaptive’ in Adaptive Boosting comes from the ability of this algorithm to adapt to instances difficult to classify and learn from mistakes, and represents one of its main strengths. This model generally shows high predictive accuracy when confronted with simpler models, while reducing overfitting compared to single BCT. On the other hand, it may be sensitive to outliers in the data, has lower interpretability than the weak learners being a combination of multiple models, and it is computationally expensive especially when using a large number of weak learners.

## K-Nearest Neighbours

K-Nearest Neighbors (KNN) is an instance-based supervised learning algorithm, with instance-based meaning that it stores the entire training dataset and makes predictions based on the similarity of new instances to existing data points. The assumption is that similar data points tend to have similar labels or outcomes, so the algorithm assigns the data points close to each other in the feature space to the same class. KNN was first developed by Fix and Hodges (1951) and expanded by Cover and Hart (1967). KNN can be used for classification or regression, with the former being the choice for banking crisis prediction. In the training phase, KNN stores the entire training dataset and for each new, unseen point of data, it calculates the distance between that point and all other points in the training dataset. Then, it identifies the  $k$  nearest neighbours to the new data point based on the calculated distances where ‘ $k$ ’ is a discretionary parameter. For classification tasks, the algorithm assigns the class label most common among its  $k$ -nearest neighbours (based on majority voting).

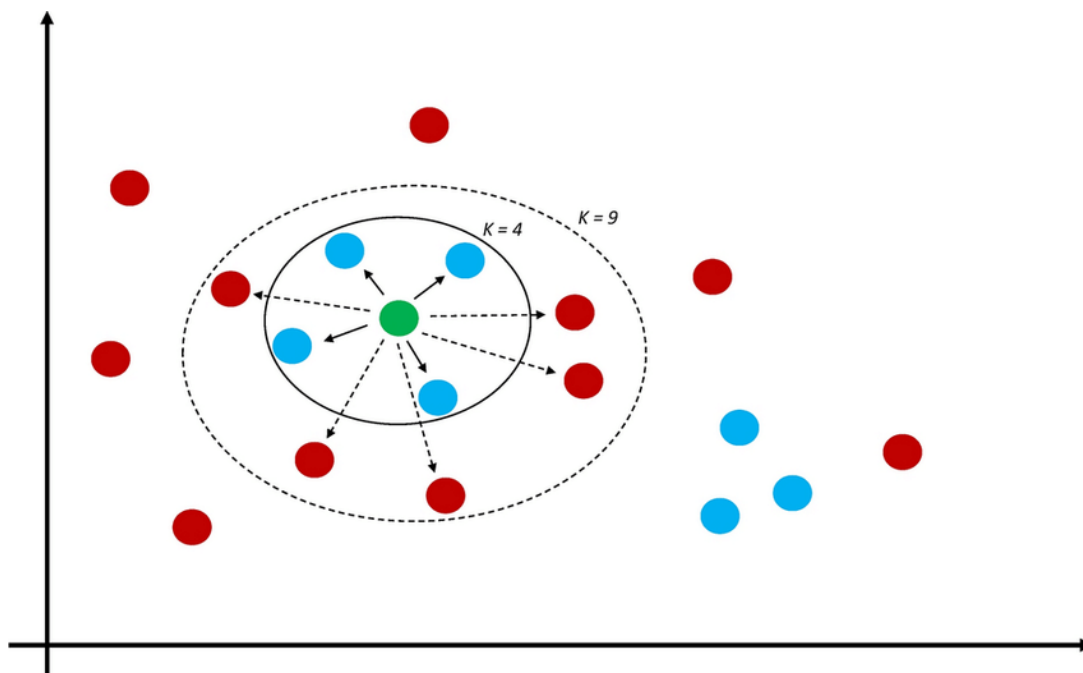


Figure 2.6: KNN representation. With  $K=4$  the green observation is classified as ‘blue’, with  $K=9$  it is classified as ‘red’.  
Retrieved from researchgate.net

As can be seen from Figure 2.6, KNN is very sensible to the choice of the parameter ‘ $k$ ’ which can significantly impact the results. If it is too small, the estimation could be poor given data scarcity, if it is too big, the estimations can be too smooth lowering prediction power. The author of this Thesis implemented KNN from a dedicated library from scikit-learn, *KNeighborsClassifier*, and the hyperparameters are chosen among the following:

```

#Define the parameter grid for tuning
param_grid = {
    'n_neighbors': [15,25,35,45],
    'weights': ['uniform', 'distance'],
    'metric': ['euclidean', 'manhattan', 'minkowski'],
    'p': [3]
}

```

Where the parameter '*n\_neighbors*' is equivalent to the number of neighbours '*k*' as described above, and '*weights*' defines the weight function used in prediction. It can take values '*uniform*' (all neighbours contribute equally) or '*distance*' (closer neighbours have more influence). Parameter '*metric*' is the distance measure used by the algorithm to determine the distance value between the *k* data points and the test sample. '*euclidean*' refers to Euclidean distance and is the most common metric, measuring the straight-line distance between two points in space measured as follow:

$$d(P, Q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

where  $P = (p_1, p_2, \dots, p_n)$  and  $Q = (q_1, q_2, \dots, q_n)$  are the coordinates of the two points in the *n*-dimensional space. '*manhattan*' refers to the Manhattan distance, also known as the L1 distance or taxicab distance, and is computed as follows:

$$d(P, Q) = \sum_{i=1}^n |q_i - p_i|$$

where  $P = (p_1, p_2, \dots, p_n)$  and  $Q = (q_1, q_2, \dots, q_n)$  are the coordinates of the two points in the *n*-dimensional space. The last parameter, '*minkowski*', is a generalised distance metric that includes both the Euclidean distance and the Manhattan distance as special cases. For two points *P* and *Q* in an *n*-dimensional space, the Minkowski distance '*d*' between these points is calculated using the following formula:

$$d(P, Q) = \left( \sum_{i=1}^n |q_i - p_i|^p \right)^{\frac{1}{p}}$$

The '*p*' parameter defines the degree of the Minkowski distance when '*metrics*' value is '*minkowski*'. When *p*=2, it represents the Euclidean distance, and when *p*=1, it represents the Manhattan distance. The author decided to set it equal to 3, so obtain a metric different from the other two when the value '*metrics*' is set to '*minkowski*'. KNN is simple to implement and understand, providing transparency in predictions but at the cost of sensitivity to irrelevant or redundant features, as it considers all data points equally important.

# Support Vector Machines

Support Vector Machines (SVM) is an advanced and versatile supervised machine learning algorithm used for both classification and regression tasks, first developed by Cortes and Vapkin (1995). It's particularly effective in high-dimensional spaces and is widely used in various domains such as image classification and text classification. In the training phase, the primary goal of SVM is to find the best possible decision boundary (hyperplane in higher dimensions) that separates the dataset points belonging to different classes while maximising the margin, which is the distance between the hyperplane and the nearest data points of each class. These nearest data points are the 'support vectors', and their location influences the position and orientation of the decision boundary. SVM works well for linearly separable data, and when the dataset is not linearly separable then separability is achieved (or at least enhanced) by employing different 'kernel tricks'. These kernel tricks implicitly map the data into a higher-dimensional space where it becomes linearly separable, so to make it possible to find the optimal hyperplane. The optimal hyperplane is that which maximise the margin while minimising classification errors. Once the training phase is complete, new data points are simply classified based on which side of the hyperplane they fall.

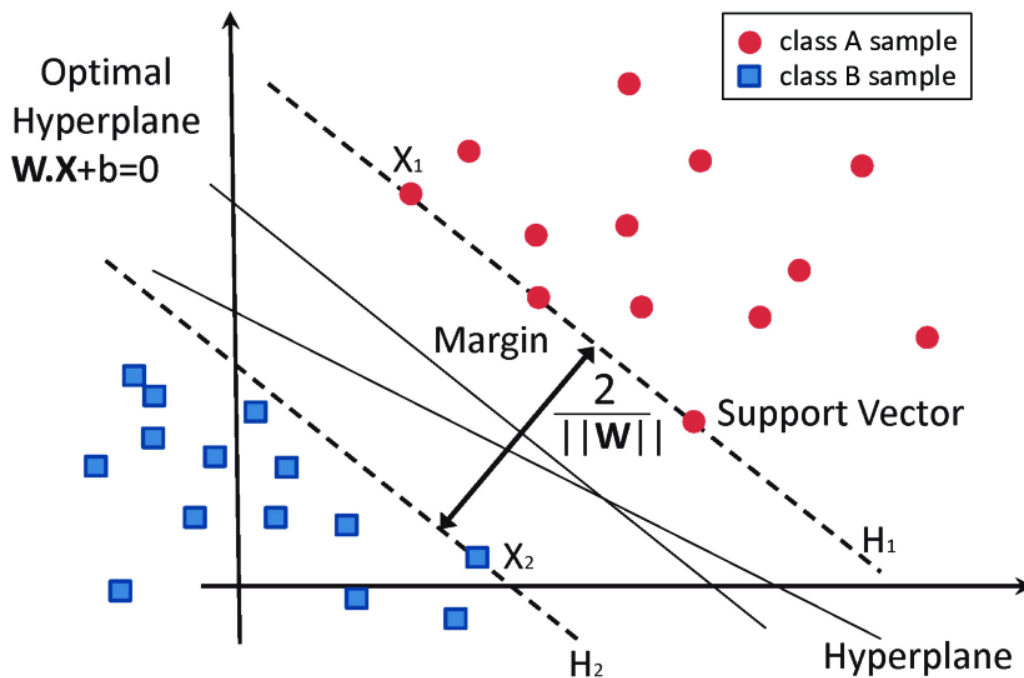


Figure 2.7: Representation of a two-dimensional SVM. Retrieved from researchgate.com

The author of this study implemented SVM algorithm in Python using a dedicated library in scikit-learn, *SVC*. The hyperparameters are selected from:



```

# Define the parameter grid for tuning
param_dist = {
    'C': [0.75, 1, 1.25],
    'gamma': ['scale', 'auto'],
    'kernel': ['linear', 'rbf', 'poly', 'sigmoid']
}

```

Value ‘*C*’ is the regularisation parameter. It controls the trade-off between maximising the margin and minimising the classification error by introducing a misclassification penalty. Adding tolerance towards misclassification (lower ‘*C*’) allows for larger margins and then more robust classification towards perturbations of the original data, for example when predicting the label of new data points (Beutel et al., 2018). ‘*kernel*’, as explained previously, is the type of function that maps the data onto higher-dimensional space and can assume four different values: ‘*linear*’ is the simplest kernel and represents a linear relationship between input features:

$$K(x, x') = (x^T x' + c)^d$$

Value ‘*rbf*’ points to the Radial Basis Function, the most favoured in the literature, which uses a Gaussian function to project data into an infinite-dimensional space:

$$K(x, x') = e^{(-\gamma \|x - x'\|^2)}$$

and in which  $\gamma$  (gamma) is a hyperparameter that controls the influence of each training example and is set through the value ‘*gamma*’ in the scikit-learn library. ‘*poly*’ is the Polynomial Kernel:

$$K(x, x') = (x^T x' + c)^d$$

where  $d$  represents the degree of the polynomial, and  $c$  is a constant term. By default, using the scikit-learn library, these values are equal to 3 and 0 respectively. The last kernel is ‘*sigmoid*’, which corresponds to the Sigmoid kernel based on the hyperbolic tangent function:

$$K(x, x') = \tanh(ax^T x' + c)$$

Given its nature, SVM can effectively manage large feature sets and work well with both linearly separable and non-linearly separable data thanks to the different kernels available. On the other hand, choosing the appropriate kernel could be challenging and introduces discretion into the model. SVM is also very computationally expensive, especially with large datasets, and lacks interpretability.

## Artificial Neural Networks – Multi-layer Perceptron

Artificial Neural Networks (ANN) are a fundamental concept in machine learning and artificial intelligence, inspired by the structure and functioning of the human brain. They are a set of algorithms which try to mimic the way a biological brain processes information. They are one

of the most researched techniques in AI development during recent years, even though their origin can be traced back to the 1940s, with McCulloch and Pitts (1943) being credited with creating the first mathematical model of an artificial neural network used as a logic gate. These algorithms achieved impressive results in recent years, ranging from image recognition to generative AI. ANNs are composed of interconnected nodes called neurons arranged in layers: input layer, hidden layers (if multiple), and output layer. Each neuron receives inputs, processes them using an activation function, and produces an output. Activation functions introduce non-linearity into the network, enabling it to learn complex patterns, and could be of many different variations based on researchers' discretion. Since the author of this dissertation uses this model to predict financial crisis with a set of variables, the input layer will be composed of a neuron for each predictor, and the output layer will consist of just a single neuron to produce a continuous value representing the probability of a crisis. Neurons are connected through links that have an associated weight, each representing the strength of the connections, that are adjusted during the training process impacting the network's behaviour.

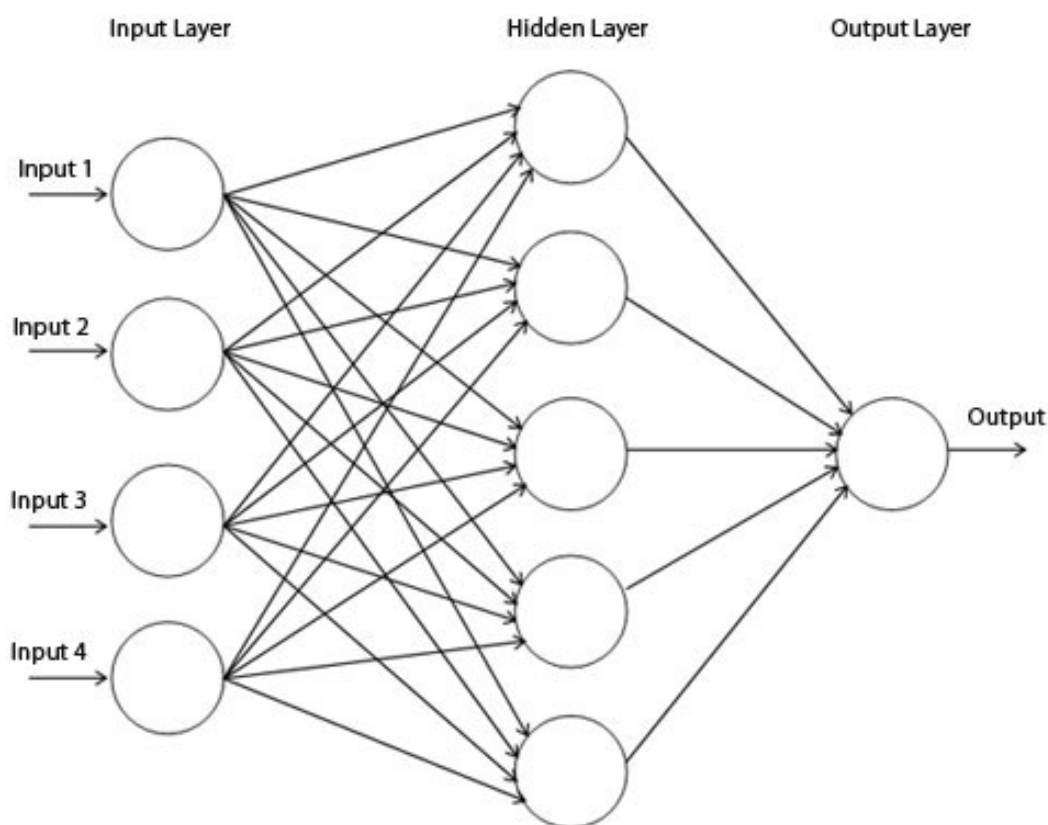


Figure 2.8: Example of a single hidden layer ANN with 4 variables. Retrieved from nicolamanzini.com

In the training phase, data is fed into the network through the input layer and proceeds to the following layers in a process called 'forward propagation'. Each neuron's weighted inputs are summed, passed through an activation function, and forwarded to the next layer.

Backpropagation is then used to adjust the weights of connections so to minimise the difference between predicted and actual outputs using a loss function (cross-entropy for classification purposes). Once training is completed after a pre-determined number of iterations, the final layer (output layer) produces the network's prediction or output based on the learned patterns in the data. ANNs with this kind of structure in the literature often take the name of ‘multi-layer perceptron’ or MLP. Several hyperparameters could be changed to affect the behaviour of the network, and in the dedicated scikit-learn library *MLPClassifier* used in this Thesis are drawn from the following:

```
# Define the parameter grid for tuning
param_grid = {
    'hidden_layer_sizes': [(5,), (10,), (5, 5), (10,10)],
    'activation': ['relu', 'tanh', 'logistic'],
    'solver': ['adam', 'sgd'],
    'max_iter': [10000]
}
```

‘*hidden\_layer\_sizes*’ represents the architecture of the neural network by defining the number of neurons and the number of hidden layers. Value *(10,)* specifies a single hidden layer with 10 neurons, while *(5, 5)* specifies two hidden layers with 5 neurons each. This is the most important parameter since it dictates the ability of the network to capture patterns and to map non-linearities in data. As shown by Cybenko (1989), a single hidden layer containing a finite number of neurons (even a single neuron) can approximate a wide range of continuous functions, even if a more recent paper by LeCun, Bengio and Hilton (2015) highlights the benefits of deep architectures in learning intricate representations. Regarding the number of nodes in each hidden layer, there is no rule for an optimal number since too few neurons might lead to underfitting (the network might not capture complex patterns) and too many nodes might result in overfitting (the network might learn noise in the data and fail to generalise). A commonly suggested rule of thumb is to use a number between the number of input and output nodes. ‘*activation*’ refers to the activation function present in each neuron, and which determines how the weighted sum of inputs received from the previous layer is transformed and then passed to the successive layer. ‘*relu*’ stands for Rectified Linear Unit and is the simplest activation function. It outputs 0 for negative inputs and the input value for positive inputs:

$$f(x) = \max(0, x)$$

‘*logistic*’ corresponds to the logistic (or sigmoid) function and squashes the output to values between 0 and 1:

$$f(x) = \frac{1}{1 + e^{-x}}$$

‘*tanh*’ is similar to the logistic function but outputs values between -1 and 1:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

The ‘*solver*’ parameter specifies the optimisation algorithm used to train the Multi-Layer Perceptron (MLP) neural network, influencing how the network weights are updated in the training phase. ‘*adam*’ stands for Adam Optimiser, which computes adaptive learning rates for different parameters from estimates of first and second moments of the gradients. ‘*sgd*’ stands for Stochastic Gradient Descent and implements the standard stochastic gradient descent algorithm. The ‘*max\_iter*’ parameter specifies the maximum number of iterations (or ‘epochs’) the neural network will undergo during the training phase. High values could imply the need for heavy computational efforts, while low values might result in underfitting. In conclusion, MLP models constitute a versatile and universal category of algorithms, capable of extracting higher-level abstractions from raw input data even in the presence of non-linearities. Nevertheless, MLPs are prone to overfitting, especially with a limited dataset, are sensitive to hyperparameter tuning, and most importantly for the purpose of this research, they act as a ‘black box’ making it difficult to understand how the model reaches its final predictions.

## Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of artificial neural networks specifically designed to work with large sequential data by retaining memory or state information. They were initially proposed by Werbos (1988) to be used on a model of natural gas market. Unlike traditional feedforward neural networks, RNNs have connections that form direct cycles, allowing them to exhibit temporal dynamic behaviour. They excel in handling sequential data such as time series, text, speech, and video frames. Connections between layers are constructed so to loop back on themselves, allowing information to persist and be passed from one step of the sequence to the next, maintaining memory of past information. The internal representation or memory of the network at a particular time step during sequence processing gets the name of ‘hidden state’. At each time step  $t$ , the RNN takes an input  $x_t$  and combines it with the previous hidden state  $h_{t-1}$  to calculate the current hidden state  $h_t$ . The update equations in a basic RNN are often formulated as:

$$h_t = \text{Activation} (W \cdot [h_{t-1}, x_t] + b)$$

where  $W$  represents weights,  $x_t$  is the input at time  $t$ ,  $b$  is the bias term, and ‘Activation’ is an activation function.

# Recurrent Neural Network

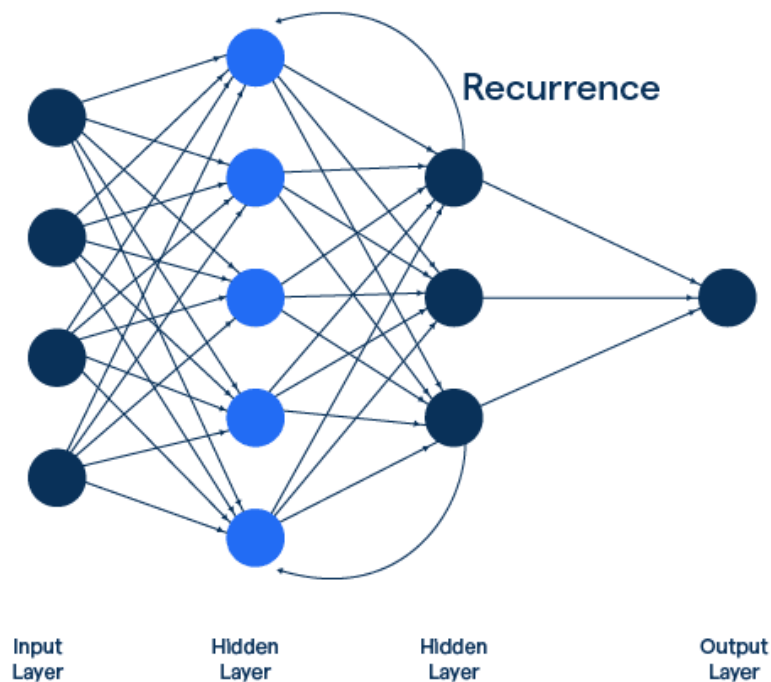


Figure 2.9: RNN structure with two hidden layers. Retrieved from [botpenguin.com](http://botpenguin.com)

This basic form of RNN, although very capable in handling sequential data, suffers from some negative aspects, such as the difficulty in capturing long-range dependencies in sequences due to their short-term memory, and the vanishing and exploding gradient problem, where gradients become too small or too large during training. More sophisticated variations of the basic RNN can be implemented to try overcoming these shortcomings. Long Short-Term Memory (LSTM) is a type of architecture designed to address the issues in handling long-term dependencies within sequential data. The key innovation in LSTM is the incorporation of memory cells, input gates, forget gates, and output gates, allowing them to effectively capture and maintain information over extended sequences. These four different components functions are:

- Memory cells store and regulate the flow of information through the cell state, acting as a conveyor belt, allowing information to persist, or be discarded.
- Input gates determine how much new information should be stored in the memory cell. It selectively updates the cell state based on the current input.
- Forget gates decide what information should be removed or forgotten from the cell state. It selectively removes irrelevant or outdated information.

- Output gates control how much information from the memory cell should be revealed or used to make predictions for the current time step.

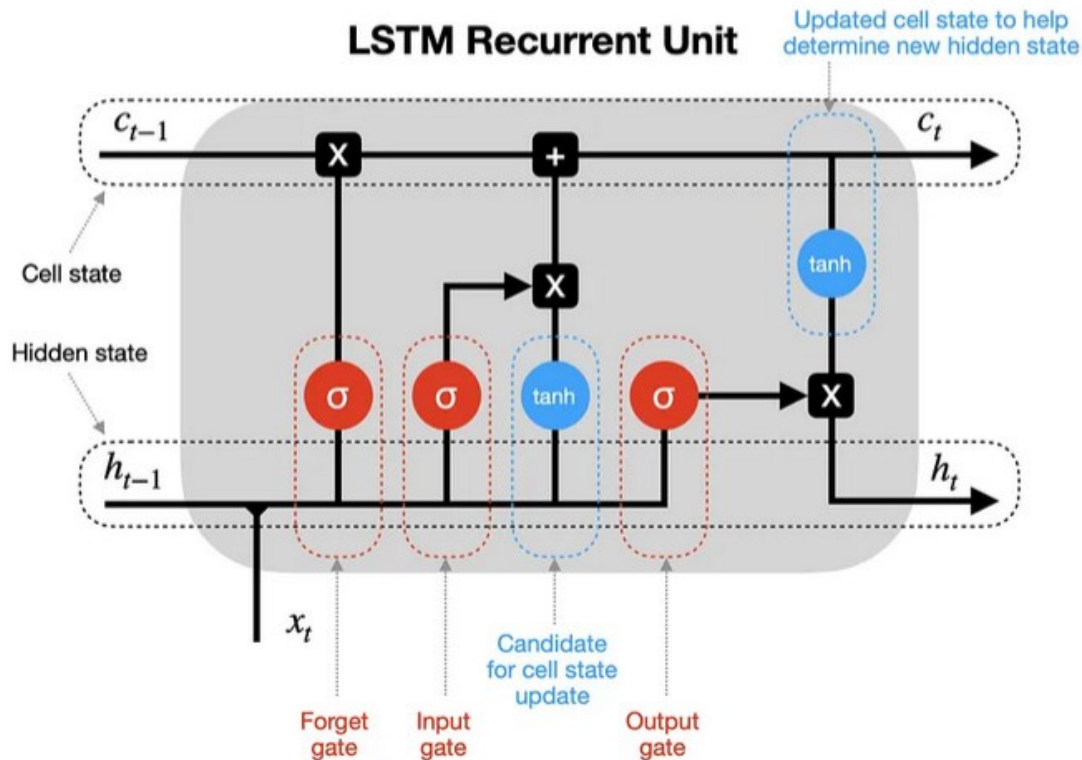


Figure 2.10: Representation of LSTM. Retrieved from [towardsdatascience.com](https://towardsdatascience.com)

In python, LSTM can be implemented through the dedicated library *TensorFlow*, thanks to the dedicated module. In conclusion, by regulating the flow of information through the memory cell, LSTM can avoid the vanishing gradient problem and capture important information across extended sequences more effectively than traditional RNN models.

## Hyperparameters tuning with k-fold cross-validation

As explained in the previous section, each machine learning model incorporates multiple parameters which values can significantly alter the behaviour of the algorithms and, consequently, the predicted probability of being in a pre-crisis state. These settings are not learned during the training process but are set prior to training. They can be assigned either discrete values, such as the kernel type used in SVM, or continuous values, for example the 'C' value in the same model. The choice of parameters is therefore potentially infinite. k-fold cross-validation is an essential technique for hyperparameter tuning as it helps in assessing how well a set of parameters make a model able to generalise to new information, maximising the use of available data. It splits the test dataset into  $k$  subsets or folds of approximately equal size,

training the model on  $k-1$  folds, and validating it on the remaining fold. This process is repeated  $k$  times, with each fold used exactly once as a validation set. After each iteration, the performance metrics are computed using the validation set. These metrics are recorded or averaged across all  $k$  iterations.

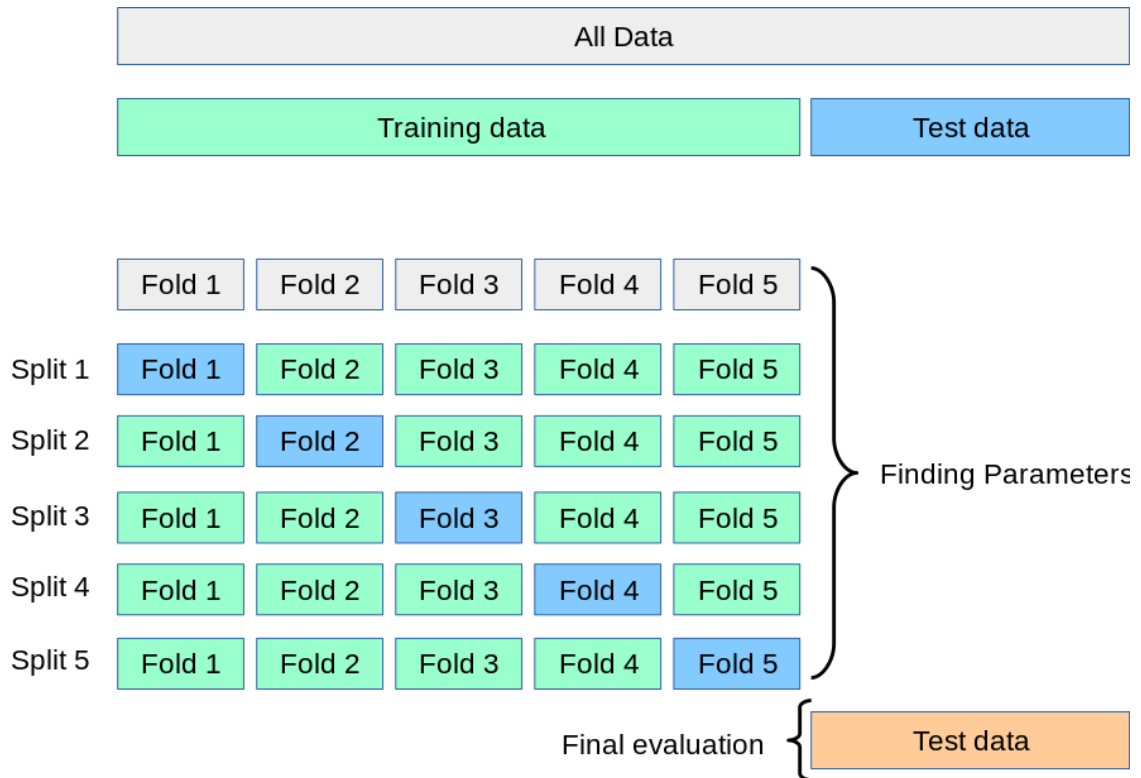


Figure 2.11: 5-fold Cross-validation diagram. Retrieved from scikit-learn.org

The most common performance metrics used by the literature when estimating the best hyperparameters are accuracy and the AUROC score, both of which will be explained in detail in the following chapter. Compared to a single train-test split, cross-validation allows for more reliable estimates of the model’s performance, and reduced variability. The value of  $k$  is on researchers’ discretion, but in the literature the dominating values are  $k=5$  and  $k=10$ . Smaller values of  $k$  lower the computational cost while increasing the performance estimate’s variance, and vice versa for larger  $k$  values. The author of this Thesis applied  $k=5$  to cross-validation along all models and used AUROC as the evaluation metrics.

# PERFORMANCE EVALUATION METRICS

Comparing the performance of different machine learning models is a crucial step in the model development process. It plays a pivotal role in selecting the best-performing models for a particular problem or task, and it is for these reasons that the author of this dissertation proposes a detailed summary of the main metrics used in the literature, as well as in this Thesis. These measures serve the role of comparing completely different models' performances, as well as comparing the performances of the same model when different hyperparameters' values are implemented, such as in k-fold cross-validation. Having at disposal multiple metrics also allows for a more nuanced assessment beyond simple accuracy, taking into account trade-offs between true positives, false positives, true negatives, and false negatives. Having clearly defined metrics helps in identifying the models which will be used by policymakers in a real-world framework, implying the critical importance and deep impact of this step.

## Contingency Matrix

A contingency matrix, also known as a Confusion matrix, is a tabular representation that allows the visualization of the performance of any classification algorithm. It compares the actual values of the target variable (in this Thesis, the observed pre-crisis/tranquil-period state) with the predicted values produced by the model. In this binary classification problem (two classes: positive and negative) the contingency matrix is a 2x2 table in which each prediction is univocally assigned to either one of the following categories:

- True Positives (TP): Instances that are correctly predicted as positive.
- True Negatives (TN): Instances that are correctly predicted as negative.
- False Positives (FP): Instances that are predicted as positive but are actually negative (Type I error).
- False Negatives (FN): Instances that are predicted as negative but are actually positive (Type II error).



		Actual class $C_j$	
		1	0
Predicted class $P_j$	1	<i>True positive (TP)</i>	<i>False positive (FP)</i>
	0	<i>False negative (FN)</i>	<i>True negative (TN)</i>

Figure 3.1: Contingency Matrix representation. Retrieved from Alessi et al. (2015)

Once the Contingency Matrix is filled-in with the predictions of the model correctly classified in the four categories, it is possible to compute five different metrics to be used as performance indicators, each of which with its shortcomings:

- Accuracy:  $(TP + TN) / (TP + TN + FP + FN)$   
Accuracy is the share of correctly predicted observations to the total number of cases. It is widely used, but it can be misleading in imbalanced datasets where one class dominates as in this dissertation. In this scenario, accuracy can achieve high values even with poor performance in identifying the minority class. It also does not distinguish between types of errors.
- Precision:  $TP / (TP + FP)$   
Precision is the share of correctly predicted crises over total predicted crises. It neglects false negatives, so that high precision doesn't guarantee a good recall or ability to catch all positive instances.
- Recall (Sensitivity or True Positive Rate):  $TP / (TP + FN)$   
Recall is the share of correctly predicted crises over actual crises that verified. It neglects false positives, decreasing the ability to avoid false alarms.
- Specificity:  $TN / (TN + FP)$   
Specificity is the proportion of actual negatives correctly identified. It ignores false negatives, so that high specificity doesn't ensure good performance on positive instances.
- F1-Score:

$$2 * (Precision * Recall) / (Precision + Recall)$$

F1 is the harmonic mean of precision and recall. It is useful in imbalanced sets as it considers false positive and false negatives. However, it assumes that Type I and Type II errors have the same impact, which is highly debatable in the framework of this Thesis.

There is not a best metrics for an all-around use, so that choosing the most relevant metric(s) depends on the problem context, class distribution, and specific objectives.

Now that the structure of the contingency matrix and metrics are defined, there is a major problem to tackle. Many of the presented ML models have as output a continuous number comprised between 0 and 1, indicating the predicted probability of a crisis in the near future. These numbers must be ‘translated’ into an integer, either 0 or 1, in order to be assigned to a specific sector of the confusion matrix. A commonly used threshold is  $c = 0.5$ , so that probabilities  $\geq 0.5$  are assigned as class 1, and probabilities  $< 0.5$  are assigned as class 0. This criterion is also used by default in the scikit-learn library *GridSearchCV*, used for the implementation of k-fold cross-validation in Python. A more complex threshold widely used in the literature, for example by Alessi and Detken (2011), can be computed by maximising a so called ‘Relative Usefulness Function’ which includes a Loss function representing the interiorised preferences of the policymaker between Type I and Type II errors, as in the framework elaborated by Sarlin (2013). Type I errors and Type II errors can be defined as follow:

- Type I:  $T1 \in [0, 1] = FN/(TP+FN)$
- Type II:  $T2 \in [0, 1] = FP/(FP+TN)$

So that the Loss function can be written as:

$$L(\mu) = \mu T_1 + (1 - \mu) T_2$$

where the coefficient  $\mu$  is the relative preference between missing crises and issuing false alarms. The Relative Usefulness score is then a function of the value  $\mu$  defined as:

$$U_r(\mu) = 1 - \frac{L(\mu)}{\min(\mu, (1 - \mu))}$$

so to measure the difference in algorithm’s performance against a perfectly predicting model. The literature sets the value of  $\mu$  as 0.5 as standard, like in Casabianca et al. (2022). However, as argued by Alessi and Detken (2018), “after the global financial crisis policymakers’ preferences are likely to have become biased against missing crises, implying a lower threshold”. Knedlik (2013) estimated  $\mu$  from the European Commission’s scoreboard of

macroeconomic imbalances and found that the European Commission has a higher relative preference for avoiding type 1 error than type 2 error, so that  $\mu$  would be higher than 0.5. The author of this Thesis will set  $\mu = 0.5$  wherever the Usefulness Function is employed, following the literature standard.

## **Receiver Operating Characteristic – Area Under Curve**

As seen in the previous section, despite the many advantages offered by the Contingency Matrix, the discretion in imposing a threshold value when categorising continuous probabilities introduces critical factors that significantly impact the construction of the matrix and subsequent performance evaluation metrics. Another discretion factor introduced by the contingency matrix is the choice of the metrics, among accuracy, precision, Relative Usefulness, and the others. For these reasons, the literature predominantly adopts the ROC curve as the evaluation metrics. The Area Under the Receiver Operating Characteristic curve (AUROC or ROC-AUC) is a widely used performance evaluation metric for binary classification models in many fields, from finance to medicine and meteorology. ROC was introduced during World War II for analysing radar signals, when it was used by the United States Army to measure the ability of their radar receiver (hence the name) to correctly identify the Japanese aircrafts against signal noise. It assesses the model's ability to distinguish between the positive and negative classes across various threshold settings. The ROC curve itself is a graphical representation of the trade-off between true positive rate (Sensitivity) and false positive rate ( $1 - \text{Specificity}$ ) at all different classification thresholds. AUROC quantifies the entire two-dimensional area under the ROC curve, summarising the model's performance across all possible thresholds, assuming values between 0 and 1. An AUROC of 1 implies a perfect prediction model which achieves perfect separation between positive and negative classes, while a score of 0 implies a model which consistently predict the opposite outcome. An AUROC close to 0.5 implies a model which is not more informative than a naïve choice (the so-called coin toss). In general, the AUROC represents the likelihood that a classifier will prioritise a positive instance over a negative one in its ranking.

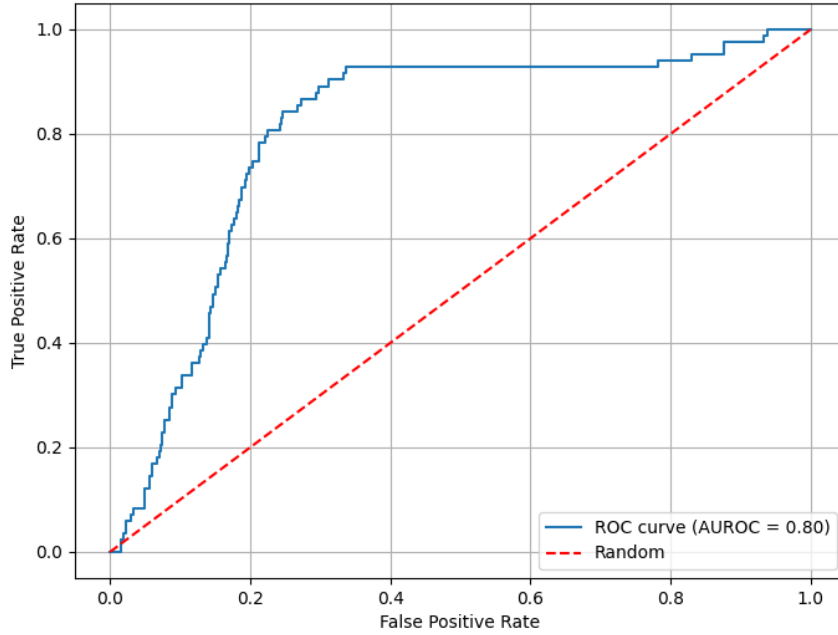


Figure 3.2: AUROC example. Author's elaboration

The author of this dissertation used AUROC as the main comparison metrics, given the *ad hoc* nature of setting a specific preference parameter  $\mu$  equal for all policymakers, and took a more general approach following the example of many researchers who chose ROC as main metrics. More formally, AUROC is:

$$AUROC = \int_0^1 TPR(FPR) dFPR$$

where TPR and FPR are the True positive Rate and False Positive Rate functions, respectively. In Python, the author implements AUROC using a dedicated module of the scikit-learn library, `roc_auc_score`.

## Brier Probability Score

A less common evaluation metrics in machine learning literature is the Brier Probability Score (BPS). BPS was firstly introduced by Brier (1950), and measures the mean squared difference between predicted probabilities and the actual outcomes for binary classification tasks. As a first step, for each prediction  $i$  a single Brier Score is computed as:

$$B_i = (P_i - O_i)^2$$

with  $O_i$  corresponding to the real observed binary outcome (either 0 or 1), and  $P_i$  corresponding to the predicted probability (continuous value between 0 and 1). After this, the Brier Score for

a single prediction is computed as the squared difference between the predicted probability and the actual outcome:

$$Brier\ Score = \frac{1}{N} \sum_{i=1}^N B_i$$

With  $N$  being the total number of predictions. Brier Score ranges from 0 to 1 and, inversely to AUROC, a lower score means better accuracy of predicted probabilities since Brier Score can be thought of as a cost function. In Python, the author of this work implements Brier Score using a dedicated module of the scikit-learn library.

## Shapley Values

The previous analysed metrics have the role of comparing different methods or choosing among different hyperparameters. However, for the scope of this dissertation, a further metrics to measure the impact of each individual predictor is needed. This is of course import for academic reasons, but it serves also a more pragmatic goal. Having such an indicator could help in pinpointing in a transparent way specific financial indicators or variables that significantly impact the likelihood of a financial crisis. This assists financial institutions, regulators, and policymakers in focusing on crucial risk factors for monitoring and mitigating potential crises. It also provides stakeholders with an explanation of how predictions are made, increasing the trustworthiness and acceptance of the model's predictions. Shapley values (Shapley, 1953), derived from cooperative game theory, are a concept used to fairly distribute the marginal contributions of players (or predictors, in this Thesis) in a coalition (or collection of predictors) to the overall payoff (the prediction). So, in the context of machine learning and model interpretability, Shapley values quantify the impact of each variable on a specific prediction made by a model. These metrics satisfy the property of additivity, meaning the sum of Shapley values for all features equals the difference between the model's prediction for a specific instance and the average prediction of the model across all instances. The computation of Shapley value  $\phi$  for observation  $j$  is carried out as follows:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [f(S \cup \{j\}) - f(S)]$$

where  $N$  is the total number of variables,  $S$  is the coalition of features excluding the predictor  $j$ ,  $f(S \cup \{j\})$  is the model's prediction when including feature  $i$  in coalition  $S$ , and  $f(S)$  is the model's prediction when excluding feature  $i$  from coalition  $S$ . A positive Shapley value signifies that an

escalation in the predictor's value augments the predicted probability of being in a pre-crisis, whereas a negative Shapley value suggests that an increase in the variable aligns with a reduction in the probability of being in a pre-crisis. In Python, the author implements Shapley Values using a dedicated library called *shap*.

# LITERATURE SURVEY

In this chapter, the author reviews the literature about predictive modelling of financial crises, separating the traditional models from newer and more complex methods.

## Statistical and Signal approach

While machine learning and neural network algorithms have become predominant in the banking crises forecasting literature in the last decade, the possibility of predicting such phenomena with quantitative means had been already explored by researchers in the past. This was done mainly using the signal approach, which studies the behavior of economic indicators both before and during a crisis event, and traditional econometric tools, namely Probit and Logit Regression. Specifically, a wave of research effort in this field can be observed at the end of the 1990s, probably induced by the financial crises that hit East Asian countries in the preceding years. The different authors state that results obtained are significant most of the time and allow to identify the main drivers of crisis episodes across different time spans and different countries.

**Eichengreen and Rose (1998)** tried to identify the variables that best anticipate the arising of banking crises in 105 developing countries from 1975 through 1992 using Probit Regression. They find a highly significant correlation between the increase in industrial countries' interest rates and banking crises in emerging countries, implying that domestic macroeconomic problems do not provide the entire explanation for banking crises. Specifically, they explain that the rise in external interest rates is observed the year prior to the onset of the crisis. Domestic factors remain important nonetheless, with overvalued real exchange rates and slowing output growth playing a significant role. Eichengreen and Rose also find that, contrary to part of the literature, there is little evidence of an independent role for domestic credit booms. The authors test the model out-of-sample by analysing the wave of banking crisis that hit East Asia in the second half of 1997, finding that the fit is not exceptionally good.

In the same year, **Demirgüç-Kunt and Detragiache (1998)** conducted a study using Logit Regression on all market economies for which data was available over the period 1980-1994, using exclusively domestic variables. Their findings show that the most impactful predictors of a banking crisis are slow GDP growth, real interest rate and inflation, confirming the well-known vulnerability of the banking system to these kinds of shocks, but also find that exchange rates do not have an independent effect once other variables are controlled for. The authors test

the predictive capabilities of the model computing the share of crises it can correctly classify, finding that it performs well with an overall classification accuracy between 67 percent and 84 percent. It is however to be noted that the tests were conducted in-sample.

**Kaminsky and Reinhart (1999)** examine currency and banking crises episodes in industrial and developing countries in a period that spans from the 1970s through 1995, using a signal approach on 16 indicators divided on financial, external, and real sector. The authors set a threshold value for each indicator such that the noise-to-signal ratio is minimised. The results obtained are significant, with no banking crises with less than 20 percent of the indicators signalling. For banking crisis, it is to note that the real sector (output and stock prices), the occurrence of a credit boom and lower terms of trade give the most consistent warning signals.

**Hardy and Pazarbasioglu (1998)** analyse 50 countries in the period 1980-1997. Again, the considered variables were split in three categories: real sector, banking sector, and potential shocks, analysed using a multivariate Logit Regression. They obtain reasonable predictive power in-sample, that is more than half of the episodes of banking system distress are predicted correctly, even though the predictors' significance decreases when trying to forecast the pre-crisis periods. The authors use the model to predict out-of-sample the East Asian crisis of 1997, correctly predicting three of the four episodes. The paper indicates as main events leading to banking distress: a fall in real GDP growth, a rise followed by a sharp fall in inflation and credit to the private sector, a fall in deposits at banks, a rise in real interest rates, a decline in the real effective exchange rate, and a sharp slowdown in the real growth in imports.

## **Machine Learning and Neural Networks approach**

At the beginning of the 2010s the literature attention switched to a more sophisticated approach in predicting financial woes. Researchers implemented a variety of machine learning algorithms to improve their prediction against the classic econometric approach. One of the main advantages of these algorithms is that, unlike Logit/Probit Regression, they can factor in the complex interactions between different variables and reflect these interactions in the output, which is the estimation of crisis probabilities. On the other end, some of these machine learning algorithms operate as a black box, making difficult to provide the marginal contribution of a particular variable and defend their predictions. Of all papers examined, the great part reaches the conclusion that these innovations in data analysis can bring a significant improvement in the forecasting capabilities. Nonetheless, a minority portion of research reached the conclusion that Logit Regression still maintains the highest chance of predicting a future banking crisis in



real-world out-of-sample observation. Here follows a summary of the main papers about the argument with a focus on the description of the methodology used in each case.

**Duttagupta and Cashin (2011)** are the first, to their knowledge, to use a Binary Classification Trees (BCT) to analyse banking crises, using as a test field 50 emerging markets around the world during 1990-2005, for a sample size of 711 observations. Industrial countries were excluded due to the perceived different nature of their financial crises. The crisis episodes are drawn from three sources: Caprio and Klingebiel (2003), Kaminsky and Reinhart (1999), and Carsten et al. (2004). Using this data, the authors have a disposal 38 crisis episodes. The economic indicators are drawn on an annual basis from the previous literature already mentioned, namely Demirgüç-Kunt and Detragiache (1998) and Kaminsky and Reinhart (1999), and comprises macroeconomic fundamentals, external environment and liquidity, monetary conditions, and financial sector health, for a total of 19 variables both domestic and international. A quick analysis of the indicators made in a signal approach fashion shows that macro-fundamentals variables are generally worse on the year prior to a banking crisis relative to tranquil time. Furthermore, both external and domestic monetary conditions tighten, and real domestic growth is much higher. The banking sector health is also worse off in the year preceding an event. An exception is the non-performing loans (NPL) ratio, which is counterintuitively lower than average before a crisis. Regarding the BCT specification, variables are lagged and all crisis observations after the first year of crisis are removed to avoid affecting the behaviour of the indicators. The resulting tree has 8 terminal nodes and predicts 37 out of the 38 episodes in-sample. The terminal node that predicts the most events is also the simplest and represents a state of macro-instability where inflation is above 18.7% and terms of trade (TOT) growth below 3.3%. In this scenario the authors see a situation of an already fragile banking sector due to inflation, that might ultimately break down when TOT deteriorates affecting trade prospects and the quality of banks' trade credit. The authors stress the improvements brought by BCT over traditional methods, in particular the ability to explain the interactions of different factors and to come up with a well-defined threshold or combination of thresholds which can act as a warning for regulators and supervisors if crossed. The indicators are then classified using the variable importance index, and find that the best performers are nominal depreciation, interest profitability, inflation, liability dollarisation, and bank liquidity. All these indicators are also splitters in the computed BCT.

**Alessi et al. (2015)** compare multiple different early warning systems. Crisis datasets were drawn from Babecký et al. (2012) and are comprised of quarterly observations of EU-27 and other OECD countries economic data, over a period from 1970 Q1 to 2010 Q4. Quarters which

fall within the period from three quarters before the onset of an event up until the end of the crisis are omitted from the analysis. Predictors are drawn from a variety of sources, including BIS, IMF, OECD, ECB, and others, and are lagged of one quarter to account for publication delay. The analysis is conducted in a real-time fashion, meaning that only information that is available at a particular point in time is used. Performance is evaluated for a homogeneous time window, specifically 20 to 4 quarters before a banking crisis. A performance measure is then obtained with the use of contingency matrix and absolute usefulness, other than the AUROC. It is important to note that different researchers of this paper elaborated different models, using different subsets of the data available or treating the same data with different transformations. The models implemented are Probit Regression, Logit Regression, dynamic Probit (in which lags of the dependent variable are added, accounting for possible time dependence), multivariate logit framework, Bayesian Random Coefficient Logit Panel model, univariate signaling approach, Bayesian model averaging, Binary Classification Trees, and Random Forest. The results of in-sample testing suggest that, even though every considered model was somewhat capable, Classification and Regression Trees (CART) models offer a better and stronger predictive capability with respect to traditional Logit and Probit models, and that every multivariate approach considered offers considerable improvements over univariate signalling variables in terms of crisis prediction performance. The best performing models, which are Binary Tree and Random Forest, identify as most significant indicators of an upcoming banking crisis a shallow yield curve coupled with high money market rates and low bank profitability (for Binary Tree), and house price, bank credit and public debt (for Random Forest).

**Alessi and Detken (2018)** present a technique based on decision tree learning, with a focus on credit development. Quarterly data is provided mainly by the dataset assembled by Babecký et al. (2014) and includes banking crises episodes in all EU countries over the period 1970-2013, even if the analysis is carried out only on euro area countries plus UK, Denmark, and Sweden. In this period 25 separate crisis events are recorded. The paper aims at building a model that correctly signals the event in the preceding four years, excluding from the analysis the three quarters immediately preceding the crisis and the crisis period itself. Early warning indicators consist of financial and macroeconomic variables, both domestic and global, as well as real-estate-based indicators, and are lagged of one quarter to proxy for publication lags. Credit-related indicators are also transformed using a backward-looking-slowly-adjusting Hodrick-Prescott filter, a year-on-year rate of growth, or a ratio to GDP. Other macroeconomic and market-based indicators are transformed on a year-on-year change or as a ratio to GDP. The authors run a Binary Classification Trees and a Random Forest analysis with 3 parts split for

out-of-bag observations but drawn results mainly from the Random Forest model given its greater robustness when additional predictors or observations are included. Testing the model in-sample, they obtain very strong results, with the model able to misclassify an incoming quarter of data only 6% of the time and with the ROC curve achieving a score of 0.94. Two logistic regressions are then estimated for comparison, even though the dataset cannot be the same as in the CART models due to the lack of a complete balanced panel required for the estimation of a pooled Logit Regression. The first Logit model prioritises the time dimension, while the second aims at extending the types of regressors included. The two regressions are assessed using AUROC on a 3-fold cross-validation to have a fair comparability with Random Forest, obtaining a score of 0.86 and 0.94 respectively, slightly lower than the Random Forest AUROC. Subsequently, the models are tested out-of-sample. A Random Forest is grown with data up to 2006 and tasked to predict the upcoming global financial crisis. Even if the models cannot achieve the same precision as in the in-sample exercise, the authors interpret the results as fairly good in identifying particularly vulnerable and safe countries and argue that the implementation of such a model would have at least triggered a discussion for borderline cases prior to the Lehman collapse. In particular, the predicted probability of a crisis assigned to Denmark, France, Greece, Ireland, and the UK was above 30%. Due to the nature of Random Forest classifier, the authors are not able to pin down the contribution of each predictor but argue that it is possible to interpret crisis probability trends based on latest developments in raw indicators, hinting at a possible narrative approach backed by quantitative analysis.

Less enthusiastic performance of the machine learning methods is measured by **Fricke (2017)**. The panel dataset used by the author comes from Schularick and Taylor (2012), who previously implemented a Logit model to predict financial crises and covers the period 1870-2008 (139 years) for 14 developed countries. The author applies to data the same filters as in Schularick and Taylor (2012) and uses the same lagged explanatory variable, real credit growth, with 5 lags. In total, 53 crisis events are analysed among 1,253 observations once world wars periods are removed. Data is then analysed using seven different supervised learning models: Logistic Regression, Classification Trees and Forest, K-Nearest Neighbors, single-layer Neural Networks, Quadratic Discriminant Analysis (QDA), and Support Vector Machines. As a first exercise the author evaluates the models' performance in-sample using ROC AUC, finding that all prediction models are better than random guesses ( $AUC > 0.5$ ) and that ML methods perform substantially better than the Logit Regression. In particular, they find that Classification Trees and KNN (with  $k=1$ ) are generally the best models, with the former yielding a score of 1, which implies perfect prediction. Moving to out-of-sample performance, the author uses 4 different

K-fold cross-validation approaches (with  $K = 2, 3, 4, 5$ ) with the added characteristic that each model is trained using information present on previous (in a temporal sense) blocks only. By doing so, the author deals with the time-series nature of data, making sure that portions of data are not used to predict events of the past ruining the goodness of out-of-sample results. The results are substantially lower compared to in-sample values, with some model achieving a score below 0.5. It is noteworthy to notice that the models performing best in-sample are the same with lowest scores out-of-sample, suggesting that overfitting is indeed an issue. To be specific, with a 5-fold split for training, Classification Trees now achieves an AUC of 0.461 and KNN-1 an AUC of 0.539. In this setting, the authors conclude that Logit Regression achieves better out-of-sample performance than the ML algorithms due to overfitting. Fricke then carries out a deeper analysis by adding further explanatory variables with 5 lags each for a total of 35 indicators ranging from broad money growth to inflation. Using Classification Forest, the author computes the variables importance by quantifying for each of them how much it reduces the overall classification error when it serves as a tree branch. The resulting most important feature in financial crises prediction is the second lag of credit growth, in line with the findings of Schularick and Taylor (2012). Nonetheless, credit growth's lags 3 and 5 appear to be relevant features despite labelled as insignificant in many specifications from Schularick and Taylor (2012).

**Joy et al. (2017)** implement CART methodology on an unbalanced dataset ranging from 1970 to 2010 consisting of quarterly observations on 36 advanced countries drawing on the results of the survey from Babecký et al. (2014). The authors aim to predict whether a crisis will occur 4-8 or 8-12 quarters ahead. To do so, data included in the period from 3 quarters prior to the crisis is removed from the analysis, as well as the observations gathered during the crisis itself to avoid "post-crisis bias". 20 potential macroeconomic and financial predictor variables are chosen, both domestic and international. Using this data, the Random Forest algorithm identifies as the most important variables the current account, short-term interest rate, and the yield curve slope, based on their score in terms of "mean decrease in the Gini coefficient". Of the original 20 predictors, only the best 10 identified by Random Forest are kept for Binary Tree analysis with a tree depth of only three levels to avoid overfitting. When trying to predict a crisis 4-8 quarters ahead, researchers find that net interest rate spread in the banking sector is the key predictor, and that crises are more likely when its value is low. On the contrary, when the value is high, then a flat or inverted yield curve becomes the crucial predictor. When trying to predict 8-12 quarters ahead, the most important predictor becomes house prices. The overall predictive power of the tree is relatively low, with overall correctness of in-sample fit estimated

at 40%. The authors proceed in the analysis by adding further variables that have limited variation across time but vary significantly across countries so to represent structural characteristics of each country. These added predictors range from financial development to overall tax burden. The analysis is carried out as before, with the best predictors selected by Random Forest and then analysed further with Binary Tree. This new baseline setting identifies as important information the country's financial development, trade openness, and industry share of GDP, while Binary Tree identifies a shallow yield curve as the primary splitter. The authors finally extend once again the initial baseline with international factors that have only time variation and no cross-country variation, such as commodity price inflation and world real GDP growth among others. For banking crises, world inflation and short-term interest rate are classified as most important, while world GDP growth is the main splitter in the Binary Tree elaborate. Surprisingly, the results imply that most of the banking crises were preceded by very strong world economic growth above the 80% quantile of its distribution.

**Holopainen and Sarlin (2017)** present a so called “horse race” of different conventional statistical methods and more recent machine learning algorithms, other than multiple aggregations of these models. The data collected is on a quarterly basis from Laeven and Valencia (2013) and Babecký et al. (2012) and covers the period 1976-2014 for 15 EU countries and includes 15 banking crisis events. The chosen early warning indicators cover a range of macro-financial variables, such as house prices, private loans, current account deficits, and many other, and are retrieved from Eurostat, OECD, ECB Statistical Data Warehouse, and the BIS Statistics. Most indicators are expressed as a ratio of GDP or in annual growth rates, while deviations from trend of indicators such as credit gap and asset prices are captured through a backward-looking HP filter. As comparing metrics, the authors choose a Usefulness indicator that considers the Loss Function of the policy maker, other than the usual ROC-AUC. The methods presented by the authors are signal extraction, Linear discriminant analysis, Quadratic discriminant analysis, Logit Regression, Logit Least Absolute Shrinkage and Selection Operators (LASSO), Naive Bayes, K-nearest Neighbors, Classification Trees, Random Forest, single-layer Artificial Neural Networks, Extreme learning machines, and Support Vector Machines (SVM). These models are trained so to forecast a crisis occurrence 5-12 quarters in advance, and post-crisis and crisis bias are accounted for by not including periods when a crisis is present or in the two years thereafter, as well as observation 1-4 quarters prior to the event. The authors use 10-fold cross-validation for two distinct purposes: selecting optimal free parameters and provide objective assessments of generalisation performance under out-of-sample testing. To test models in a real-time analysis fashion, Holopainen and Sarlin use a

recursive exercise that derives a new model at each quarter using only information available up to that point in time, also considering a publication lag of 1-2 quarters depending on the variable. Moreover, any given quarter is known to be tranquil only when the forecasting period has passed, so a window of equal length as the forecast horizon is dropped at each quarter. The algorithm estimates a model at each quarter with available information up to that point, evaluates the current vulnerability of each country, stores them, and at the end collects all probabilities and evaluates how well the model has performed out-of-sample. The authors extend the research to aggregation procedures and identify four different methods: *best-of* simply choose the most accurate model each time; *voting* decides the predicted outcomes based on the output of most of the models; *arithmetic* computes the means of predicted probabilities; and *weighted average* weights predicted probability with the Usefulness score obtained in-sample. The in-sample results show that all machine learning methods achieve impressive results, the best being KNN and SVM with AUC score of 0.988 and 0.998 respectively. In the recursive real-time estimations, results are still impressive with KNN and ANN achieving an AUC score of 0.979 and 0.969 respectively. Looking at aggregation of models, the authors conclude that the simultaneous use of many models yield in general good results, with weighted aggregate and non-weighted aggregate reaching the top spots with a AUROC score of 0.970 and 0.953 respectively. The authors argue that such good results depend on the necessity of using multiple modelling techniques in order to collect information of different types of vulnerabilities.

A paper which highlights the weaknesses of modern ML techniques is presented by **Beutel et al. (2018)**, in which a benchmark Logit model is compared against several machine learning approaches. They use a comprehensive quarterly dataset encompassing systemic banking crises for 15 advanced economies in the period 1970-2016 covering a total of 22 banking crises, retrieved from multiple sources including Laeven and Valencia (2013), Babecký et al. (2014), and Reinhart and Rogoff (2008). The models are trained to predict a financial crisis starting between the next 5 to 12 quarters conditional on not already being in an acute crisis period. Explanatory variables are comprised in 4 categories: asset prices, credit developments, macroeconomic environment, and global factors. The authors stress that data availability is a real issue, and that some plausible predictors had to be dropped due to a lack of long enough series for all countries. Anyway, through later robustness analysis they state that “for the purpose of comparing different prediction methods having a sufficient number of observations in the sample appears to outweigh the benefits of using a complete set of all potentially important early warning indicators”. Credit-to-GDP, real residential real-estate prices, as well

as other time series are transformed with the use of a one-sided HP filter. The ML methods employed are K-nearest Neighbors, Classification Trees, Random Forest, and Support Vector Machines. Hyperparameters are chosen in a cross-validation exercise, in which only information before the start of the out-of-sample window is used, using the relative Usefulness score as optimising criterion. Differently to Holopainen and Sarlin (2017), Beutel et al. do not use cross-validation to evaluate models as this could lead to serious over-estimation of the models' performance, as explained by Neunhoeffler and Sternberg (2018), in a political science paper regarding civil wars predictions. They instead use cross-validation only for hyperparameters selection and perform a classic out-of-sample prediction experiment. Predictions are then evaluated using relative Usefulness, AUC, and Brier probability score. Following Holopainen and Sarlin, Beutel et al. set the out-of-sample window between 2005 Q3 and 2016 Q4, splitting total data approximately in half, and performing the predictions in a recursive way, quarter-by-quarter. In-sample results unsurprisingly show the dominance of ML algorithms, with Random Forest achieving an AUC score of 0.999 and relative Usefulness of 0.990, against a Logit's score of 0.810 and 0.511 respectively. When looking at out-of-sample results, on the other hand, Logit outperforms almost always the ML methods. It obtains a relative Usefulness of 0.605 against the second-best Binary Trees with 0.126, and a ROC score of 0.852 against the second-best SVM with 0.629. Brier scores also confirms these findings. ML algorithms appear to perform better than Logit out-of-sample only on a limited subset of macroeconomic indicators. The authors explain these striking results with ML overfit on training data, and the sufficient flexibility of the Logit model. They also explain that previous results from Alessi and Detken (2018) indicates ML algorithms as best performers because they did not run an out-of-sample analysis, but a k-fold cross-validation one resulting in an out-of-sample score similar to those obtained in-sample. The same conclusions can be drawn regarding the work of Holopainen and Sarlin (2017) and their use of cross-validation out-of-sample.

**Ristolainen (2018)** focuses mainly on single-layer artificial neural networks. A monthly dataset assembled by Kaminsky (2006) from January 1970 to June 2003 including 18 countries, both developed and developing, is used to predict banking crises out-of-sample with a 24-month pre-crisis window. The dataset includes 14 indicator candidates, both regional and global, and describes a total of 32 separate banking crises. During tests, the 5 years post-crisis observations are removed to deal with post-crisis bias, and only data until three years before the start of a crisis is used, so to be able to adhere correctly to pure out-of-sample predictions. The ANN model is trained with 10-fold cross-validation on the training dataset so to choose the best hyperparameters, and then confronted with a traditional Logit Regression using ROC as the

scoring metrics. In-sample, ANN can fit the data almost perfectly as expected, while Logit obtains a still good but lower score. The out-of-sample results using the test set are based on the ability of the models to correctly classify the pre-crisis periods of eight specific events in eight different countries. The results are mixed, with the two models prevailing in different countries, with ANN achieving better out-of-sample results for five out of eight countries. Using a method proposed by Garson (1991) and Goh (1995) the author can estimate the relative importance of each indicator in the ANN. He finds that the most significant predictors are domestic credit, M2/reserves ratio, real GDP growth, inflation, and oil price.

Different recurrent neural networks are used in the work of **Tölö and Eero (2019)**, specifically RNN, RNN-LSTM, RNN-GRU and multilayer perceptron. Data is collected from Jordà, Schularick and Taylor (2015) and covers the financial crises and relevant annual macroeconomic data for 17 advanced economies in the period 1870-2016, for a total of 2499 observations. Observations during crisis periods and on the 5 following years are removed from the computation. The models are trained so to predict events with a horizon from 1 to 5 years using a country-by-country cross-validation approach. The main results are drawn using only 5 predictors with 5 lags each, even though 8 additional variables are used later in the sensitivity analysis. The main out-of-sample testing is done splitting the sample in two parts, with the earlier (until 2002) used for training and the latter (2003-2016) for testing so to preserve the temporal structure, and ranking the models based on ROC score. Overall, RNN-LSTM and RNN-GRU outperform the other methods, including a Logit Regression, by a “significant margin”. As an example, considering a forecast horizon of 2 years, RNN-LSTM achieves a ROC of 0.742 while Logit scores 0.521 in the sequential validation. The lowest position is that of the multilayer perceptron with a score of 0.431. The authors conclude stating that RNNs generally outperform simpler ANNs when dealing with time series and multiple lags because of their structure which efficiently counteracts overfitting.

**Bluwstein et al. (2023)** uses the same database from Jordà, Schularick and Taylor (2015) to predict financial crises one to two years in advance. The observations occurred during a crisis episode and on the following four years are excluded to avoid post-crisis bias. For the same reasons the authors decided to also remove all observations from the later years of the Great Depression, from 1933 to 1939, and world wars. After these corrections, a total of 1249 observations remains. The models tested are a benchmark Logistic regression, Decision Trees, Random Forests, Extremely Randomised Trees, SVM, and ANN. Ten different domestic economic indicators are considered, plus two more global variables. Out-of-sample performance is first evaluated with 5-fold cross-validation and then on a forecasting approach



using only data available until the given point in time, acknowledging that the cross-validation approach does not reflect the real-time performance of an early warning model as shown by Beutel et al. (2018). Hyperparameters are chosen with nested cross-validation out-of-sample, that is each training set of the 5-fold cross-validation is used in a further 5-fold cross-validation to assess the performance of all possible combinations of hyperparameters, and then the best setting is used to train the model on the complete training set. As in most papers, results are compared in the ROC space. In the cross-validation approach, machine learning algorithms have an advantage against Logit Regression, with Extreme Trees and Random Forest being the most accurate. These results are also confirmed by several additional robustness check exercises run by substituting or adding a small number of variables, or by changing the transformations applied to the time series. Only Decision Trees performed worse than Logit, due to its tendency to overfit based on the authors' opinion. The authors then run an out-of-sample forecasting experiment, finding that Logistic regression performed poorly again, and that Neural Networks was the model with the highest AUC score in predicting the period 2004-2016. Nonetheless, results were generally poorer than in the cross-validation exercise and the share of false alarms prompted by the models substantially higher. The authors proceed by examining the relative importance of the single predictors elaborated by the Extreme Tree model through Shapley values. The variables with the largest predictive shares are the global yield curve slope and global credit growth, consistently across the five models. Performing a Shapley regression, the authors are also able to determine the statistical significance of the predictors by regressing the crisis indicators on the Shapley values. Consistent with previous results, they find that global and domestic yield curve slope obtain the highest coefficients and lowest p-values. Predictors that surprisingly give a signal that cannot be statistically differentiated by the null are public debt, current account balance, and house prices.

**Fouliard et al. (2020)** experimented with several machine learning algorithms in a recursive “online” out-of-sample approach. They used quarterly data for seven advanced countries from 1985 Q1 to 2019 Q3. Systemic crises episodes are retrieved from the Official European Database by the ECB, and financial conditions and macroeconomic indicators are drawn from OECD, BIS, and Cross Border Capital databases, using, whenever possible, the vintage data available at the time of collection and detrending variables using only data of the estimation sample to avoid look-ahead bias. The authors focus on predicting systemic events from twelve quarters ahead, also considering the delayed feedback the forecaster receives: in time  $t$ , the forecaster cannot know if the observations in  $t-1$  belong to a pre-crisis period. The models implemented are General Additive Model (GAM), Random Forests, SVM, and several

Logit/Probit models with different combinations of variables. The main result is then drawn from an EWA (exponentially weighted average) aggregating rule. Results are evaluated by observing the forecasting capabilities on four countries, France, UK, Germany, and Italy, by observing if the signal is monotonically increasing before a crisis provided it does not have too many false alarms. For France, the dominating models picked by the EWA are GAM and a Logit combination, in UK it was again GAM and SVM, in Germany it's a dynamic Probit model, and for Italy a Logit model (even though results for Italy are not very good on average). The authors conclude by arguing that their model aggregation technique works, stating that the sample crises are all predictable ahead of time based on the observed probability increasing significantly and monotonously, even if they admit some heterogeneity across countries in terms of which models and variables forecast better. A further study conducted by the same authors, Fouliard et al. (2021) using the Jorda-Schularick-Taylor (2017) dataset reaches the same conclusions.

A study by **Casabianca et al. (2022)** analyses the macroeconomic determinants of banking distress using the AdaBoost algorithms with a dataset of over 100 countries retrieved from Laeven and Valencia (2008, 2013 and 2018), Reinhart and Rogoff (2008), and Jordà et al. (2017), both advanced and developing, over the period 1970-2017, for a total of 142 banking crisis. Selected predictors are both of domestic and global nature and are detrended with a 1-sided HP filter in most cases. The authors try to predict an event by defining the pre-crisis period as the three years preceding the event and removing from the pool of observations those drawn during active crisis periods and from the following three years so to avoid post-crisis bias. Dataset is split in two, with the 1970-2005 split used to train the models and to choose the hyperparameters so to optimise relative Usefulness in a 5-fold cross-validation, while the 2006-2017 split is used for out-of-sample testing and models evaluation. Furthermore, advanced and developing countries are analysed separately given the important differences in the average values of the predictors. In-sample fit is evaluated running a panel-block bootstrap in which the length of each block is five consecutive years. AdaBoost displays a good fit with better performance than Logit, even if both models suffer a decrease in performance when analysing developing countries. In the recursive out-of-sample test the AdaBoost model consistently outperforms the traditional Logit, which achieve performance that is barely better than a random guess, using as evaluation parameters AUROC, sensitivity, specificity, and relative Usefulness. To be more specific, for advanced economies, Logit achieves an AUROC score of 0.481 against 0.874 for AdaBoost, and 0.636 and 0.822 respectively for emerging economies. AdaBoost is able to predict probabilities of pre-crisis in advanced countries well, but on the other hand

probabilities fit poorly for emerging economies. The authors then carry out an analysis of the contribution of each individual predictor using the Shapley values. In advanced economies the most significant predictor was the US 10yr Treasury rate (implemented as a global variable) acting against the build-up of crisis, except for the years 2005-2007. A similar pattern of the effect of the 10yr rate is shown in emerging economies. Inflation stood out as the leading domestic indicator for both developed and developing countries.

From the general literature about the argument, it is possible to draw the conclusion that machine learning and neural networks have the potential to bring improvements in this research field. Most of the cited researchers state that these algorithms have a great potential in crisis prediction. In particular, these more sophisticated techniques show to have an impressive in-sample ability to fit data. However, when testing on a true real-time out-of-sample approach, which exclude the use of cross-validation out-of-sample since it uses data that would not have been available at the time of the real-time testing, sophisticated methods show a general decrease in performance and less enthusiastic forecasting capabilities, with a significant variability based on the specific method, dataset and evaluation metrics implemented. In the mentioned papers, of those who carried out a true real-time out-of-sample analysis, it is interesting to highlight the results of Fricke (2017) and Beutel et al. (2018), since these works argue that the traditional econometric methods still preserve the role of best forecasters given their ability to generalise and their lower tendency to overfit data. Another observation could be done about the use of periods immediately preceding a crisis in the training dataset, for only a minority of examined dissertations excludes these from out-of-sample analysis. In a proper real-time application those could not have been available since it would have been too early to correctly determine if they were classifiable as pre-crisis or tranquil period. The author of this Thesis conducts the analysis on data following a true recursive out-of-sample strategy, so to mimic an actual prediction of future unknown outcomes, since the main purpose of this field of research is to provide policymakers with an effective tool to predict banking woes. Even if testing the models with an in-sample approach to show the algorithms' impressive properties is tempting, the author of this dissertation argues that evaluating models with these strategies would not guarantee the replicability of results in a real-world implementation.

# PREDICTORS

Following the results of the examined literature, the author collected yearly data about the economic variables that seem to best predict upcoming financial distress. It is important to notice that, given the large base of countries included in the study and the relative long time-series, data collection is a real issue. By choosing a large number of economic variables, the total number of complete information could shrink significantly due to data availability, posing to the author a trade-off between the number of variables and the number of observations to be included in the study. Following the argument proposed by Beutel et al. (2018), the author will implement only predictors that showed their relevance in previous research and for which data is available in a manner that does not truncate substantially the number of observations, prioritising total observations quantity. An example is given by domestic interest rates, which despite being an important variable, are available only for a limited panel of advanced countries. It is important to notice that including 204 countries and 50 years one would expect a total number of observations of 10.200. However, obtaining this number is unfeasible due to certain circumstances. For instance, within the 50-year timeframe considered, some countries have not existed (most notably the Soviet Block). Additionally, certain micro-nations (e.g., British Virgin Islands) and nations enduring persistent unrest or experiencing unreliable data collection (e.g., South Sudan) lack essential data such as real GDP, leading to their exclusion from the analysis. Missing data is also an issue when computing the trends for de-trending a time series. The literature makes extensive use of the HP filter given its ideal properties for de-trending economic data, but this method requires a long time series without any missing point. However, data collected from the author presents some discontinuity which makes the application of HP filter inconvenient. The same problem was faced by Ristolainen (2018), who chose to not apply the filter. In this dissertation, time series needing such treatment will be instead de-trended simply by subtracting to the observation the average of the last 5 years. If in the last 5 years an observation is missing, the average will be simply computed on the available values. De-trending allows to make data stationary, which might help regression analysis, improves comparability, and makes so that the long-term trend does not obscure short-term fluctuations. Other transformations applied to the time series include winsorizing, which consists in substituting the values above the 99<sup>th</sup> percentile and below the 1<sup>st</sup> percentile with the values corresponding to the 99<sup>th</sup> percentile and the 1<sup>st</sup> percentile respectively. This transformation is applied to variables that may present outliers in rare instances, such as inflation, and avoid the adverse effects of extreme values in the prediction capabilities of some machine learning algorithms. Winsorizing allows for the adjustment of these extreme values without completely

removing them from the dataset, thus preserving the information they carry while decreasing their distorting influence on statistical analyses. A last transformation is standardisation, which is achieved by subtracting from an observation the mean of all previous observations (from that and every other country), and then dividing by the standard deviation. By doing so, the time series has an average of 0, and a standard deviation of 1. The main advantage brought by standardisation is that it enables easier comparison and analysis. It helps in interpreting the relative importance of different variables in the dataset by eliminating the influence of scale differences, and when performing regression analysis or similar techniques, standardisation also helps in interpreting the coefficients. It also helps the computational processes in some machine learning algorithms, resulting in reduced computing time. Standardised coefficients indicate the importance of predictors relative to each other, since they show the change in the dependent variable in terms of standard deviations for a one standard deviation change in the independent variable. In later experiments, all these transformations are done with a backward-looking approach, using only data that would have been available at the time, so to avoid any bias and stick to the real-time approach.

## **Domestic variables**

The nine domestic predictors, the sources, and the transformations applied in this Thesis are reported in this chapter.

- **Inflation**

High inflation is usually associated with macroeconomic mismanagement. Since data on interest rates are not available in an extensive way across countries and time, inflation also serves the role as a proxy since it is likely to be correlated with high nominal interest rates, as explained by Demirgüç-Kunt and Detragiache (1998). Inflation also decreases the real return on assets and discourages saving while inducing more borrowing, as stated by Duttagupta and Cashin (2011). A further explanation of the effect of high inflation is given by Jiang (2008): “higher inflation reduces the return to domestic asset (currency plus capital) inducing the bank to invest more on dollars and less on capital, which makes it less likely that the residual claimants prefer keeping bank deposits instead of cashing out and increases the likelihood of banking crises”. In this Thesis, inflation data is proxied by the GDP deflator, which is retrieved by the World Bank (WB) database. The time series is then winsorised and standardised. In this study, figures report GDP deflator as ‘*GDPdefl*’.

- Nominal depreciation

Kaminsky and Reinhart (1999) show that banking crises have often occurred at the same time, or immediately after, currency crises. Banks may be largely exposed to foreign exchange risk, affecting the value of banks' assets and liabilities in foreign currency, so that a strong depreciation of the exchange rate might imply a situation of upcoming banking distress. As explained by du Plessis (2022): "Whereas a sharp currency depreciation reverses capital flows and reduces asset values, higher import inflation and a weakening in the terms of trade could increase cost-push price growth, cause an outflow of working capital, a contraction in domestic liquidity, and, in turn, reduce the ability to settle debt obligations". In this dissertation, nominal depreciation is computed from the exchange rates drawn from the databases of Organization for economic cooperation and development (OECD), WB, and Federal Reserve Economic Data (FRED). Yearly exchange rates are expressed as the period-average national currency amount needed to buy one USD. Nominal depreciation is computed as:

$$NomDepr = (Fx(t0) / Fx(t-1) - 1) * 100$$

So that an increase in the exchange rate (national currency is losing value) leads to a positive value of nominal depreciation. The predictor is then winsorised and standardised and represented in future tables as '*NomDepr*'.

- M3 / Reserves

In the literature, some authors such as Ristolainen (2018) attributes a significant importance to the M2 to reserves ratio indicator in predicting financial crises. However, this Thesis is based on a much wider scope timewise and country-wise, and M2 indicators for a substantial number of countries for a long enough period are not available. To solve this issue, the author used M3 as a proxy for money supply, given the availability of this indicator in the WB database. M3 has also been used as a predictor by Alessi et al. (2015). M3 data series are sourced from the International Monetary Funds (IMF) and from International Financial Statistics (IFS) and are defined as "Absolute value of liquid liabilities in 2010 US million dollars. Liquid liabilities are also known as broad money, or M3. They are the sum of currency and deposits in the central bank (M0), plus transferable deposits and electronic currency (M1), plus time and savings deposits, foreign currency transferable deposits, certificates of deposit, and securities repurchase agreements (M2), plus travellers' checks, foreign currency time deposits, commercial paper, and shares of mutual funds or market funds held by residents". The denominator, national reserves, is drawn from IMF and IFS databases

and is defined as “holdings of monetary gold, special drawing rights, reserves of IMF members held by the IMF, and holdings of foreign exchange under the control of monetary authorities. The gold component of these reserves is valued at year-end (December 31) London prices. Data are in current U.S. dollars”. The M3/reserves predictor could contribute to the formation of a financial crisis in two paths since it could indicate a monetary expansion and/or the depletion of national reserves. In the following exercises, this time-series is winsorised and standardised, and reported as ‘*M3/Res*’.

- Real GDP growth

During periods of robust economic growth, businesses and individuals tend to experience higher incomes, increased profitability, and improved creditworthiness. Banks tend to see higher demand for loans, increased deposit levels, and improved profitability during periods of economic expansion, and, as explained by Demirgüç-Kunt and Detragiache (1998) and Reinhart and Rogoff (2009), higher GDP growth affects the bank’s share of NPLs. This improved financial health of banks reduces the likelihood of systemic banking crises. However, it's important to note that sustained high growth can also create potential risks, such as overheating, asset bubbles, or excessive risk-taking, which if left unchecked might contribute to future banking crises. Real GDP growth time series are retrieved from WB, FRED, and Jordà et al. (2017), and the WB database reports it as “annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2015 prices, expressed in U.S. dollars”. In the following exercises, these time series are winsorised, standardised, and reported as ‘*rGDPgr*’.

- Public debt

Excessive public debt levels increase a country's vulnerability to sovereign risk. If investors perceive the government's ability to repay its debt as questionable, it can lead to higher sovereign borrowing costs, credit rating downgrades, and capital outflows. Sovereign debt crises can spill over to the banking sector, causing financial instability and banking crises due to exposure to government bonds or because of broader economic shocks. High debt levels can also strain fiscal policies, leading to austerity measures, reduced government spending or tax increases, which could dampen economic growth, increase unemployment, and impair borrowers' ability to service debts, potentially leading to increased NPLs in the banking sector. Excessive government debt could also postpone measures to strengthen bank balance sheets, according to Lindgren, Garcia, and Saal (1996), since the fiscal situation could be “too

weak” to allow for any consideration of banking problems by supervisors. Public debt may also be a proxy for countries’ vulnerability to solvency and liquidity shocks as in Casabianca et al. (2022). Public debt computed as a ratio of GDP is drawn from multiple sources, including IMF, Jordà et al. (2017), WB, FRED, Statista, and CEIC and is reported typically as on the last day of fiscal year. Time series are then standardised and reported as ‘*GovDebt*’.

- Investments

Investment growth is closely linked to economic cycles. During periods of robust investment, economies tend to expand, driving growth and prosperity. Investments often require external financing, and if a country relies heavily on foreign capital to finance its investment projects, sudden reversals or disruptions in capital inflows can increase vulnerabilities in the banking sector. According to Beutel et al. (2018): “high levels of investments in fixed capital may be driven by overly optimistic expectations, leading to problems when future returns are lower than expected”, so that when investment growth slows or contracts sharply, it can lead to economic downturns, rising unemployment, and decreased consumer and business spending. Such economic volatility can strain banks’ balance sheets, impacting their profitability and potentially leading to banking crises. This is particularly true in the case of a rapid investments growth since it might lead to excessive lending and inadequate capital buffers. Investments data series are retrieved from WB and defined as gross fixed capital formation: “Gross fixed capital formation includes land improvements; plant, machinery, and equipment purchases; and the construction of roads, railways, and the like, including schools, offices, hospitals, private residential dwellings, and commercial and industrial buildings”. It is expressed as a ratio of GDP, de-trended, standardised, and reported with the acronym ‘*GFCF*’.

- Current Account Balance

Current Account Balance (CAB) has widely been used as a proxy for external vulnerabilities. A sustained deficit implies reliance on external financing, such as borrowing or capital inflows, to fund the shortfall. High and persistent deficits may lead to increased external debt levels, making the country more vulnerable to sudden changes in investor sentiment or disruptions in capital flows, which can impact the banking sector's stability. As argued by Kaminsky and Reinhart (1999), large capital inflows from abroad may support asset price booms and induce a reversal in asset prices when these inflows decline or stop. In this dissertation, the CAB time series are retrieved from WB, IMF, and OECD databases. WB defines CAB as “the sum of net exports of goods and services, net primary income, and net secondary income” and expresses it as a GDP



ratio. The author then standardised the data series. Current account balance is expressed in later tables as ‘*CAB*’.

- Loan-to-deposit ratio

A high bank credit to deposit ratio indicates that banks are extending more loans relative to the deposits they hold. This might indicate aggressive lending practices, where banks extend a substantial amount of credit without adequate deposits as a funding base. This aggressive lending behaviour can increase banks' risk exposure, leading to higher levels of NPLs. If many banks are heavily leveraged and face liquidity or solvency challenges simultaneously, it can propagate financial instability and create systemic risks that have the potential to lead to broader banking crises. The denominator of this ratio, bank deposits, is also suggested as an indicator of impending crises by Kaminsky and Reinhart (1999). The time series used by the author is drawn from IMF. IMF defines loan-to-deposit ratio as “The financial resources provided to the private sector by domestic money banks as a share of total deposits. Domestic money banks comprise commercial banks and other financial institutions that accept transferable deposits, such as demand deposits”. The series is winsorised and standardised before analysis and labelled ‘*LoanDep*’.

- Credit growth

The literature is rich in studies regarding the effects of credit growth on financial crises. The most frequent explanation is that high bank credit growth may fuel asset price bubbles, such as in real estate or stock markets, as increased lending provides more liquidity for investment. If these asset bubbles burst due to changes in market conditions or economic shocks, it can lead to a sharp reversal in asset values, as showed by Jordà et al. (2015) and by Schularick and Taylor (2012). Banks heavily exposed to these assets face significant losses, potentially causing distress in the banking sector as a whole and triggering financial instability (Borio and Lowe, 2002). This is confirmed by Eichengreen and Rose (1998) who state that expansionary monetary and fiscal policies fuel a lending boom, and that eventually monetary policy must be tightened to contain inflation, pricking the bubble. Borrowers are unable to repay, forcing banks to curtail their lending, further depressing property, and securities markets. In later exercises, credit growth is expressed as a ratio of GDP. It is retrieved from IFS and IMF and defined as “Private credit by deposit money banks and other financial institutions to GDP”. The author de-trended, winsorised and standardised the time series and labelled it ‘*Credgr*’.

## Global variables

Other than domestic variables, this dissertation includes three global predictors, equal for each country in the dataset, following a frequent practice in the literature.

- Global real GDP growth

Global GDP is often used in the literature as a predictor for financial crises. During periods of robust global economic growth, countries tend to experience increased trade, higher incomes, and improved economic conditions. This can positively impact borrowers' ability to repay loans, leading to lower default rates and better credit quality in banks' loan portfolios, reducing the likelihood of banking crises. On the other hand, during economic expansion, positive investor sentiment and market optimism prevail, reducing perceived risks in banking systems. The source of dataset is WB, and values are expressed as year-on-year percentage change. Standardisation is applied to the time series, with the variable being labelled '*glGDPgr*' in later tables.

- US 10y treasury yield

This variable is often credited in the literature to be the most significant global predictor of financial distress. In fact, higher yields in the US might attract capital from other countries as investors seek higher returns. This can lead to capital outflows from other countries, potentially causing currency depreciations, reducing liquidity, and impacting the stability of banking sectors in those nations. Other than this, it is important to consider that many companies and countries, especially developing ones, borrow in international markets. Variations in US yields could significantly impact the borrowing costs for these entities straining debt sustainability. This pattern is explained by Calvo, Leiderman and Reinhart (1993), who show that the highest international financing costs incurred by banks are passed to domestic borrowers, implying eventual repaying problems. The foreign effects of higher US treasury rates are similarly analysed also by Iacovello and Navarro (2019) who highlight the impact in vulnerable emerging economies. The author retrieved the US 10 years treasury yield from the European Central Bank (ECB) website, expressed as average of the period. Standardisation is applied to this variable, which is labelled as '*gl10y*'.

- Global yield curve slope

Yield curve slope is measured by the spread between 2-year and 10-year US Treasury yields and is often cited as one of the main predictors of future economic conditions, since it reflects market's expectations about interest rates (Estrella and Hardouvelis, 1991 and Wright, 2006). It is particularly important for the banking sector, given that

banks profit from maturity transformation. Banks typically borrow at short-term rates and lend at long-term rates, implying positive net interest margins when the yield slope is positive (long-term rates are higher than short-term rates). However, a flattening or inverted yield curve can compress margins, potentially impacting banks' profitability and weakening their ability to absorb losses, which can contribute to banking crises as shown by Joy et al. (2017). An inverted yield curve slope could also signal negative market sentiment, which can lead to volatility and a risk-averse environment. In this paper's exercises, the yield curve is computed as:

$$Yield\ curve\ slope = \log \left( \frac{10y\ rate}{2y\ rate} \right)$$

which implies that a negative value indicates an inverted yield curve. Historical US interest rates are drawn from the Federal Reserve (FED) and ECB websites expressed as average of the period. Standardisation is then applied to the global variable 'gLYC'.

## Signals

Now that the predictors have been defined, a signal approach is used to analyse the fluctuations of the average value of each domestic predictor in the preceding years of a crisis event and in the years immediately following. The average value of the predictors, including both crisis and tranquil periods, is also considered for comparison. Using the whole panel of countries, the results are as in *Figure A.1* in the Appendix. Anyways, it might be more interesting to observe the same signals when developing and advanced economies are separated, to observe the different behaviour of the variables before and during a financial crisis. In *Figure A.2* in the Appendix, the domestic predictors of developing economies are shown. Results are somewhat similar to the overall database, probably given that the majority of observations in the database come from developing countries. For developing economies, the most noticeable take-away is that the three years preceding a crisis are characterised by rising credit, both in terms of credit growth and loans to deposits ratio, and by a sharp upward trend of investments, all of them above the average values. Another interesting point is the decline in the current account balance and the nominal depreciation rate from high values to a point close to the average value, and the drop in inflation from higher than usual values in the year immediately preceding the event. All cited predictors then appear to take a sharp turn to the opposite direction once the crisis has burst. All these observations are coherent to the literature. Regarding advanced economies, results are shown in *Figure A.3* in the Appendix. As before, an increase in credit and investments

is clearly noticeable, with the difference that the ratio of loans to deposits does not seem to be particularly affected, even if it is stable above the mean line. It is interesting to observe the depreciation, which increases until the year before the crisis, when a sharp decline (meaning an appreciation of the national currency) happens, and the slow but steady decline of inflation from above to below average values. An upwards trend in the M3 to reserves ratio up to the year preceding the crisis is visible, as well as the slow but constant decline of current account balance well below average values. In both developing and developed countries, the crisis provokes a decrease in GDP growth below mean rate and pushes Government debt upwards, most probably due to the recovery expenses incurred and the decrease in GDP which acts as the denominator. Both these adverse effects are most noticeable in developing countries. Looking at the global indicators, for developing economies, it is interesting to notice that the US 10-year Treasury rate is high but decreases slowly and constantly, and that the yield curve slope is stable but with very low values, implying a flatter than usual curve, which later grows during the crisis. Regarding advanced countries, the yield curve decreases until the point of being almost flat two years prior to the event, and then increases rapidly. The US 10-year rate shows slightly above average values up until the crisis year, after which it moves to the mean line. In all cases, global GDP growth is quite stable before the event, and drops significantly below average in the year after the crisis, indicating that banking distress often strikes many countries simultaneously. In the Appendix are also present the correlation matrices for both advanced and developing countries for all predictors. Given the different dynamics occurring in developing and advanced economies and the different crises' nature they will face in the test sets, these two categories will be analysed both together and separately in the following baseline experiment proposed by the author of the Thesis to observe the change in the models' performance.

# BASELINE EXERCISE

Now that the models' characteristics, evaluation methods and variables have been clarified, an explanation for the baseline exercise run by the author is given. First, to avoid crisis-bias, every observation recorded in a country during a crisis is removed from the computation as it is standard in the literature. Other authors from the literature also removed three to five years after the crisis so to avoid post-crisis bias. Instead, the author of this Thesis did not remove post-crisis years but removed the observations from the whole duration of the crisis from the analysis. In the baseline experiment, the crisis duration is chosen according to the indications retrieved from the authors of the databases implemented from Laeven and Valencia (2020) and Reinhart and Rogoff (2008), while for events pointed out exclusively by Jordà et al. (2017) a duration of three years is assumed by the author of this dissertation. This implies that the number of observations removed depends on the crisis duration, while in the literature it is usual to remove a fixed number of years/quarters for all events.

The objective is to determine whether any year is to be classified as a pre-crisis or tranquil period, and to do so a year is defined as pre-crisis if in any of the following three years a crisis event is present. In other words, to each observation is attached a Boolean variable, equal to 1 if at any point in the following three years a banking crisis erupts in that specific country, and 0 otherwise. Attached at each observation at  $t0$ , there are also the values of each variable at time  $t-1$ ,  $t-2$ , and  $t-3$ , representing the lagged variables. Even if these lags are not implemented in the baseline experiment, they will be useful for later experimentations. In these later exercises, when an observation is missing data for a lag, or when that lagged data is recorded during a crisis, the observation must be dropped from analysis completely, at the cost of a reduced database length. Following this set of rules, for the baseline experiment a total of 3576 observations are collected, 2284 for emerging economies and 1292 for advanced ones. Of these observations, 330 are accounted for as pre-crisis period, 203 in developing economies and 127 in advanced ones.

The focus of this dissertation is a recursive real-time analysis and is achieved by proceeding with the following approach. The models are firstly trained on the set of data from 1970 until 1995, using the Python programming language and various scikit-learn libraries. This portion of data will serve as the first training set for each model, including the choice of the hyperparameters using 5-fold cross-validation. Next, the trained models are used to make a prediction whether the year 1998 is a pre-crisis period for each country, or in other words, if any of the years 1999, 2000, or 2001 will be characterised by a banking crisis, using the

predictors' values collected relative to the year 1998. To make the first prediction, contrary to what other authors in the literature have done, the author of this Thesis trained the models using data only up until 1995 and not until 1998 because in a real-world framework it would be possible to assign the label 'pre-crisis' or 'tranquil period' to the observations from the year 1998 only at the end of 2001. Placing ourselves in the role of a forecaster at the end of 1998, our models could be trained only referencing years that we know for sure being pre-crisis or tranquil periods, so that the last serviceable year for training is 1995. Data from 1996, 1997 and 1998 could be implemented for training purposes only at the end of 1999, 2000, and 2001 respectively. As explained by Fouliard et al. (2020), the feedback of the forecaster is delayed. Doing otherwise would probably bias upwards the capability of the tested predictive models.

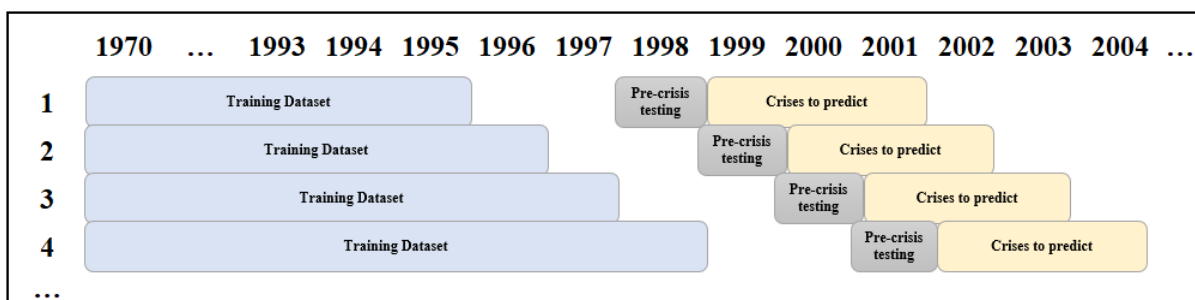


Figure 6.1: Recursive testing. Author's own elaboration

Every model is built so to have as output a continuous value comprised between 0 and 1, corresponding to the predicted probability that the year which data it was fed (1998 as first) is in a pre-crisis state. These outputs are used to fill-in a vector, while another vector of equal length is filled-in with the real values (0 or 1) that are unknown to the forecaster at the time of training. The experiment then proceeds with the same approach by analysing the following year, that is 1999, with the training set drawn from the observations up until 1996. This procedure is then repeated for each following year, with the last year analysed being 2016 (even if predictors data is collected until 2022, the database containing financial crises events is updated only until 2019). The vectors containing the prediction values and actual observations are extended with each analysed year, so that at the end they contain the true observations from 1998 to 2016, accompanied by the predicted values from 1998 to 2016 computed in a real-time recursive way. At the end, these two vectors are used to compute multiple different comparing score: AUROC, Relative Usefulness, and Brier Score, other than scores retrieved from the Confusion matrix. All cited models are tested and compared in the baseline experiment, except RNN given that this model is particularly well suited for dealing with many lags, therefore it will be the subject of a separate chapter. The initial training period 1970-1995 was chosen by the author since it contains more than half of crisis in the complete dataset, and roughly one third of total complete observations. This gives an adequate initial training dataset to be used by the algorithms in the

first prediction. The author thinks that this approach allows for an honest evaluation of models' performance in a simulated real-world recursive implementation lasting 19 years, starting from 1998 until 2016, but presents a major drawback. Due to the temporal distribution of financial crises and data availability, the testing period contains a far smaller density of pre-crisis periods compared to the testing portion of data, which might impair the models' performance. Moreover, the crises present in the testing dataset are in great part represented by the Great Financial Crisis of 2007-2008, which impacted almost exclusively advanced countries and had a common origin, making the crisis episodes in the test dataset very homogeneous. An ideal test dataset would include many different crisis episodes around the world with different causes and dynamics so to observe the different reactions of the models in every situation.

In the baseline experiment, developing and advanced economies are analysed both together and separately, and every variable is included without any lag, so that every predictor is assigned the value recorded at  $t0$ . Shapley values are also computed and recorded at any year for each variable and compared to the Logit regression coefficients. In *Table 6.2* are reported the different performance scores relative to the recursive testing on the complete dataset.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,7405	0,0518	0,368	0,09	0,50	0,08	0,88	0,49	0,14
BCT	0,5321	0,0678	0,090	0,10	0,81	0,07	0,25	0,83	0,11
RF	0,5836	0,0473	0,140	0,19	0,88	0,11	0,23	0,91	0,15
AdaBoost	0,5624	0,1660	0,129	0,34	0,34	0,05	0,81	0,32	0,10
KNN	0,6584	0,0517	0,272	0,08	0,51	0,07	0,77	0,50	0,12
SVM	0,5894	0,0502	0,168	0,13	0,54	0,06	0,64	0,53	0,11
MLP	0,6695	0,0505	0,366	0,14	0,61	0,08	0,76	0,60	0,15

*Table 6.2: Performance score, baseline experiment, complete database*

As expressed before, the main evaluation parameter of this dissertation is AUROC, given its objectivity and lack of a specific threshold to be set by the author.

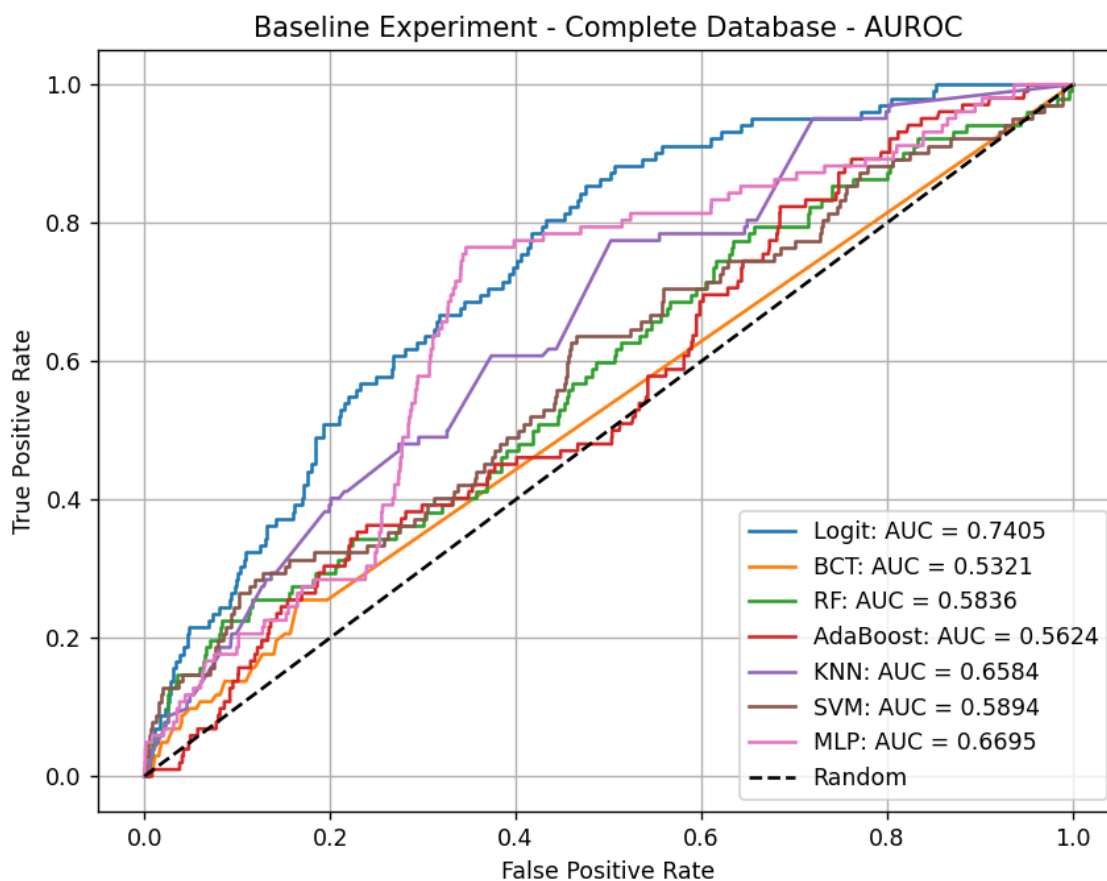


Figure 6.3: AUROC, baseline experiment, complete database

As shown in Figure 6.3, every model performed better than a random guess, even though for most models the difference from a random choice is not particularly significant. The best model is the Logit Regression, and the worst is BCT. Anyway, it is important to notice the low value of Precision (the share of correctly predicted crises over total predicted crises) for each model. This indicates that these models are not able to correctly predict a crisis when it is coming, and that they obtain a high Usefulness value simply by setting a low threshold and assigning almost all observations to the ‘tranquil period’ category, exploiting the fact that crises are rare event so to minimise the Loss function. It is also important to note that the traditional Logit Regression performed better than the average ML model in many evaluation metrics, confirming the hypothesis of Beutel et al. (2018): “It turns out that the logit approach is surprisingly hard to beat, generally leading to lower out-of-sample prediction errors than the machine learning methods”.

The Shapley Values obtained recursively with each analysed year are reported in the Appendix (Figure A.9), as well as the coefficients of the Logit model (Figure A.8).

Looking at the Logit coefficients, they are mostly of the expected sign throughout the whole testing period. Inflation, credit growth, nominal depreciation, M3/reserves, investments, and loans/deposits, all have positive signs, meaning that they positively contribute towards the



build-up of financial distress. This confirms previous findings in the literature, according to which an environment of great credit expansion could often be followed by credit crunches and market freezes, as explained by Boissay et al. (2016) in their article “Booms and Banking Crises”. In their model, they explain that favourable supply shocks initially lead to credit expansion and economic growth. Banks compete to satisfy credit demand by competing on interest rates and conditions, but the situation may reversal if doubts about loans quality, the interbank market, and economic prospects arise. At this point, the likelihood of returning to average productivity rises, slowing corporate demand and inducing a household saving excess. Stating the authors: “the larger the credit boom relative to the possibilities for productive use of loans, the larger the fall in interest rates, and the higher the probability of a bank run in —and therefore of a disastrous freeze of— the interbank market”. On the other hand, Gorton, and Ordoñez (2016) using their model in “Good Booms, Bad Booms” argue that not all credit booms are destined to turn into a financial crisis, but that if at the end of the credit expansion the decrease in productivity is avoided, for example thanks to technological improvements, the crisis may be averted.

High inflation and rising M3/deposits ratio might proxy for a loose monetary policy by the central bank. A monetary expansion can be achieved in different ways, for example by reducing the reserve requirements for commercial banks or by lowering interest rates, causing both money supply and inflation to increase. This could lead to a credit boom and bust cycle as explained above. To this regard, Grimm et al. (2023) stated that an over accommodative monetary policy protracted over an extended period increases the likelihood of a financial crisis considerably for the medium-term, mostly through credit creation and asset price overheating. These findings are not new to the literature and find feedback way back to Wicksell (1898), who hypothesised back in the 19<sup>th</sup> century the effects of low interest rates over a long period on house prices and the boom-bust cycle. Dell’Ariccia et al. (2017), using US data as reference, also find a negative correlation between ex-ante risk taking by banks and increases in short-term policy interest rates, due to increased leverage and search-for-yield behaviour.

Moving forward, the current account balance shows negative coefficients’ sign, as expected. A strong current account deficit implies an increasing indebtedness toward foreign entities and reliance on capital inflows from abroad, which may eventually raise doubts about external solvency. The macroeconomic mechanisms through which a worsening current account balance could turn into a financial crisis are manyfold. It could be a “sudden stop”, as in Calvo and Reinhart (2000), causing a serious depletion of national reserves and GDP loss. In the framework of Obstfeld (2012), current account deficits could also deteriorate the net

international investment position (NIIP) putting pressure on external debt solvency or implying unsustainable macroeconomic imbalances which will be corrected with a painful reversal of the current account. This last point is also clearly visible in the reversal of CAB values as crises erupt in *Figures A.1, A.2 and A.3* in the Appendix. However, it is interesting to notice how both current account balance and GDP growth lose significance after 2008, with their values approaching zero, possibly indicating that the Great Financial Crisis brought a shift in paradigm for the predictors' behaviour in the years preceding crises.

Government debt has a counterintuitive negative sign, meaning that a highly indebted state has lower chances of incurring in banking woes. This could be misleading though, since it might just indicate that advanced economies, which have usually a higher public debt to GDP ratio, are also the most resilient to financial crises. In fact, looking at the same coefficient computed on the high-income only database, its value fluctuates around zero implying low significance of this predictor.

Nominal depreciation could stress banks' balance sheets when a large share of their liabilities is denominated in foreign currencies, which is often the case in developing economies. As found by Aghion et al. (2000), firms are also negatively affected by domestic currency depreciation by increased difficulties in repaying bonds denominated in foreign currencies. This reduces firms' profit, output, and investments, further reducing demand for the national currency in a vicious cycle.

Looking at global variables, contrary to what expected, world GDP growth contributes to crises formation (even if its behaviour is quite erratic). Global yield curve and US 10-year rates have the expected signs, negative and positive respectively, but only starting from the years preceding the GFC. Before that, they had the opposite signs indicating that a steep yield curve and low interest rates in international markets lead to financial distress, opposed to what pointed out by the literature such as in Calvo et al. (1993) and Wright (2006). The yield curve reflects markets' expectation about the future, with a flat or inverted curve indicating that markets expect an upcoming interest rate cut and deteriorating economic prospects. The literature on the effects of international interest rates, especially in developing economies, is wide. For example, Arteta et al. (2022) argues that increasing domestic rates in the US could spill over to low-income countries increasing borrowing cost and the debt burden. Financial markets anticipate this and discourage capital inflows. Higher interest rates in the US also strengthen the dollar exchange rate, which implies nominal depreciation of other currencies.

According to the Logit Regression values, the most impactful variables in causing crises (based on coefficients' values) are GDP deflator and loans/deposits ratio, while the variables which mitigate risks the most are government debt (which is misleading, as previously explained) and the current account surplus.

Shapley values result to be less readable and coherent compared to Logit coefficients. Variables such as current account balance, nominal depreciation, GDP growth, gross fixed capital formation, and loans to deposits fluctuate greatly around the zero line, meaning that a clear effect over time on crises formation cannot be defined. It is also important to notice that Shapley values differ substantially among models, which sometimes point to opposite results, making so that the average line is close to zero. On the other hand, it's nonetheless clear that public debt and credit growth contribute positively toward crises formation before the GFC, as expected, but also that models produce negative coefficients for the 10-year rate variable, which strangely contradicts the Logit Regression and the consensus of the literature. In this regard a critique could be moved to the advanced machine learning algorithms, as some other authors have already done, saying that they act as 'black boxes' which makes difficult to draw conclusions about the mechanisms which led to the final predictions and the contribution of each variable.

The next step is to look at the predictions for some specific countries which experienced crises during the testing period. The predicted probabilities are plotted for four countries: USA, Italy, Dominican Republic, and Ukraine. These countries are chosen so to include different typologies of crises and different income level economies. USA and Italy were both hit by the GFC, which generated in USA in 2007 and then spread to Italy the following year. The crisis in Dominican Republic was caused by the abrupt stop to a period of strong economic growth, a lack of confidence in the financial institutions and by a botched bailout attempt (Hanke, 2004). Ukraine faced two crises in the testing period, one in 2008 as a direct consequence of the GFC and decreasing steel prices (Korduban, 2008), the second in 2014 caused by the annexation of Crimea by Russia and the loss of its major trading partner. This second crisis is used as benchmark, assuming that it is impossible for any model to predict such event and checking predicted probabilities in the previous years. In the below graphs, the blank spaces represent the crises years (which are removed from analysis to avoid crisis-bias).

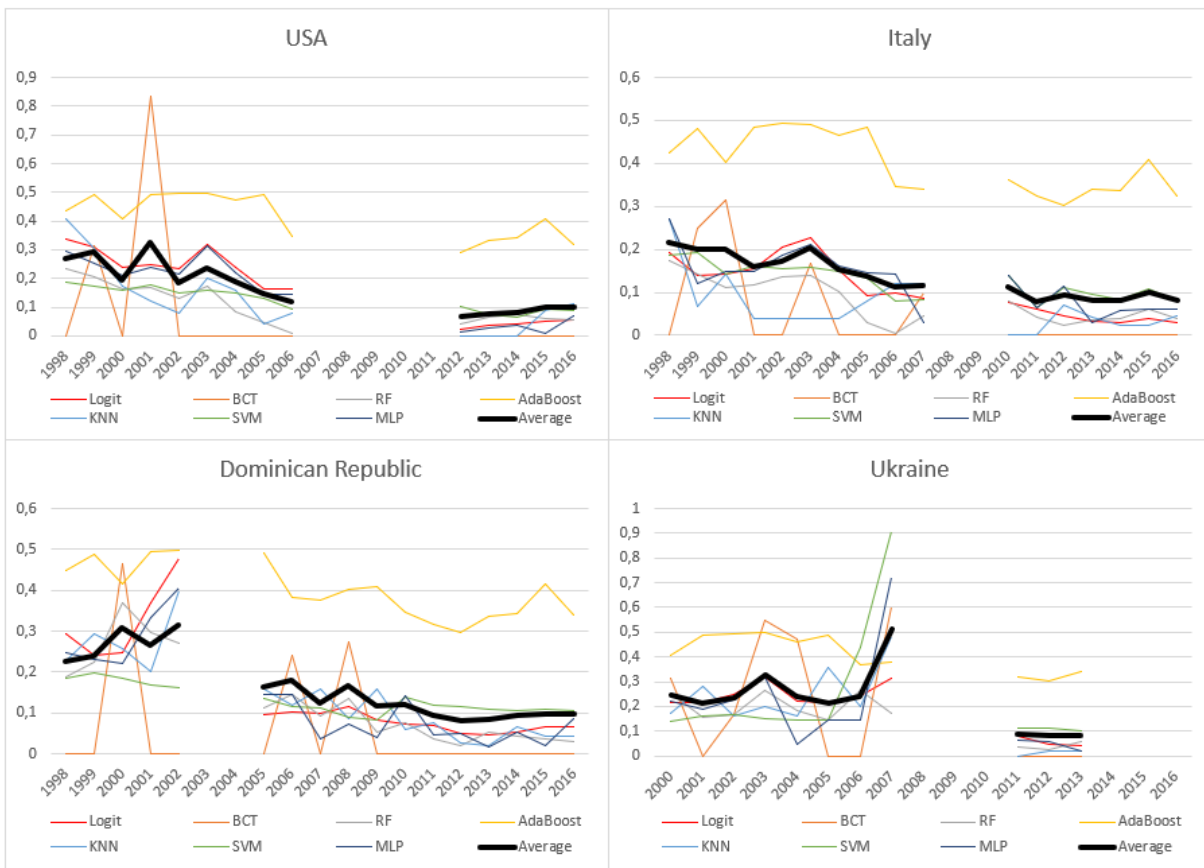


Figure 6.4: Predicted crisis probabilities by country, complete database, recursive approach

Looking at USA and Italy, in the three years preceding the financial crisis, predicted probabilities are close to the thresholds' values, but the indicators are showing a steadily decreasing probability of being in a pre-crisis state. A forecaster would have received warning signs in the previous years, only to find out that they were false alarms. The same forecaster, seeing the decreasing probabilities, would conclude that values close to thresholds in 2006-2007 are just another false alarm and that the financial situation is heading towards better prospects. This would prove catastrophically wrong in 2007 and 2008, showing how these models would have been unable to unequivocally predict that specific event in a real-world setting using the current framework. A completely different situation emerges both in Dominican Republic and Ukraine. In the Caribbean state the indicators correctly signal the rising risk of financial distress before 2003 and return to lower level once the crisis is over. In the 2000-2002 period, all indicators (except BCT) are above the warning thresholds. In Ukraine, contrary to Italy, in 2007 the indicators predict that the GFC would spread outside the USA. This is indicated especially by BCT, KNN, SVM, and MLP, which predicted probabilities shoot up in 2007. However, how it is easy to imagine, the models are unable to predict the 2014 Russian annexation of Crimea and its consequences, with all indicators below warning levels. Summing up, the number of models unequivocally signalling a crisis are only one in USA and

Italy (Logit and KNN respectively), and six both in Dominican Republic and Ukraine. No model can singlehandedly correctly predict all four crises.

It is interesting to observe how the performance change by splitting the complete database in the developing and advanced economies subsets. For developing economies, the following results are achieved.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,7698	0,0523	0,455	0,16	0,63	0,06	0,83	0,62	0,11
BCT	0,6514	0,0588	0,286	0,08	0,80	0,07	0,48	0,81	0,12
RF	0,7942	0,0343	0,517	0,14	0,75	0,08	0,76	0,75	0,15
AdaBoost	0,6729	0,1539	0,277	0,34	0,39	0,04	0,90	0,37	0,08
KNN	0,7844	0,0465	0,495	0,26	0,82	0,10	0,67	0,83	0,17
SVM	0,7521	0,0423	0,423	0,21	0,78	0,08	0,64	0,78	0,14
MLP	0,6326	0,0562	0,252	0,27	0,86	0,08	0,38	0,87	0,13

Table 6.5: Performance score, baseline experiment, developing economies

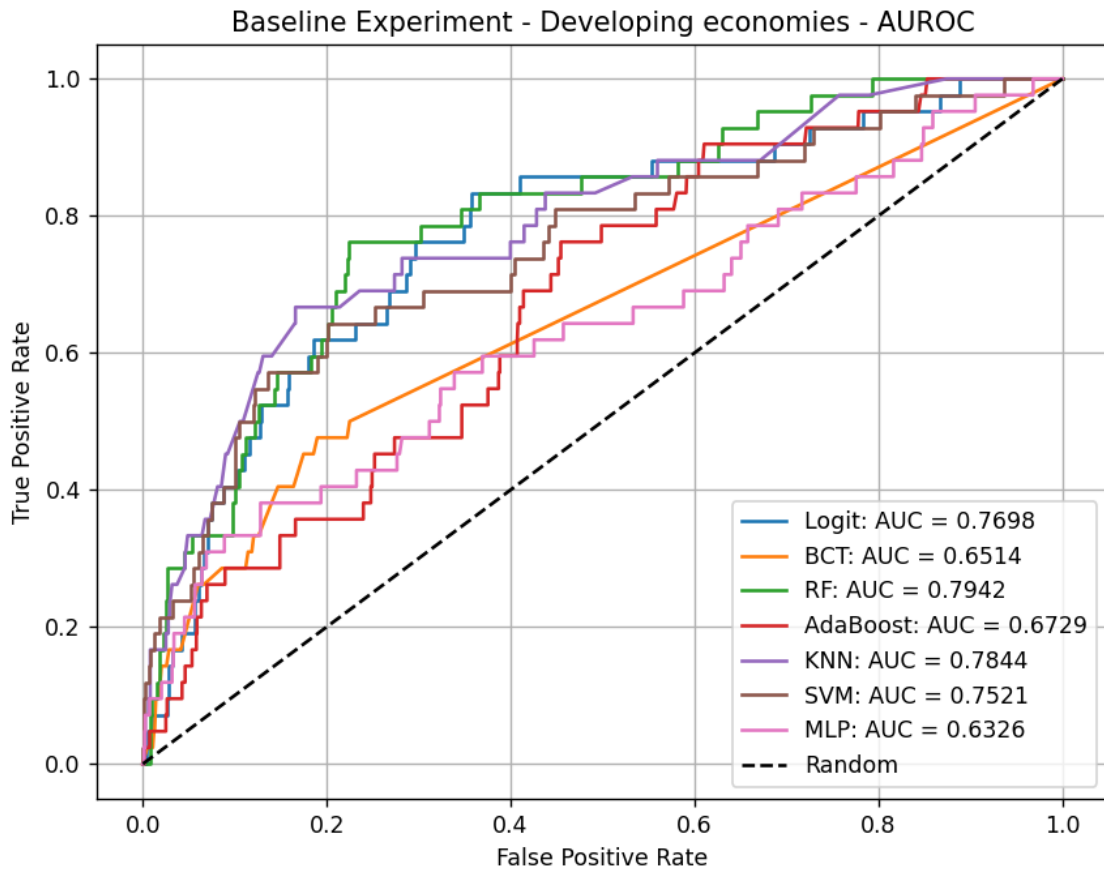


Figure 6.6: AUROC, baseline experiment, developing economies

It is immediately visible that the performance is slightly better for almost every model, even though precision scores remain equally low. The author of this dissertation hypothesizes that this difference is due to the fact that the hard task of predicting the Great Financial Crises of 2007-2008 in high-income countries is removed. Random Forest is the best performer in almost every evaluation metrics, with KNN being a close second. The Shapley values and Logit coefficients are reported in the Appendix (Figure A.10 and A.11). Looking at Logit coefficients,

they are mostly similar to what observed in the complete database and lead to the same conclusions. Two noticeable differences are observed in the current account balance, which is much less significant and even positive after the GFC, and real GDP growth which has negative coefficients throughout the whole testing period. This may suggest that banking systems in developing economies are more vulnerable to output shocks, which is a reasonable assumption, but also that they do not suffer particularly from current account deficits, opposed to what observed by Obstfeld (2012). The highest coefficients are recorded for inflation and loans/deposits ratio, and even higher for investments but only on the first years of analysis. The mitigating predictors are real GDP growth and public debt.

Shapley values again are erratic during the testing period, with the only conclusions that can be drawn being the same as in the complete dataset, regarding credit growth, government debt and international rates.

Regarding the subset containing only the observation retrieved from high-income countries, the results are presented in the following graphs.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,7086	0,0731	0,325	0,04	0,52	0,13	0,83	0,49	0,22
BCT	0,4359	0,1055	-0,001	0,95	0,92	0,00	0,00	1,00	0,00
RF	0,6099	0,0780	0,274	0,04	0,44	0,11	0,87	0,41	0,20
AdaBoost	0,7650	0,1432	0,424	0,32	0,53	0,14	0,93	0,49	0,24
KNN	0,5947	0,0761	0,188	0,03	0,52	0,11	0,68	0,51	0,19
SVM	0,3457	0,0770	0,103	0,27	0,92	0,44	0,12	0,99	0,18
MLP	0,7316	0,0784	0,423	0,03	0,68	0,17	0,75	0,67	0,27

Table 6.7: Performance score, baseline experiment, advanced economies

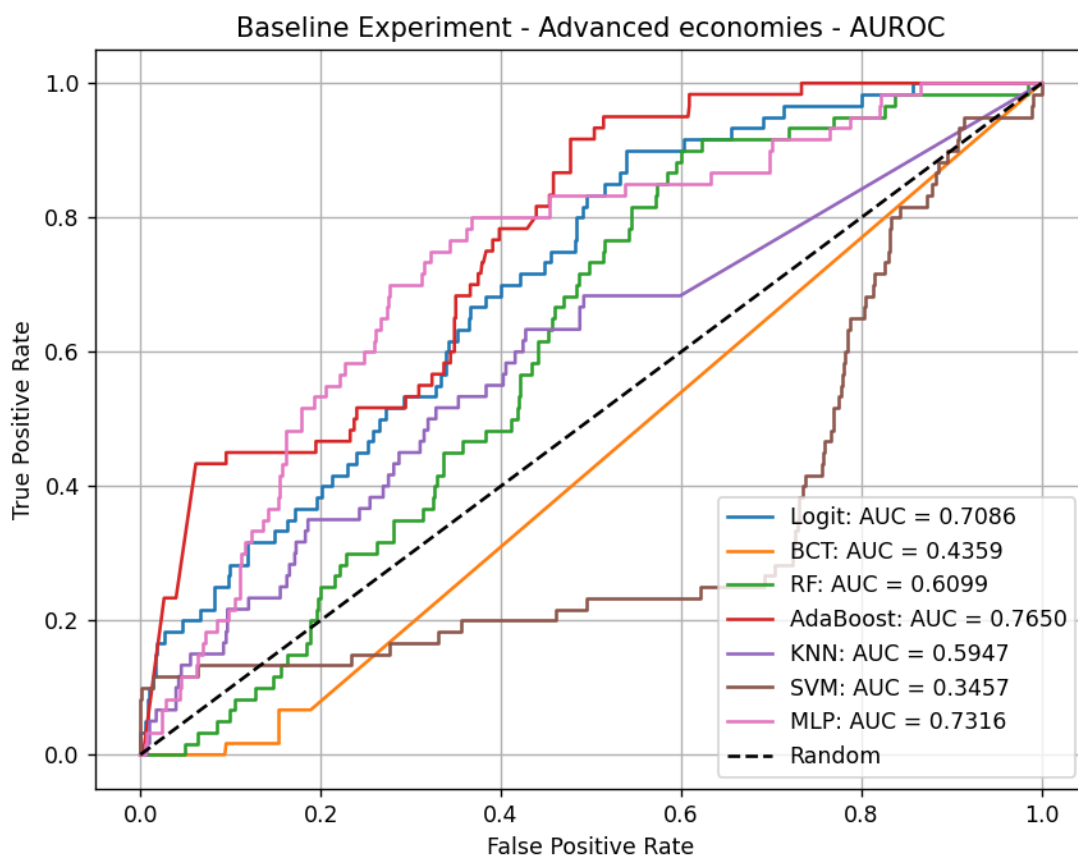


Figure 6.8: AUROC, baseline experiment, advanced economies

In this case, the same models achieved worse results on average, with some performing worse than a simple random guess and receiving a Usefulness score close to zero. The best models are Logit and AdaBoost, while the worst is also one of the most complex, SVM. It is important to note the strange thresholds set to minimise the Loss function, which are either very low (0.03 for KNN and MLP) or extremely high (0.95 for BCT). Again, Logit performed quite well against the more modern methods. The author of the Thesis hypothesises that the large difference in results in low and high-income countries obtained by the same models is due to the temporal window used for out-of-sample testing. The Great Financial Crisis (GFC) of 2007-2008 involved almost exclusively western high-income economies. The causes and dynamics of this event may be unique or somewhat different from previous banking crises in the western world, so that complex algorithms such as SVM and MLP, which are extremely good in data-fitting, struggle to predict events that do not appear in similar forms in the training dataset. This could also explain why a general model such as Logit Regression could still achieve some utility in this framework. Again, the Shapley Values obtained recursively for the high-income dataset are reported in the Appendix (Figure A.13), as well as the coefficients of the Logit model (Figure A.12). Analysing the Logit coefficients, some differences from the complete dataset catch the eye. Nominal depreciation has a negative sign up until 2008, after which it loses significance by narrowing zero. This contradicts the literature, but it is also important to

consider that many of the included high-income countries are part of the Euro Area, with the Euro acting as the second international reserve currency, which might mitigate the average vulnerability of advanced countries to the exchange rate with the US dollar compared to developing economies. Loans to deposit ratio has the highest coefficients of all predictors, twice the value from the complete database setting, confirming that financial crises can be the results of an unchecked credit bubble. International 10-year rate are also constantly higher than zero, which is a reasonable result also in accordance with the literature. The lowest coefficients are recorded for the current account balance, and for the yield curve slope, but only after 2008.

By trying to predict the crises in the USA, Italy, Dominican Republic, and Ukraine using the split datasets, this figure is obtained.

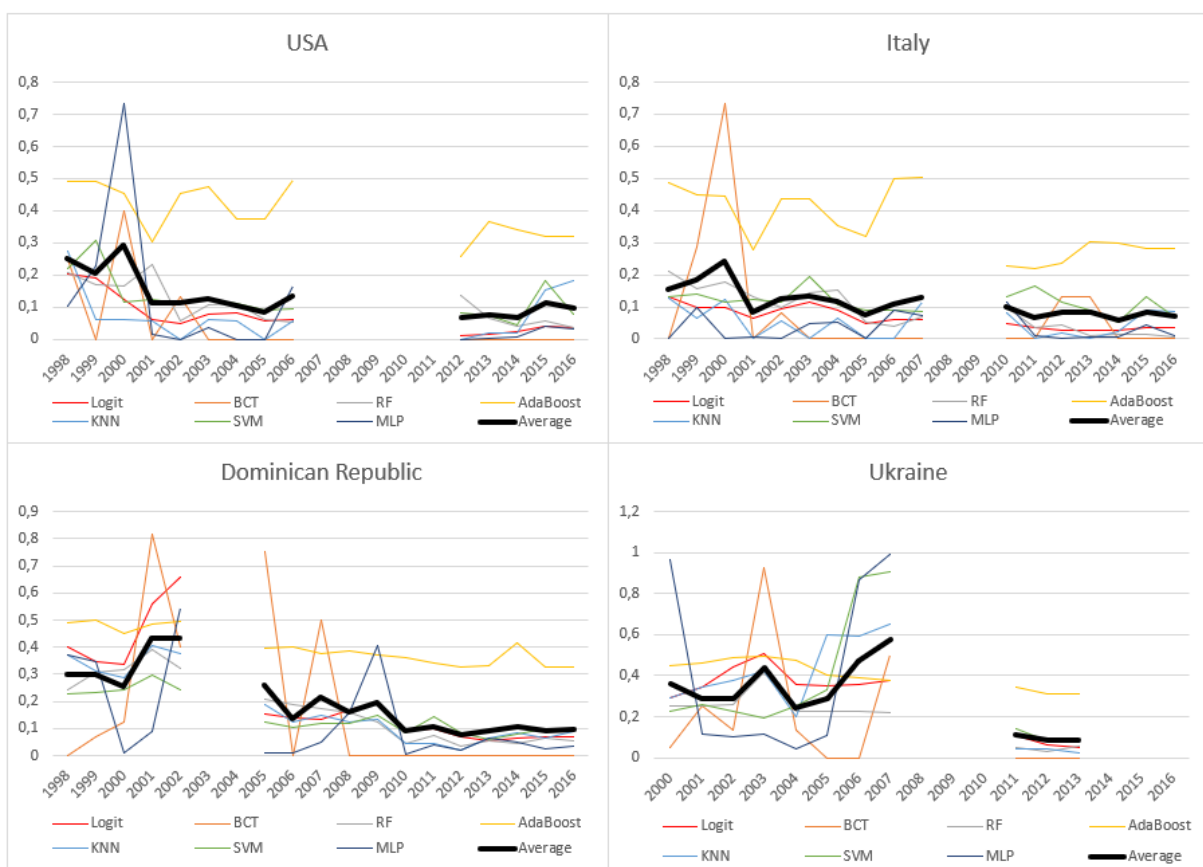


Figure 6.9: Predicted crisis probabilities by country, database split in low and high-income countries

By splitting the database in low and high-income countries, a slight improvement can be noticed in advanced economies, no longer signalling decreasing probabilities, even if the models still fail to clearly anticipate the GFC. In both USA and Italy, only the AdaBoost model correctly put up a warning sign. In the low-income countries the signals are still present, as in the complete database case. Five models correctly predict the Dominican crisis, while they are six for Ukraine. The best model appears to be AdaBoost, which correctly warns about all four upcoming crises. Overall, it can be said that splitting the database in developing and advanced



countries could bring only modest to no improvements to the models' predictive capabilities, depending on the countries income level.

The next step is to check the performance of the same models, using the exact same data, with a different approach. Instead of a real-time recursive approach, the models are tested using 10-fold cross-validation along the whole dataset period, from 1970 to 2016, as has been previously done in the literature. This procedure divides the entire dataset in ten folds of roughly equal length and containing adjacent observations, and then each fold is used as testing dataset while the other nine act as the training dataset, independently of their temporal positioning. The author implemented the module *cross\_val\_predict* from the Python library *sklearn* to run this analysis. The results for the complete database evaluated using cross-validation are presented below.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,6876	0,1126	0,331	0,07	0,57	0,15	0,79	0,54	0,25
BCT	0,5852	0,1832	0,198	0,08	0,66	0,14	0,52	0,68	0,22
RF	0,6514	0,0872	0,327	0,10	0,54	0,14	0,82	0,51	0,25
AdaBoost	0,6650	0,1515	0,331	0,34	0,46	0,14	0,92	0,41	0,24
KNN	0,7152	0,0840	0,376	0,07	0,56	0,16	0,84	0,54	0,26
SVM	0,5309	0,0851	0,113	0,09	0,56	0,11	0,55	0,56	0,19
MLP	0,7249	0,0834	0,160	0,10	0,81	0,18	0,29	0,87	0,22

Table 6.10: Performance score, cross-validation, complete database

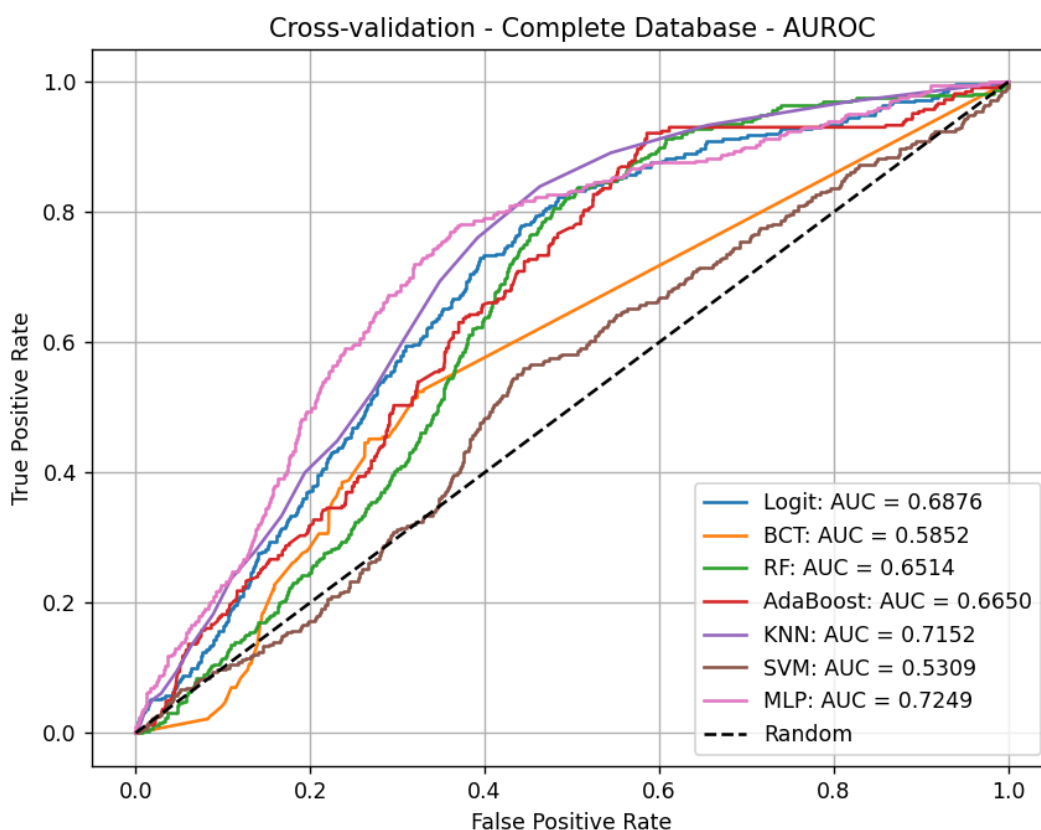


Figure 6.11: AUROC, cross-validation, complete database

The out-of-sample testing done through cross-validation leads to slightly better performance scores for most of the models, not so much regarding AUROC and Usefulness but more significantly on the other indicators. Precision and F1 scores doubled compared to the recursive approach. It is also interesting to notice how the curves in AUROC graphs now have a much more smooth and consistent shape, and how the results are more homogeneous between different models. In this framework Logit is no longer the preferred method, being surpassed by the more sophisticated MLP and KNN. The improvements in prediction capabilities may probably be explained by the fact that the training set now contains a bigger portion of data, part of which recorded after the testing year. This supplies the algorithms with a larger number of crisis events from which they could learn different patterns leading to financial troubles. The models trained with this data ‘learned’ that a similar combination of predictors’ values must be associated with a pre-crisis period, and later successfully uses this information in the testing-fold. In the Appendix, in *Figure A.15* and *A.17* and *Table A.14* and *A.16*, are reported the cross-validation results for the low and high-income datasets. As for the complete dataset, there are noticeable improvement in Precision and F1 score, but not so much in AUROC. Advanced countries dataset benefited particularly from cross-validation evaluation, with all curves in the AUROC graph now well above the random guessing line. Again, in this framework Logit Regression is surpassed by the advanced machine learning algorithms in predictive capabilities. Using cross-validation estimated crisis probabilities, the author generates the following four graphs regarding the usual four-countries panel used as example.

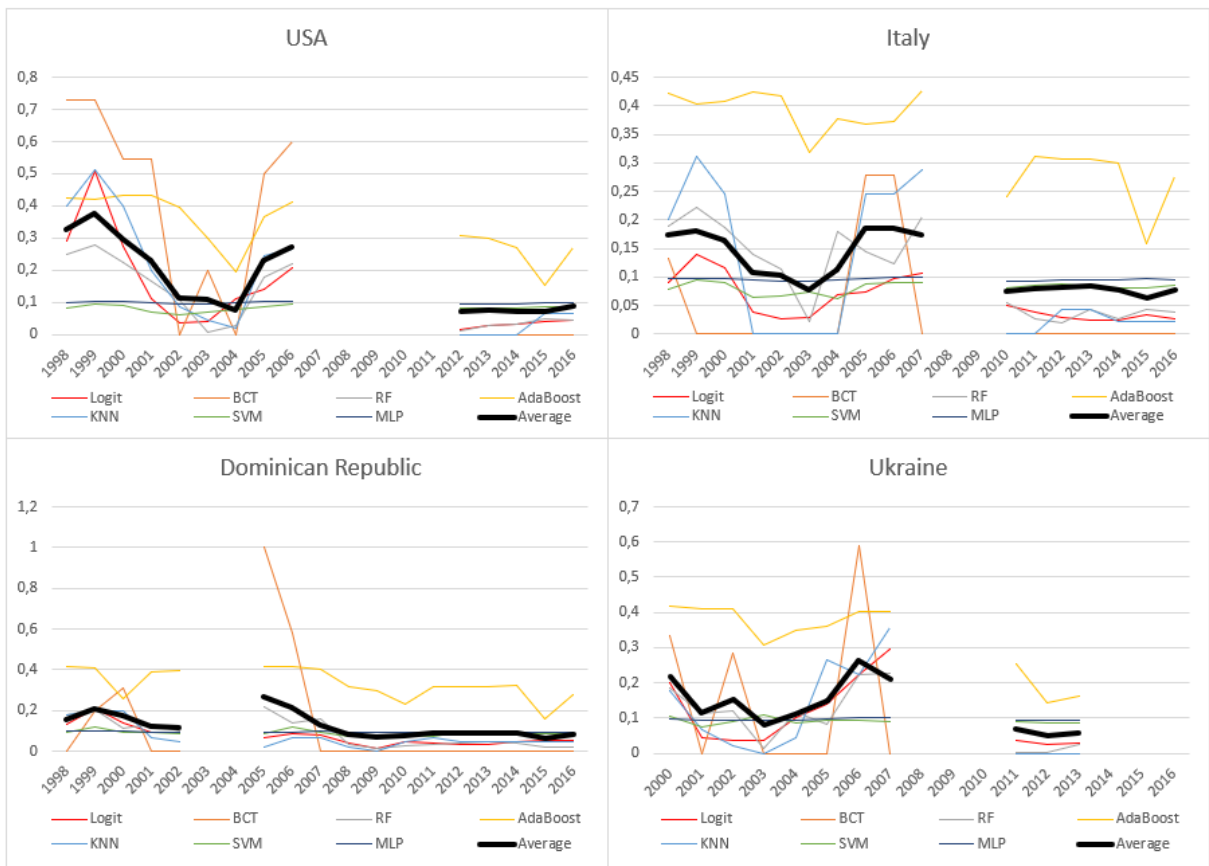


Figure 6.12: Predicted crisis probabilities by country with cross-validation

Contrary to the real-time recursive approach, it is noticeable that the models can perceive the upcoming financial crisis both in the USA and Italy two or three years in advance. Predicted probabilities are growing in both countries starting in 2004, with five and six signals above warning levels in the USA and Italy respectively. It is interesting also to observe the high values signalled in the USA before the Dot-com bubble of 2000. This is a great improvement from the recursive approach, in which the signals were flat or even slowly declining before the event, and mostly below warning levels. On the other hand, in Ukraine the signal is somewhat less strong than in the case where recursive evaluation is used (four warnings out of seven models), and in Dominican Republic the signal is mostly lost (only two warnings). This reinforces the AUROC scores computed previously, indicating that, in this specific framework, cross-validation evaluation leads to higher performance of the models when analysing high-income countries. These findings partially reinforce the conclusions of Beutel et al. (2018), who discussed the work of Alessi and Detken (2018) stating that “the cross-validation procedure may provide an inflated estimate of the performances of these methods”, even if in this Thesis this is particularly true for high-income regions, while in the case of Dominican Republic the opposite result is drawn. In *Figure A.18* in the Appendix the cross-validation evaluation is applied to the split datasets, resulting in impaired predictive capabilities in Italy and Dominican Republic, without significant gains for the other two countries of the example panel.

# VARIATIONS

In this chapter, some variations of the baseline exercise are experimented with. The objective is to try different approaches, some taken from the literature, and some elaborated by the author of this Thesis, and compare the results with the baseline variation. The analyses are run on the complete database, for brevity reason and given the modest improvement and mixed results brought by the low and high-income split. All variations are run and evaluated using the usual real-time recursive approach.

## Removing Micro-nations

The presence of multiple micro-nations in the dataset could impair the predictive capabilities of the models, given that the dynamics that drive such small nations could be non-comparable or non-appliable to larger countries. Moreover, the models do not assign a greater weight to larger countries, so that variables collected in a micro-state such Aruba could potentially affect the model as much as variables collected in the United States. The presence of micro-states could especially harm prediction for high-income countries, given that nations with small population overwhelmingly belong to the high-income specification. For these reasons, the baseline experiment is run again, filtering data to remove the smallest countries. The author decided to label a country as a micro-nation if its population is below one million individuals in 2022. Data about population numbers is retrieved from World Bank database.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,7346	0,0604	0,363	0,13	0,60	0,09	0,77	0,59	0,16
BCT	0,5528	0,0756	0,107	0,13	0,83	0,08	0,25	0,86	0,12
RF	0,5972	0,0522	0,147	0,17	0,80	0,09	0,32	0,82	0,14
AdaBoost	0,6298	0,1949	0,254	0,48	0,69	0,08	0,56	0,69	0,15
KNN	0,6550	0,0595	0,288	0,08	0,45	0,07	0,85	0,43	0,13
SVM	0,6107	0,0583	0,226	0,13	0,28	0,06	0,98	0,25	0,12
MLP	0,5668	0,0648	0,307	0,16	0,62	0,08	0,69	0,62	0,15

Table 7.1: Performance score, no micro-nations

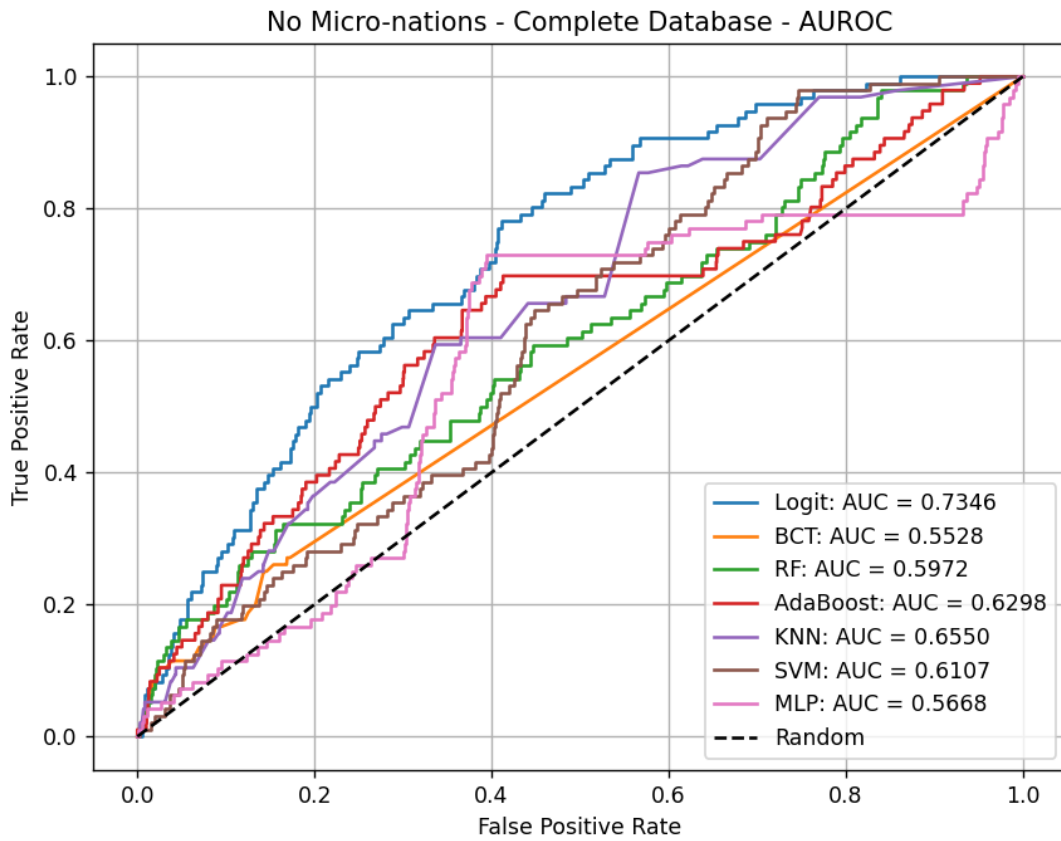


Figure 7.2: AUROC, no micro-nations

The results show that the models do not benefit in a significant way from a dataset filtered from the smaller states' data. The average performance score is just slightly higher than in the baseline experiment. The predictions on the four-countries panel are as follow.

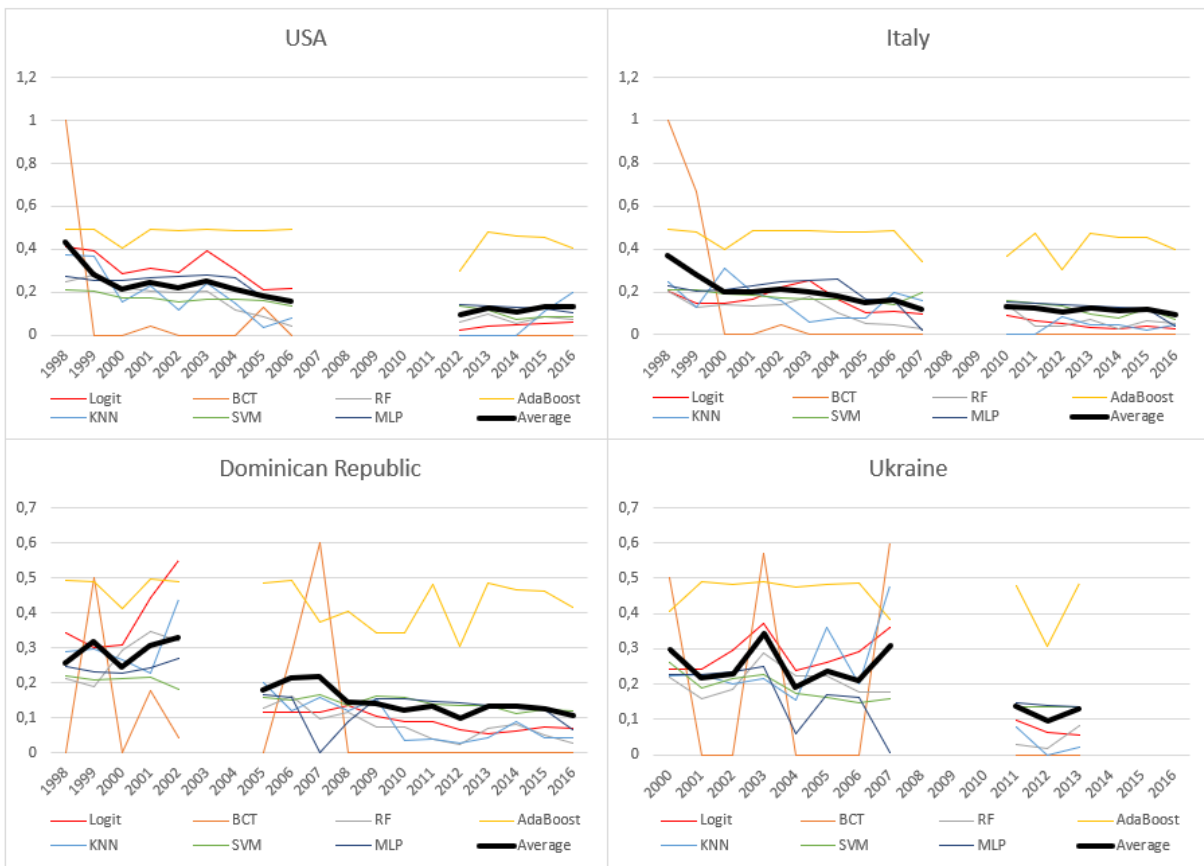


Figure 7.3: Predicted crisis probabilities by country, no micro-nations

As in the baseline experiment, the models do not show increased probabilities before the GFC in high-income countries, but both in USA and Italy four out of seven models are in a warning state in the preceding years, which is an improvement from the baseline framework. The signals appear to be slightly weaker in low-income nations, even if five models are still signalling in each country. The Logit model again presents the highest performance scores, and together with SVM is able to signal the upcoming crisis in all four examined economies. This variation of the baseline exercise brings some improvements when looking at the number of models able to correctly warn about financial distress, however it is important to notice how, especially in high-income countries, false alarms in tranquil periods (before 2004) are even more relevant than in the baseline framework and may probably lead to underestimate the correct signals and dismiss them as another false alarm. Because of this, the author reaches the conclusion that the presence of values recorded in micro-nations is not the cause of the low performance of the models when using the high-income dataset.

## Standard crisis duration

As an ulterior test, the definition of crisis period is changed. In the baseline experiment, the duration of a crisis was set as reported by the authors from which the crises databases are retrieved. This duration could span from just one year, to a decade. In the literature the standard approach is to set the duration of a crisis to just one year, the year in which the event began, and then to consider the following three (but in some case five) years as a post-crisis period to be removed from the analysis to avoid the post-crisis bias. By not considering the effective duration of a crisis, assuming a standard duration of one year and removing from the analysis all observations in the three following years, the following results are obtained.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,7551	0,0503	0,413	0,10	0,55	0,08	0,88	0,54	0,15
BCT	0,5002	0,0682	0,019	0,20	0,85	0,05	0,13	0,88	0,07
RF	0,6235	0,0456	0,203	0,06	0,55	0,06	0,66	0,54	0,11
AdaBoost	0,6585	0,1594	0,281	0,35	0,53	0,07	0,76	0,52	0,12
KNN	0,6918	0,0495	0,295	0,03	0,36	0,06	0,96	0,34	0,12
SVM	0,5544	0,0491	0,190	0,16	0,79	0,08	0,38	0,81	0,14
MLP	0,6698	0,0524	0,291	0,12	0,35	0,06	0,97	0,32	0,12

Table 7.4: Performance score, standard crisis duration

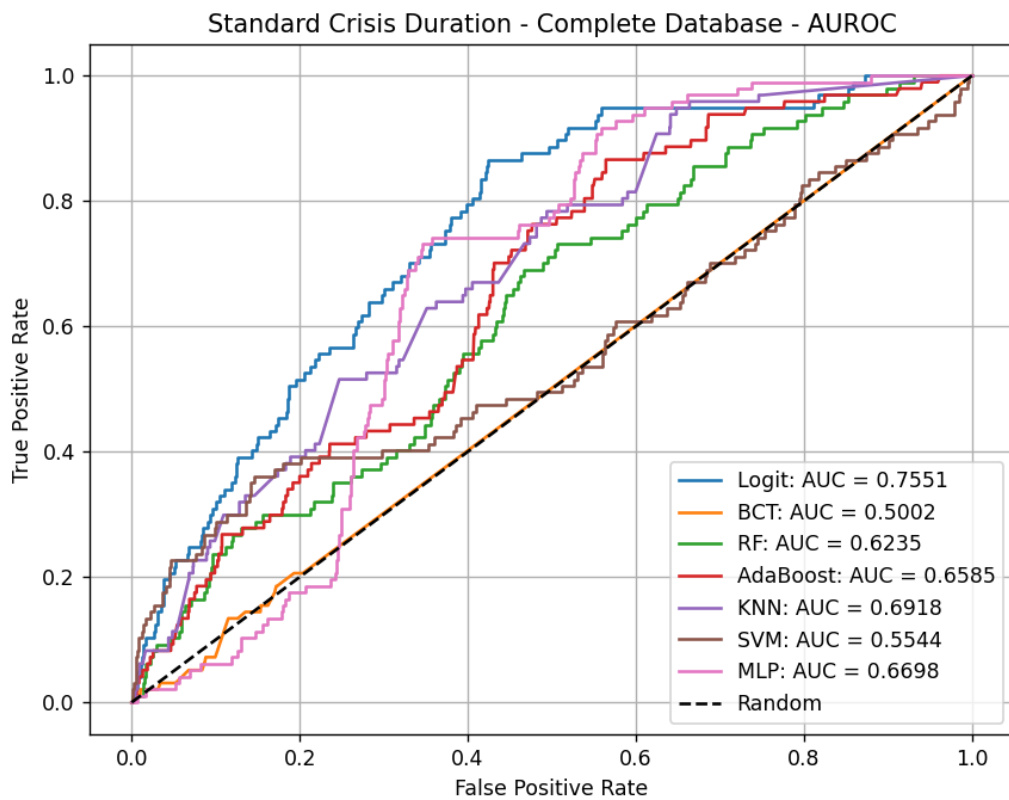


Figure 7.5: AUROC, standard crisis duration

The results show that the average score obtained by the models is similar, but slightly lower, than the baseline framework. Precision values are still very low, and the Logit Regression

achieves the best results in almost every parameter. On the other hand, the AdaBoost algorithm seems to benefit the most from this change in approach. The predictions of the four-countries panel are presented below.

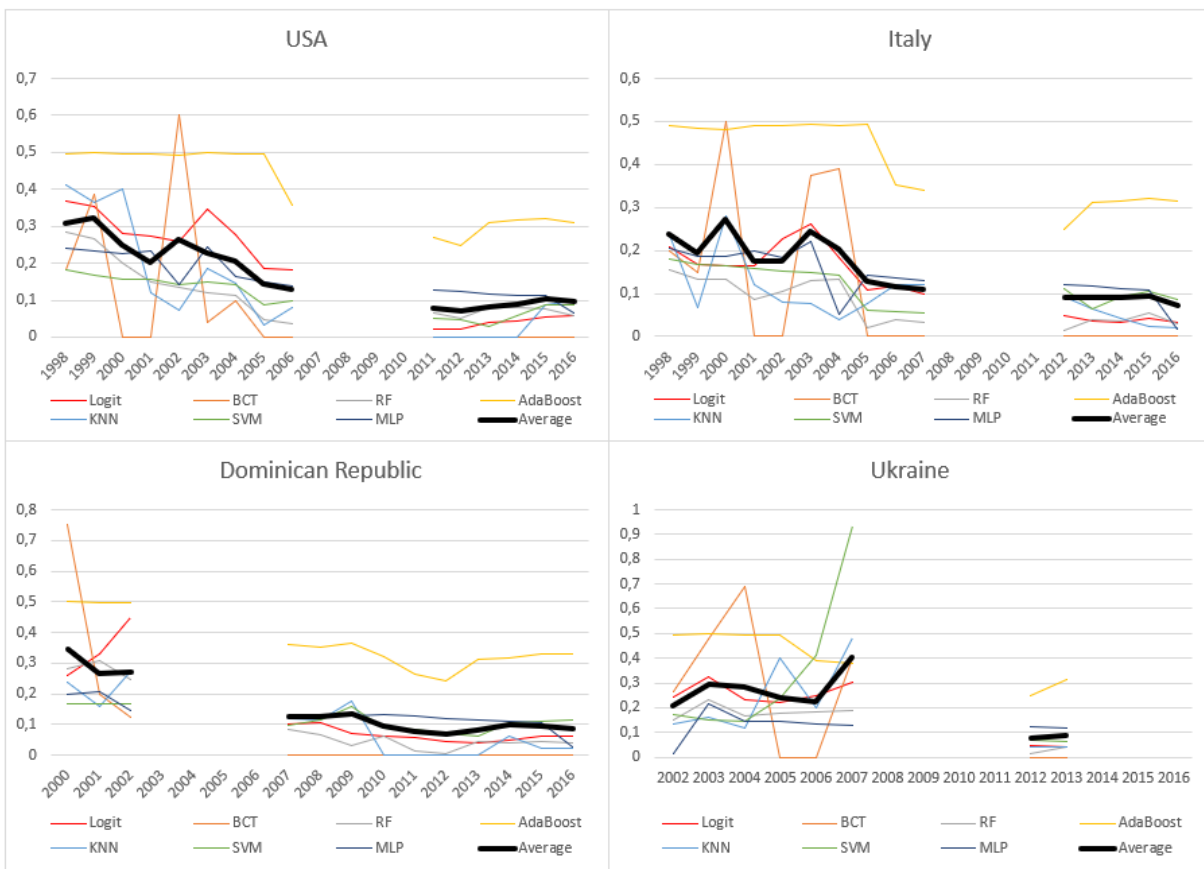


Figure 7.6: Predicted crisis probabilities by country, standard crisis duration

By looking at how the signals changed with the new crises' duration settings, it appears that no improvements are made for the USA and Italy, while the models show decreasing probabilities for the 2003 Dominican banking crisis, even if five models still produce values above the thresholds. Ukraine signals are still present, even though with weaker emphasis. Overall, the models achieve slightly better performance when only the actual period of financial distress is removed from the training dataset, and not a generalised number of years equal for all crisis events. As seen from the crises databases, the effects of such events require a very heterogeneous time to vanish, and the author of this Thesis thinks that considering a standard fixed duration when the real observations are available is a waste of useful data, which most probably is reflected on the above results.



## Models' aggregation

As a further variation on the main exercise, the author takes inspiration and follows the example of Fouliard et al. (2020) and Holopainen and Sarlin (2017) and set an aggregating rule to combine the results of all models. The method used is the Exponentially weighted average (EWA) aggregation rule, which combines the results of the seven models by assigning the weights based on performance in an exponentially increasing way. For each recursive training dataset, all models are run, and the AUROC score of out-of-sample predictions is recorded for each model. The AUROC score then determines the weights associated to each model 3 years ahead. The weights are computed with this 3-year delay so to keep a real-world setting, since it would be impossible for a forecaster to compute the AUROC score for recent prediction, given that the pre-crisis status of a year is determined only at the end of the following third year. After all models have been run, the probabilities predicted by each method are multiplied by the corresponding weight. Since the AUROC scores for the first three years are missing due to the real-world framework, the author implemented equal weights for all models in the period 1998-2000. The weighted probabilities estimated by each model for each observation are then summed to obtain a single probability, the EWA output. The formula used by the author is as follows:

$$EWA = \frac{\sum_{i=1}^n w_i \cdot x_i}{\sum_{i=1}^n w_i}$$

in which  $n$  is the number of models implemented,  $x_i$  is the probability predicted by model  $i$ , and  $w_i$  is the weight assigned to that prediction based on the AUROC score computed on the predictions up to three years before, as in:

$$w_i = e^{k \cdot AUROC}$$

Where  $k$  is the parameter used to control the exponential growth of the weights. In this dissertation  $k$  is set equal to 10. This might seem a quite high value, but after some experimenting by the author it is found to be necessary to reduce the weight of the least performing models to low values. By running the analysis using the usual database, the following weights are applied to the seven different algorithms over the testing period. As explained before, in the first three years of recursive testing the weights are set in equal proportions, since an AUROC score is not yet available to the forecaster at that time.

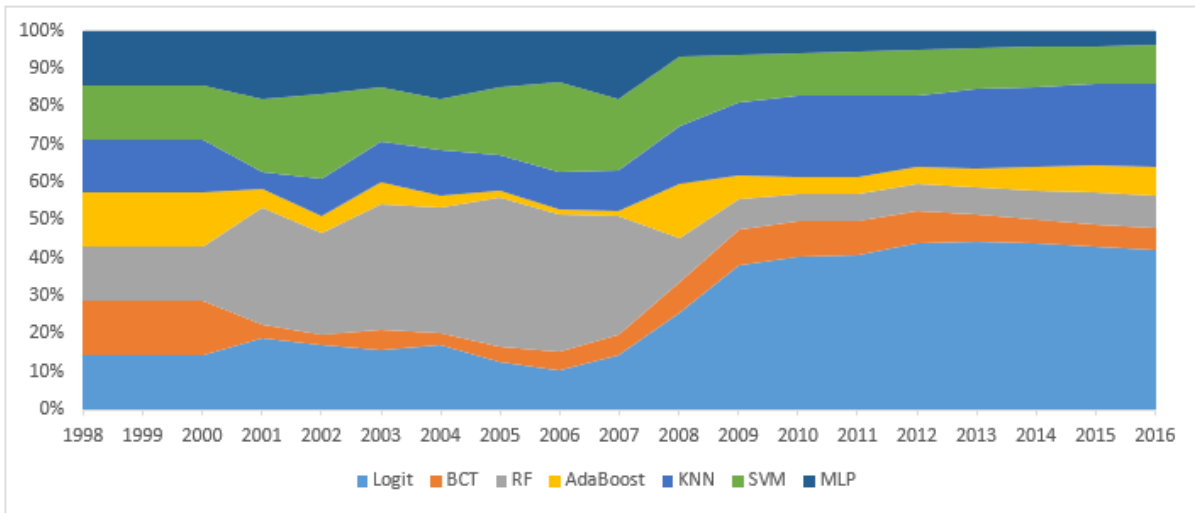


Figure 7.7: Weights computed by the EWA model

The weights computed show how Random Forest is the preferred method up to the burst of the Financial Crisis of 2007-2008, after which the Logit Regression gains larger weights. The author of this dissertation hypothesises that the timing may not be coincidental and may be due to overfitting of the machine learning models and the entrance on the scene of an unpredicted kind of crisis which may benefit the Logit model.

Here are presented the results obtained by the EWA in the usual real-time recursive approach.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
EWA	0,5703	0,0521	0,190	0,21	0,85	0,11	0,31	0,88	0,16

Table 7.8: Performance score, EWA

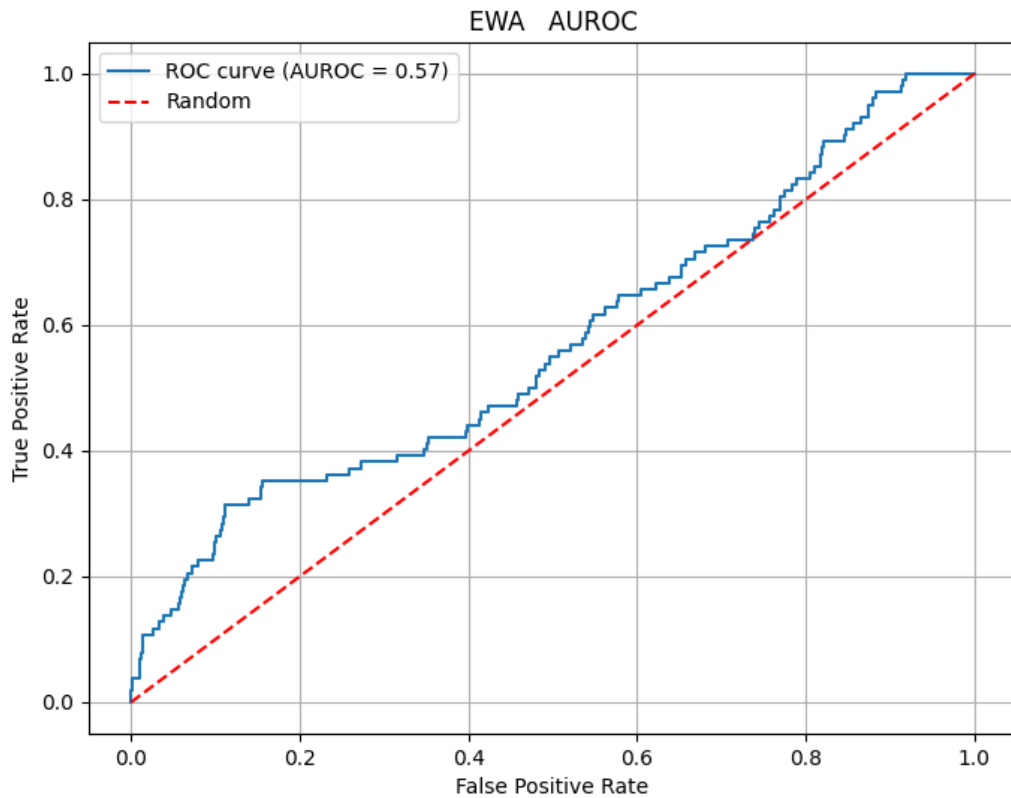


Figure 7.9: AUROC, EWA

Given the complexity of the model involving seven different models and their thoughtful combination, the results obtained are quite disappointing. This might indicate that the algorithms which achieve the best out-of-sample performance up to a certain point in time are not necessarily the ones that will also perform best in future predictions. This result is quite discomfoting, since it may imply that the search for a general best model is vane and may only lead to the detection of the model which worked best in the past. The predictions regarding the four-countries panel are presented in *Figure 7.10*.

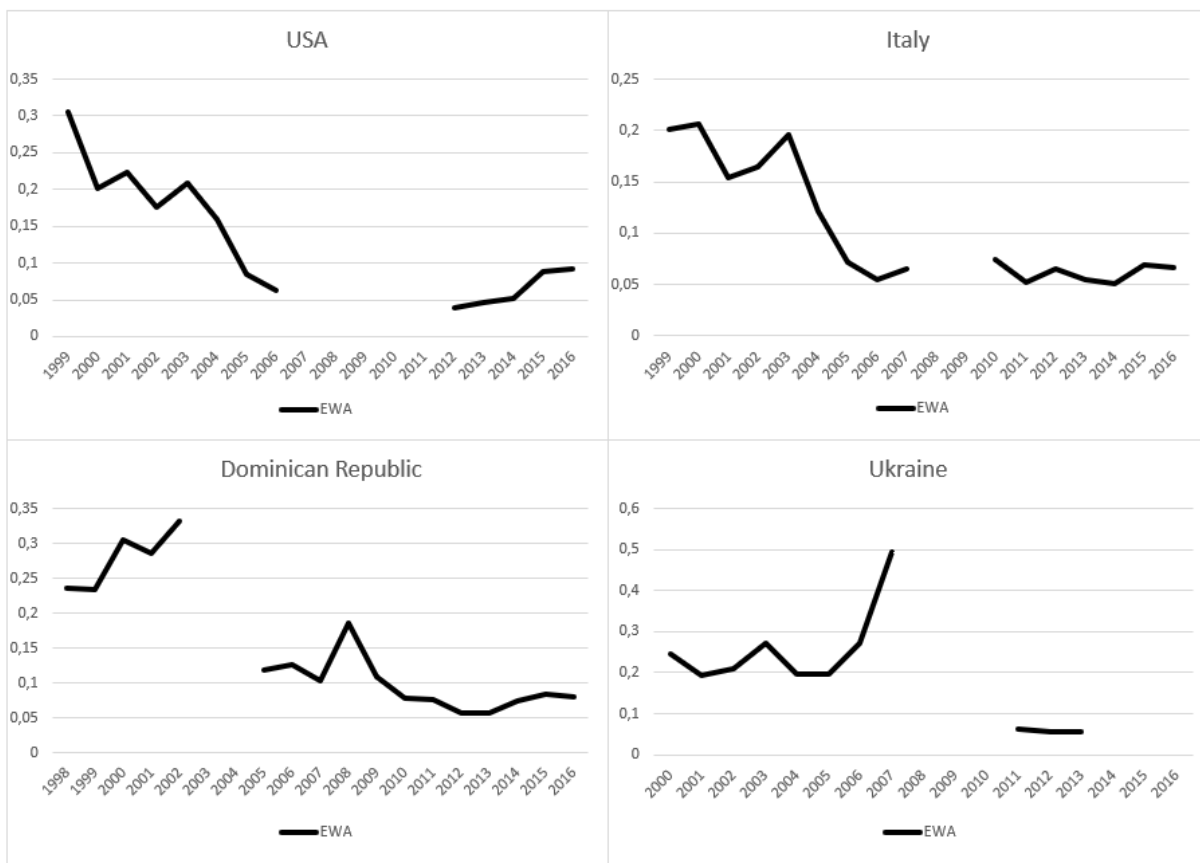


Figure 7.10: Predicted crisis probabilities by country, EWA

As in the baseline experiment, the models can predict financial distress in developing economies but erroneously signal the opposite outcome in high-income regions. In advanced economies, values are high and signal a warning until 2003, and then decrease rapidly at the eve of the GFC.

# LAGS ANALYSIS

In this chapter, the author uses the same database as in the baseline framework, with the addition of three lags for each observation. This approach has an adverse effect on the dataset length, since as explained before if any of these three lags is recorded during a crisis period, the observation must be removed from the analysis to avoid crisis-bias. This approach effectively leads to the removal of three consecutive years at the end of each crisis episodes, shortening the available dataset from 3576 observations to 3181.

## Machine Learning algorithms – Adding 3 Lags

For this variation, the author runs the same recursive test using the same seven algorithms and using the values recorded at  $t_0$  plus three more lags for each variable, recorded at  $t-1$ ,  $t-2$ , and  $t-3$ , taking inspiration from what has been previously tried in a similar fashion by Fricke (2017) and Tölö and Eero (2019). By doing so, each model works with a total of 48 variables. The results achieved are here presented.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,7976	0,0677	0,467	0.23	0,73	0,11	0,73	0,73	0,19
BCT	0,4829	0,0903	0,006	0.89	0,94	0,06	0,02	0,98	0,03
RF	0,5193	0,0511	0,055	0.17	0,81	0,06	0,22	0,83	0,09
AdaBoost	0,6866	0,1778	0,381	0.4	0,51	0,07	0,89	0,49	0,14
KNN	0,6914	0,0503	0,255	0.07	0,40	0,06	0,87	0,38	0,11
SVM	0,6203	0,0512	0,372	0.11	0,47	0,07	0,93	0,45	0,13
MLP	0,6864	0,0534	0,375	0.12	0,48	0,07	0,91	0,46	0,13

Table 8.1: Performance score, three lags

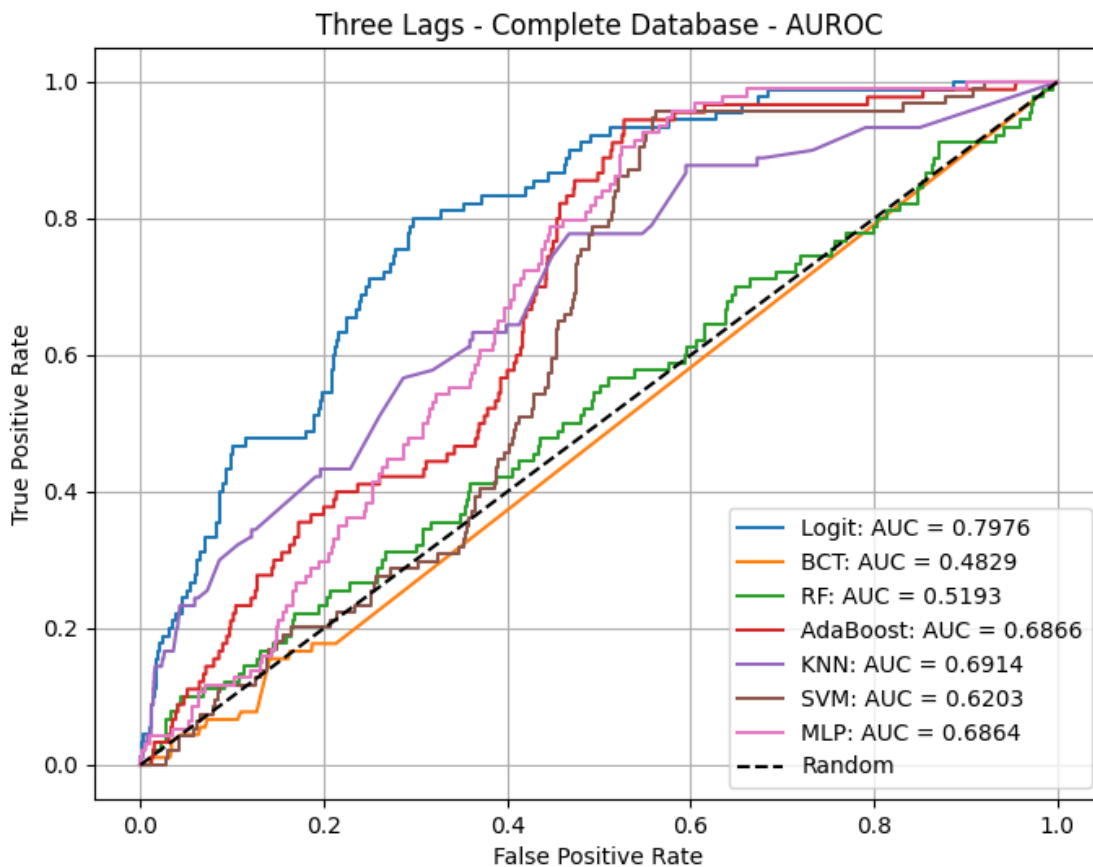


Figure 8.2: AUROC, three lags

Looking at the results, adding three lags for each predictor and analysing all lags simultaneously brings no improvements to the machine learning algorithms on average, while the Logit Regression show marginally higher scores. AdaBoost also shows improved performance compared to the baseline experiment. It is interesting to notice how adding three lags especially reduces predictive performance of the trees models, BCT and Random Forest, weakening them to the point of being predictors as good as random guessing. The predictions on the four-countries panel are presented below. The crisis duration is three years longer than in the baseline setting because of the three years lags adopted and the need to avoid crisis-bias as previously explained.

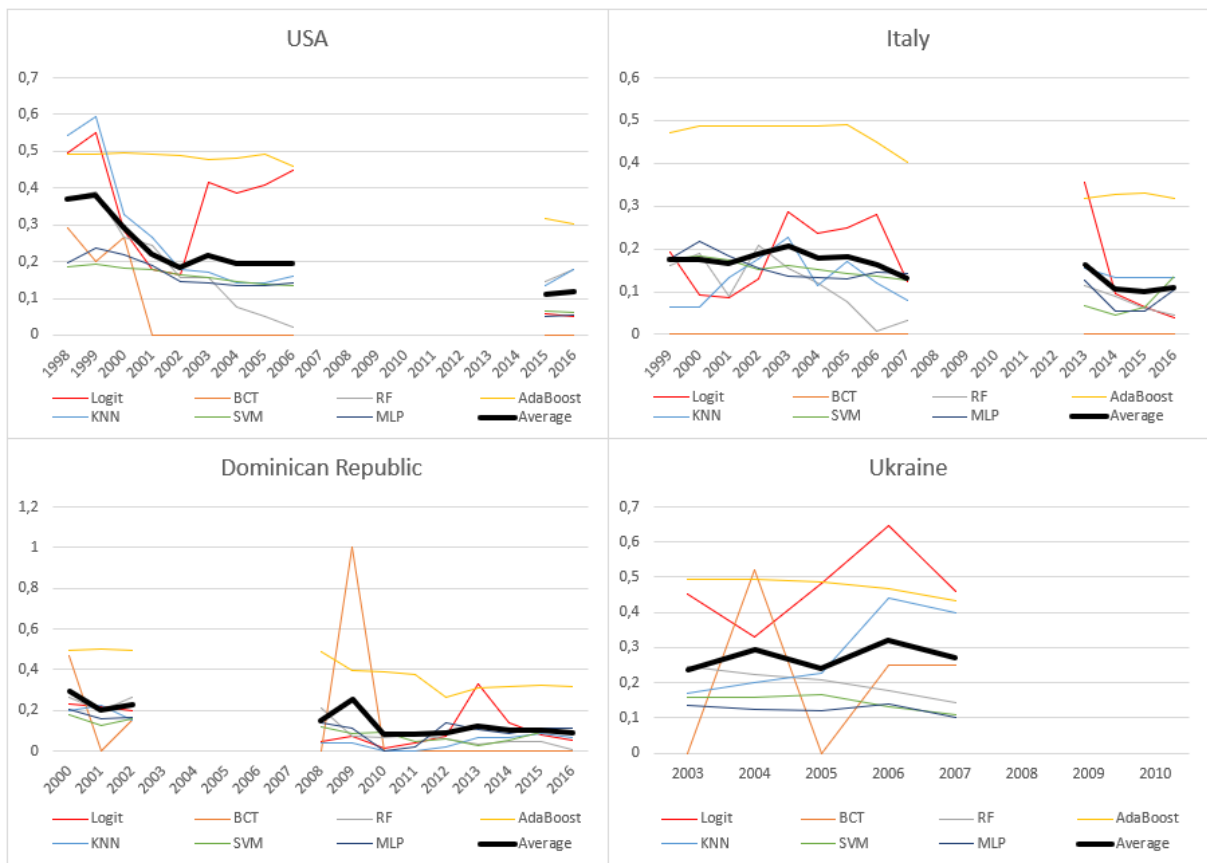


Figure 8.3: Predicted crisis probabilities by country, three lags

As can be deduced from the performance table, the predictive performance of this setting is weaker than in the baseline framework. Not only the GFC is preceded by weaker warnings in high-income countries, especially in Italy, but the signals in developing countries are now much less strong, with fewer models producing outputs above alarm thresholds. As previously highlighted, Logit and AdaBoost somewhat benefit from this setting and show higher crisis probabilities than the other models.

In the next experiment, the analysis involves multiple training stages using only one variable, with only one lag at a time:  $t_0$ ,  $t-1$ ,  $t-2$ , and then  $t-3$ . The lag which yields the highest AUROC score in a 10-fold cross-validation evaluation is recorded for each variable in each year, assuming that is the lag which best contributes to correct predictions. Then, the overall out-of-sample testing involving all predictors, each chosen based on its best lag for each testing year, is run in the same recursive approach. As a side note, in the phase in which the best lag for each predictor is estimated in the MLP model, the hyperparameters regarding the neurons and layers numbers are adjusted given that the analysis is executed with only one predictor at a time.

```

# Define the parameter grid for tuning
param_grid = {
    'hidden_layer_sizes': [(1,), (2,), (1, 1), (2, 2)],
    'activation': ['relu', 'tanh', 'logistic'],
    'solver': ['adam', 'sgd'],
    'max_iter': [10000]
}

```

The hyperparameters of all other models are kept the same. The results are presented in *Table 8.4* and *Figure 8.5* and *8.6*.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,7692	0,0529	0,403	0,17	0,73	0,10	0,67	0,74	0,18
BCT	0,5189	0,0865	0,095	0,08	0,71	0,06	0,37	0,73	0,10
RF	0,6356	0,0500	0,210	0,07	0,48	0,06	0,74	0,47	0,11
AdaBoost	0,6956	0,2076	0,343	0,47	0,59	0,08	0,76	0,59	0,14
KNN	0,6624	0,0494	0,295	0,12	0,66	0,08	0,63	0,66	0,14
SVM	0,6572	0,0515	0,321	0,13	0,41	0,07	0,93	0,39	0,12
MLP	0,7451	0,0515	0,418	0,11	0,56	0,08	0,88	0,54	0,15

Table 8.4: Performance score, best lags

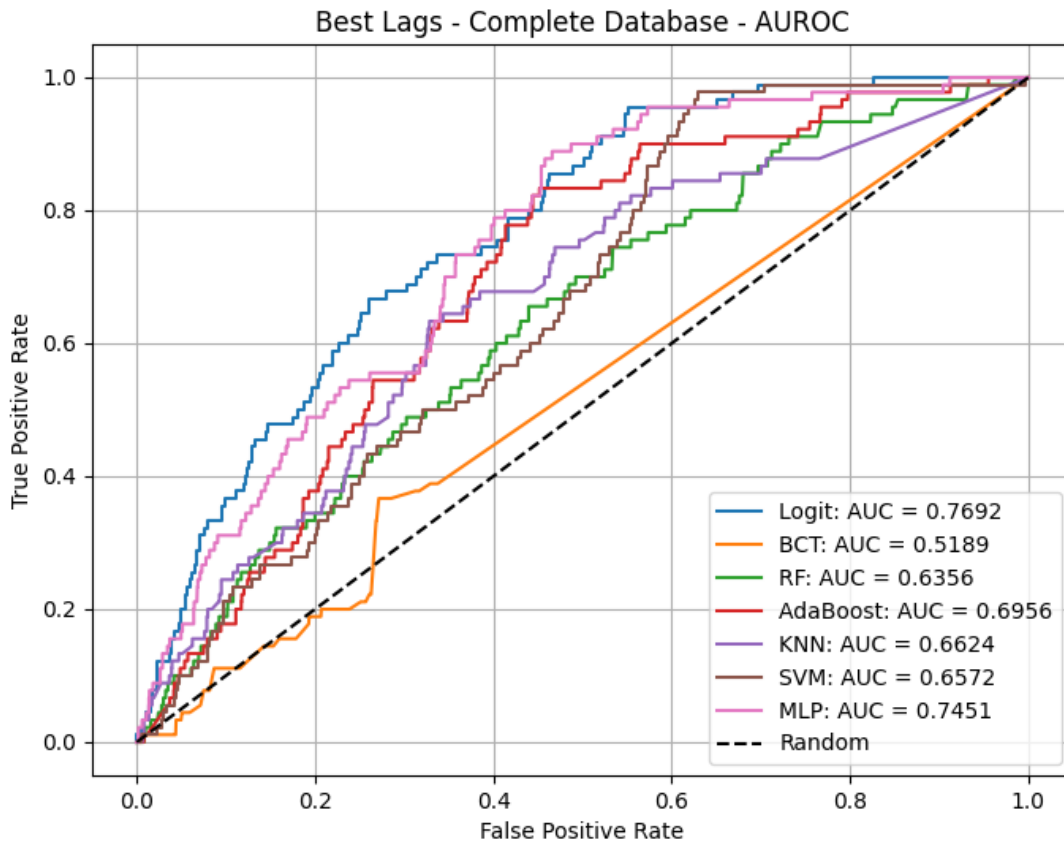


Figure 8.5: AUROC, best lags



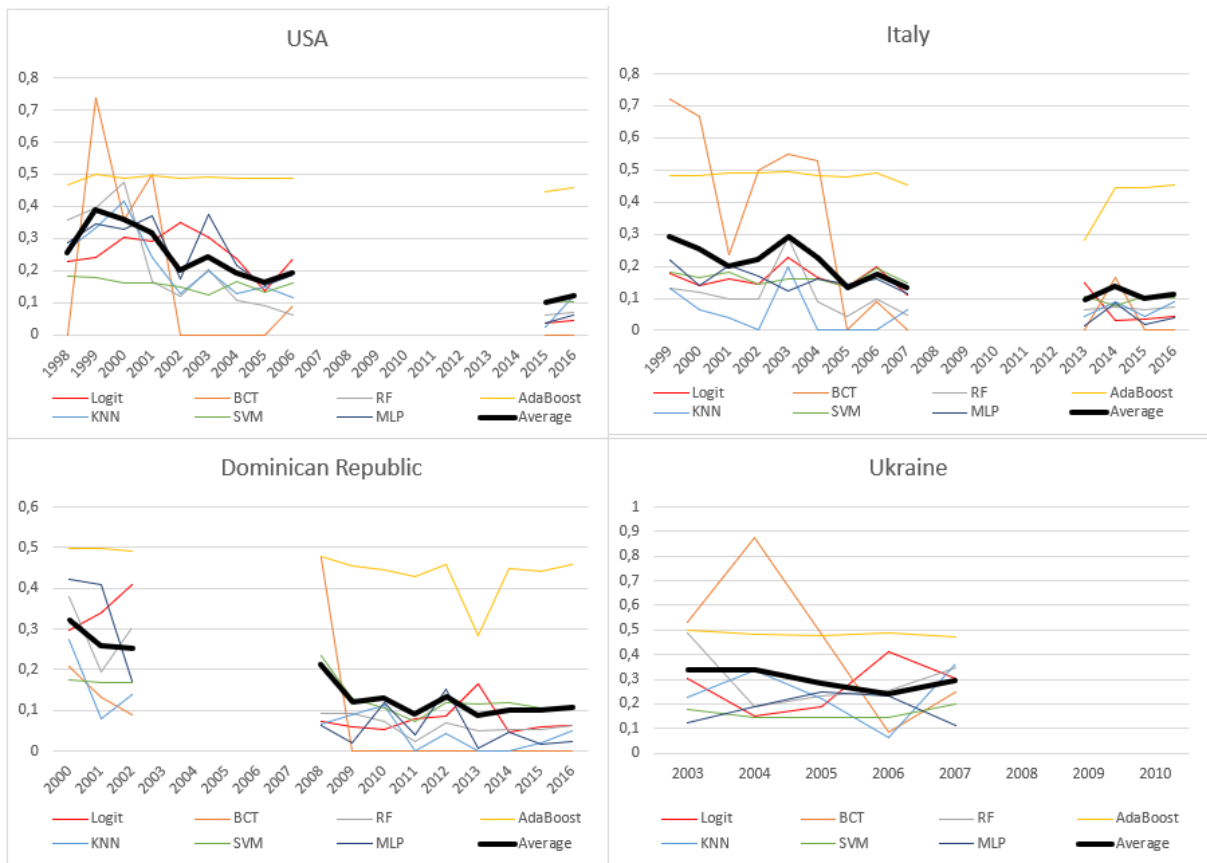


Figure 8.6: Predicted crisis probabilities by country, best lags

A quick analysis of the results suggests that only a marginal improvement is brought by the in-sample selection of the best lag for each variable. Using this approach, the upward trend of signals for the low-income countries is lost when comparing with the baseline case. Most signals are still above alarm level in USA, Dominican Republic, and Ukraine, but flat or declining trends may do so that a forecaster would dismiss them as false alarms. In Italy, only SVM produces a probability above warning level. Again, the author hypothesis that this may be due to the fact that the lags that best suit the crises dynamics in the past might not be the same that will best explain the surge of such events in the future.

## Recurrent Neural Networks

A last experiment is carried out using the most sophisticated algorithms present in this dissertation, Recurrent Neural Networks. The dataset implemented is the same as in the previous experiments using lags, which implies that RNN will be fed each observation's data at  $t_0$  accompanied by three lags. Two different typologies of RNN are used, Simple RNN and Long short-term memory (LSTM), in this order, each built with a six neuron – single hidden layer structure.

Regarding Simple RNN, the results achieved are as follow.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
SimpleRNN	0,7964	0,0719	0,466	0,15	0,68	0,10	0,79	0,68	0,18

Table 8.7: Performance score, Simple RNN

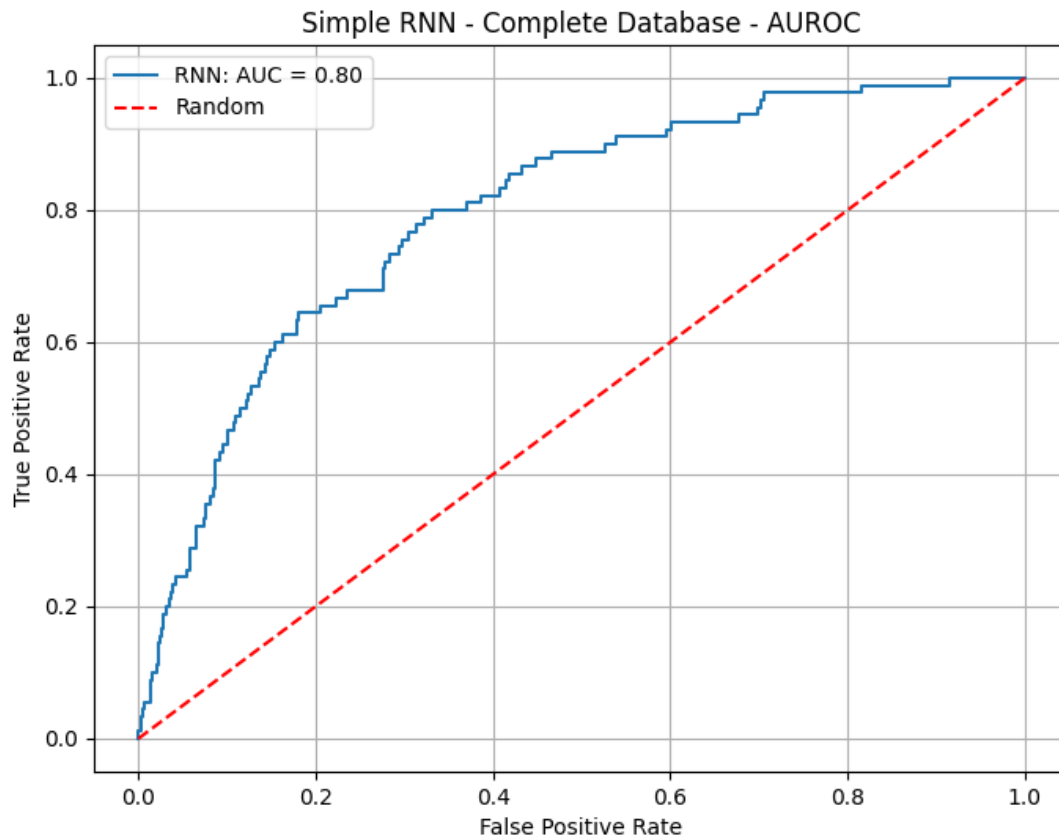


Figure 8.8: AUROC, Simple RNN

The performance of Simple RNN is substantially higher than the previous machine learning models, achieving the highest AUROC score up to now. Precision, however, is still lacking even if F1 and Usefulness scores are well above the previous results' average. The author also elaborated the corresponding signals for the four-countries panel.

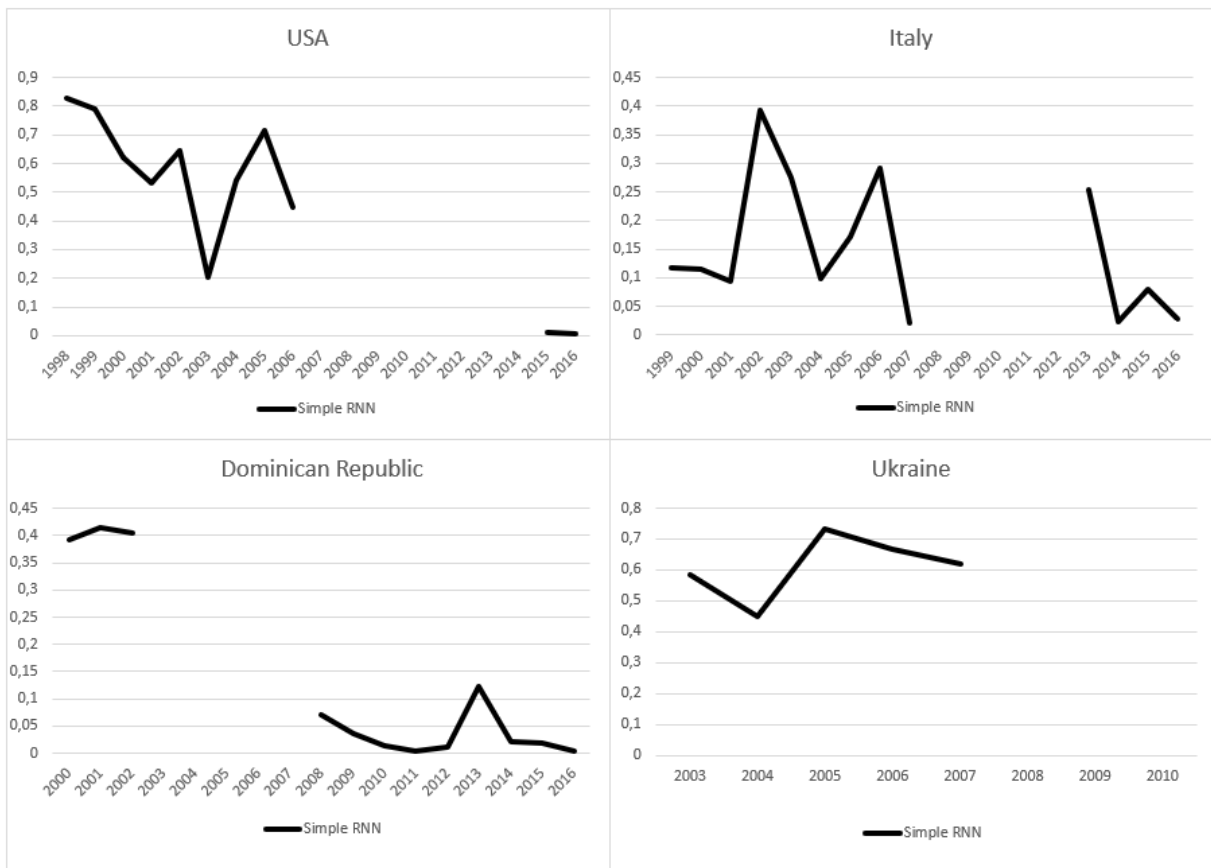


Figure 8.9: Predicted crisis probabilities by country, Simple RNN

Comparing the graphs to the result of the experiments involving multiple lags, the prediction values in high-income countries display an erratic behaviour in the three – four years preceding the event. A clear indication of what is coming is still absent, but it is an improvement compared to the constantly decreasing predicted probability signalled by previous models. Furthermore, in the USA, it is reassuring seeing that the signal is constantly above warning levels (0.15) in the three years preceding the GFC. Again, high predicted probabilities in the USA (above 80%) in 1998 and 1999 might be due to rising Dot-com bubble. Regarding Dominican Republic and Ukraine, predicted probabilities are much higher above warning level in the period preceding the crisis, compared to the results achieved in the framework in which the less sophisticated machine learning algorithms were implemented with multiple lags.

A last analysis is conducted with the same framework using the LSTM Recurrent Neural Network type. Results are displayed below.

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
SimpleRNN	0,7280	0,0598	0,373	0,10	0,58	0,08	0,80	0,57	0,14

Table 8.10: Performance score, LSTM

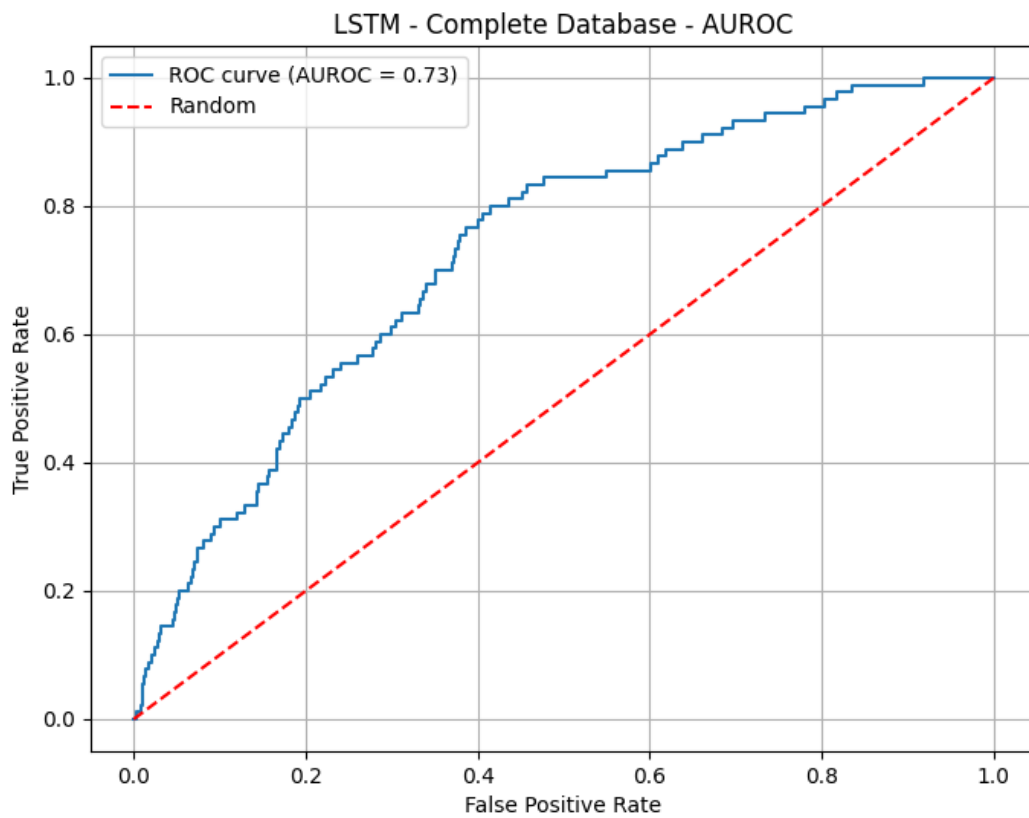


Figure 8.11: AUROC, Simple LSTM

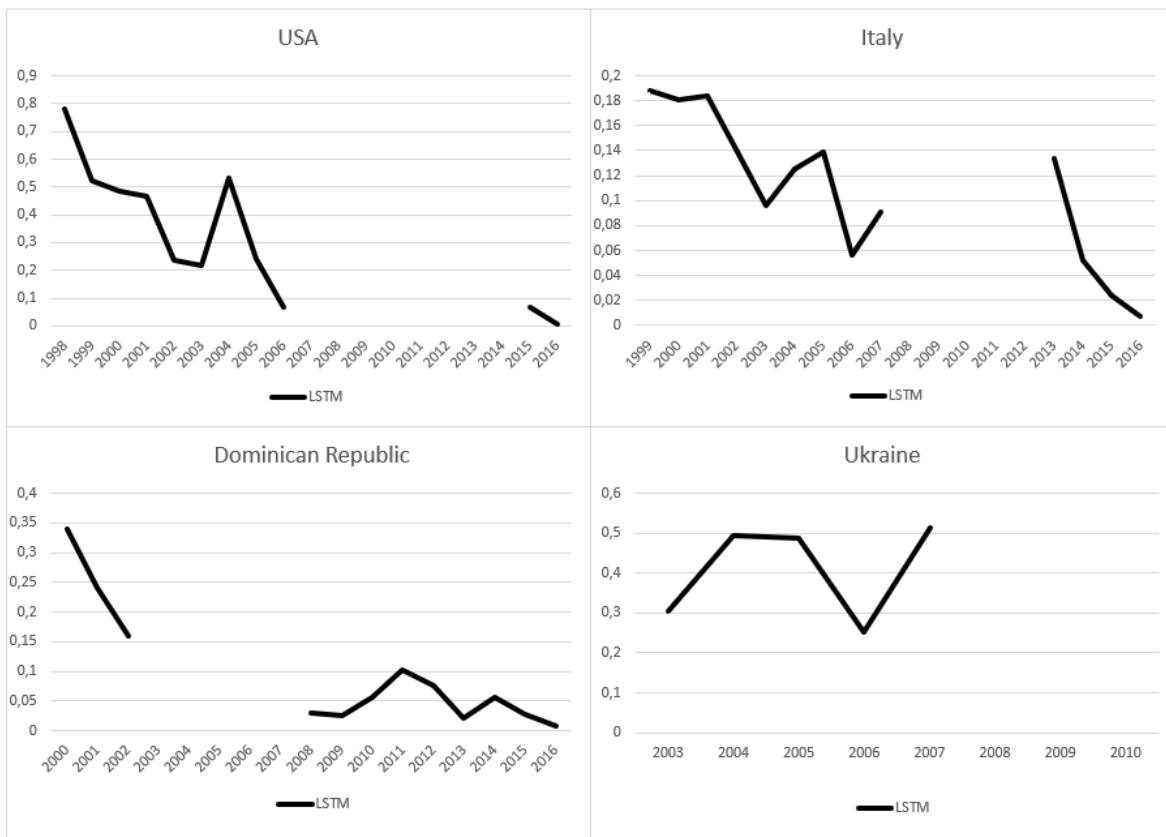


Figure 8.12: Predicted crisis probabilities by country, LSTM

The deterioration in performance compared to the Simple RNN model is clear, resulting both in decreased score measures and misleading signals for both income-level countries. Except for Ukraine, predicted probabilities have a decreasing trend, which could be wrongly perceived as reassuring by the forecaster. Overall, it is possible to say that when dealing with multiple lags of the same variables, in this specific framework, a Simple RNN could bring tangible improvement over the other machine learning methods examined, and that LSTM is not suitable for the same purpose being specifically engineered to work with much longer time series. The effectiveness of Simple RNN and its better performance compared to MLP is highlighted as well by Tölö and Eero (2019), who state that the structure of RNN counteract the harmful effects of overfitting.

## 2023 - 2025 PREDICTIONS

Given the results achieved in the previous chapters, the author of this Thesis applies the most promising strategies to observations up to 2022, the most recent year for which data is available for some countries, to seek what the models signal for the upcoming years 2023, 2024, and 2025. The testing panel is limited to the US, Italy, and China. Italy is chosen based on the author's nationality, while USA and China are chosen due to their importance in the global economy and their recent 2023 woes caused by the failure of SVB and other smaller regional banks in America, and filing for bankruptcy of Evergrande, China's second largest property developer. The models are trained with data up to 2016 using the crises database available up to 2019, and the testing dataset consists of years 2016 to 2022 observations in order to evaluate the trend of the predictions. The frameworks used for predictions are the one used in the baseline experiment, and the Simple RNN model, in this order. The author also computes the threshold which minimise the Loss function for each model in the whole training period, so to have a reference value for assessment.

The results from the baseline framework are showed in *Figure 9.1*.

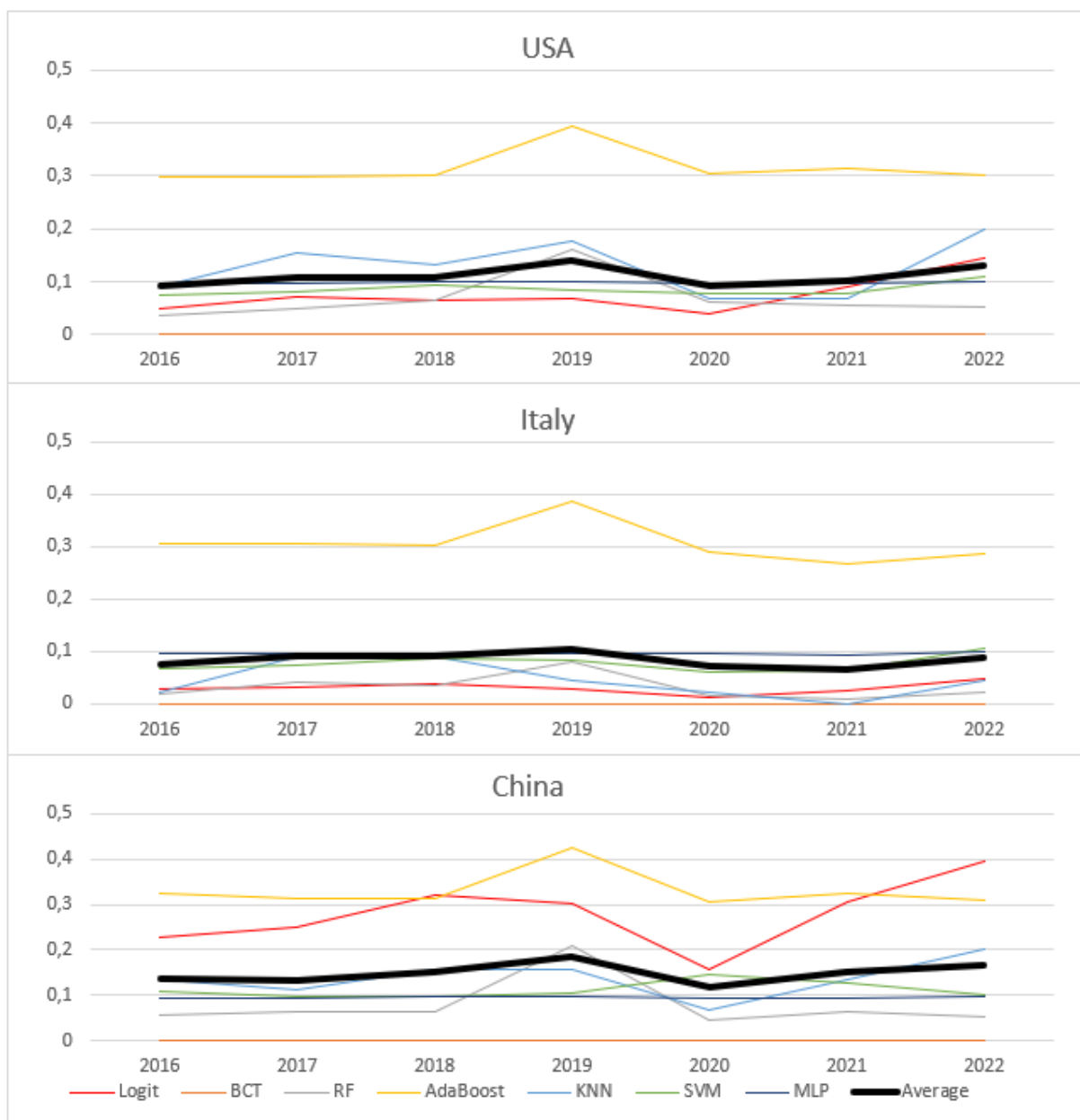


Figure 9.1: Predicted crisis probabilities by country, Baseline framework

All three countries follow a similar pattern in the considered period. The signals are quite flat, with just a marginal upward trend starting in 2020. Predicted probabilities for crisis in the Chinese economy are slightly higher compared to the other two western countries. The models which signal a coming crisis in future years are Logit and KNN both in the US and China (thresholds 0.08 and 0.14). Logit especially signals a 0.40 probability for China being in a pre-crisis state in 2022. SVM and MLP give a timid warning for all three countries (thresholds 0.08 and 0.10 respectively).

Moving to the Simple RNN analysis, the following predictions are computed.

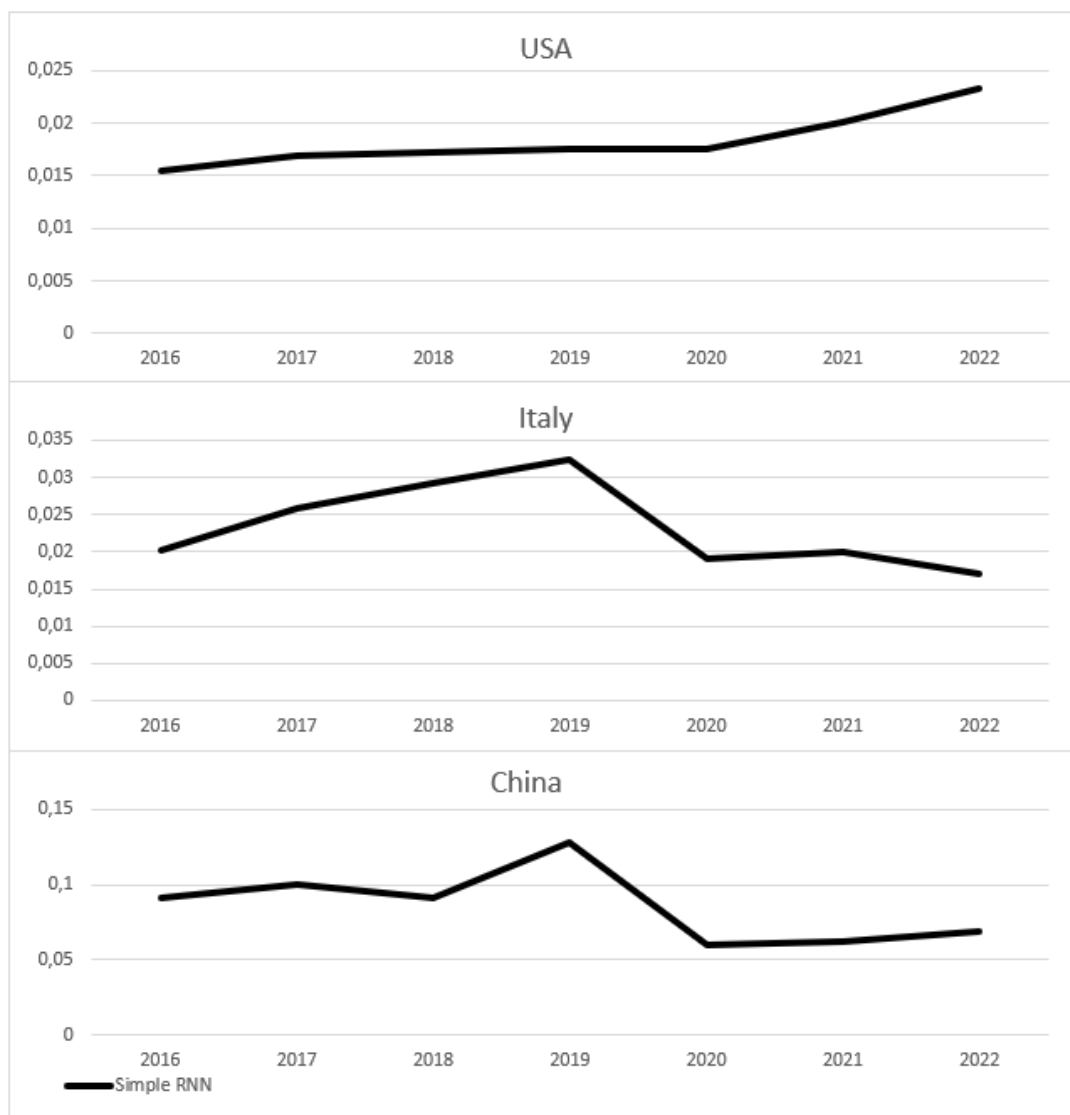


Figure 9.2: Predicted crisis probabilities by country, Simple RNN

The more advanced algorithm presents a similar picture for China, but a quite different one for western economies. The model clearly signals an increasing warning for the USA after 2020, while in Italy predicted probabilities are decreasing but still more than double respect to the baseline setting prediction. Warning levels in China are noticeably higher than in the western economies. Running the Simple RNN model using the whole training database, the threshold minimising the Loss Function is 0.11, implying that the model does not give any warning sign for coming years. Only China crossed this threshold in 2019, which may have been an early warning about the growing bubble in the real-estate sector, leading to the 2023 bankruptcy of Evergrande, or more simply, just a false alarm.



# CONCLUSIONS

This dissertation accomplished an extensive review of the literature on the subject, an explanation of the most widely implemented machine learning models and variables' effects, and then proceeded by testing different methodologies and different subsets of the available databases. A great focus is put on the attempt to simulate a continuous implementation of the models in a real-world environment, so to mimic in the most realistic way the results that would have been achieved. This implies that data used for testing must be strictly separated from the training datasets. The author achieves this by not including observations in the training subsets until the outcome (pre-crisis or tranquil period) is confirmed after the third year, by choosing the best hyperparameters with cross-validation using only the testing dataset, applying standardisation, winsorizing, and de-trending in a backward-looking fashion, and repeating all these passages for each testing year when new information is made available to the forecaster.

The results show that, using different parameters for evaluation and focussing mainly on AUROC and the ability to predict specific events in a four-countries panel, some models or methodology have an advantage over the others. In particular, as previously noticed by other authors such as Beutel et al. (2018) and Fricke (2017), in a strictly out-of-sample environment Logit Regression is clearly amongst the most reliable methods. These authors critique some results of the past literature and explain the over optimistic performance assigned to machine learning models with the evaluation techniques used, in particular cross-validation. When a strict separation of training and testing data is applied taking in consideration the temporal sequence of data, as in a real-world application, these authors state that machine learning's performance greatly decays. The author of this dissertation finds this true, but mostly for the high-income countries, and not so much for the low-income dataset. In fact, the author of this Thesis speculates that this might be a direct consequences of crises distribution in the database, with a single event, the GFC (which impacted mostly advanced economies), constituting the origin of most crises in the testing subset. Other than Logit, AdaBoost is also able to achieve good results on average, and correctly calls for warning most of the times in the four-countries panel testing. When adding lags to the predictors, it is clear how the standard machine learning methods are surpassed by the more advanced Simple RNN, which is specifically engineered to deal with long and complex time series. This model achieves the highest out-of-sample AUROC score and achieves above-average prediction capabilities in the four-countries panel sample. Logit can however still deal successfully with multiple lags and reach reasonable predictions,

reinforcing the opinion of Beutel et al. (2018), according to which machine learning algorithms tend to overfit on training data, while the Logit Regression retains sufficient flexibility.

Overall, the author of this research reaches the conclusion that the best models are either the simplest and most general, such as Logit, or the most complex ones, such as Recurrent Neural Networks with lags and AdaBoost, with the less sophisticated machine learning algorithms such as BCT and KNN being stuck in the middle and reaching non-satisfactory performance. A strong limit of these models is their apparent inability to predict events which do not follow similar patterns to crises of the past. This is expected, given the basic functioning of these algorithms which learn and repeat from a finite dataset. However, it is somewhat disappointing to see how a clear consensus of the models regarding the burst of the GFC, a credit-fuelled bubble, is missing in advanced economies. On the other hand, it is important to add that models' flawed performance could also be the fault of lacking data availability which, as stated by many other researchers, could impose significant restrictions on analysis. Having at disposal data about house prices and derivatives diffusion in each country could have made possible a more meaningful training of the models from similar, preceding crises and prompted better and clearer early warnings before the Great Financial Crisis, given the importance of those two predictors in the evolution of that event.

Regarding models' readability, Logit coefficients are mostly adherent to what the author expected and what the literature suggests. On the other hand, trying to understand how the single variable contributes to the final prediction in each ML algorithm turns out to be a much harder task, with results being very heterogeneous among different models and throughout the testing period, and coefficients often having signs opposite to what expected from widespread opinions. This has the adverse effect of rendering the adoption of the more advanced models by policymakers difficult, due to the lack of clear interpretability and causal effect of predictors.

To conclude, the author argues that a real-world implementation of the best forecasting strategies could not bring a magic sphere level of prediction, but that it would nonetheless still be able to at least spur a discussion and warn policymakers about current and future financial conditions. For example, results obtained from data up to 2022 could induce further analysis in the US and China, given the signal obtained from multiple models, and could also be reinforced by recent events in those countries. The quantitative analyses carried out with the methodologies presented in this Thesis should not be used as the sole warning indicators and should not in any case substitute an in-depth qualitative analysis.

# BIBLIOGRAPHY

- Aghion, Philippe & Bacchetta, Philippe & Banerjee, Abhijit (2001). "Currency crises and monetary policy in an economy with credit constraints," *European Economic Review*, Elsevier, vol. 45(7), pages 1121-1150.
- Alessi, Lucia and Detken, Carsten, (2011), Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity, *European Journal of Political Economy*, 27, issue 3, p. 520-533.
- Alessi, L., Antunes, A., Babecký, J., Baltussen, S., Behn, M., Bonfim, D., ... & Zigrainova, D. (2015). Comparing different early warning systems: Results from a horse race competition among members of the macro-prudential research network.
- Alessi, Lucia and Detken, Carsten, (2018), Identifying excessive credit growth and leverage, *Journal of Financial Stability*, 35, issue C, p. 215-225.
- Arteta, Carlos & Kamin, Steven Brian & Ruch, Franz Ulrich (2022). "How Do Rising U.S. Interest Rates Affect Emerging and Developing Economies? It Depends," Policy Research Working Paper Series 10258, The World Bank.
- Babecký, J., Havránek, T., Matějů, J., Rusnák, M., Šmídová, K. and Vašíček, B. (2012). Leading Indicators of Crisis Incidence: Evidence from Developed Countries. Czech National Bank, mimeo.
- Babecký Jan, Havránek Tomáš, Matějů Jakub, Rusnák Marek, Šmídková Kateřina, Vašíček Bořek (2014). Banking, debt, and currency crises in developed countries: Stylized facts and early warning indicators, *Journal of Financial Stability*, Volume 15, 2014, Pages 1-17, ISSN 1572-3089.
- Beutel, J., List, S., & Von Schweinitz, G. (2018). An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions?.
- Betrán, C., & Pons, M. A. (2013). Understanding Spanish Financial crises, 1850-2000: What determined their severity? (No. 48). EHES Working Papers in Economic History.
- Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Şimşek, Ö. (2023). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics*, 103773.
- Boissay, F., Collard, F., & Smets, F. (2016). Booms and banking crises. *Journal of Political Economy*, 124(2), 489-538.
- Borio, Claudio and Lowe, Philip, (2002), Assessing the risk of banking crises, *BIS Quarterly Review*, issue.
- Breiman, L., & Ihaka, R. (1984). Nonlinear discriminant analysis via scaling and ACE. Davis One Shields Avenue Davis, CA, USA: Department of Statistics, University of California.
- Brier, G.W. (1950) Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78, 1-3.
- Calvo, G. A., Leiderman, L., & Reinhart, C. M. (1993). Capital inflows and real exchange rate appreciation in Latin America: the role of external factors. *Staff Papers*, 40(1), 108-151.
- Calvo, Carmen, Guillermo, and Reinhart (2000), When Capital Inflows Come to a Sudden Stop: Consequences and Policy Options, MPRA Paper, University Library of Munich, Germany.
- Caprio, G. and Klingebiel, D. (2003) Episodes of Systemic and Borderline Financial Crises. In: Klingebiel, D., Ed., The World Bank, Washington DC.
- Carstens, A.G., Hardy, D.C., Pazarbasioglu, C. (2004). Avoiding Banking Crises in Latin America. Finance and Development. (September). International Monetary Fund, Washington, DC.

- Casabianca, E. J., Catalano, M., Forni, L., Giarda, E., & Passeri, S. (2022). A machine learning approach to rank the determinants of banking crises over time and across countries. *Journal of International Money and Finance*, 129, 102739.
- Cortes, C., Vapnik, V. Support-vector networks. *Mach Learn* 20, 273–297 (1995). <https://doi.org/10.1007/BF00994018>.
- Cover, T.M. and Hart, P.E. (1967) Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13, 21-27.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems* 2, 303–314.
- Dell’Ariccia, Giovanni, Luc Laeven, and Gustavo A. Suarez. (2017). Bank Leverage and Monetary Policy’s Risk-Taking Channel: Evidence from the United States. *Journal of Finance* 72(2): 613–654.
- Demirgüç-Kunt, A., & Detragiache, E. (1998). The Determinants of Banking Crises in Developing and Developed Countries. *Staff Papers (International Monetary Fund)*, 45(1), 81–109.
- du Plessis, Emile (2022). Multinomial Modeling Methods: Predicting Four Decades of International Banking Crises (March 27, 2022). *Economic Systems*, Vol. 46, No. 2, 2022.
- Duttagupta, Rupa and Cashin, Paul, (2011), Anatomy of banking crises in developing and emerging market countries, *Journal of International Money and Finance*, 30, issue 2, p. 354-376.
- Eichengreen Barry & Rose Andrew K., 1998. "Staying Afloat When the Wind Shifts: External Factors and Emerging-Market Banking Crises," NBER Working Papers 6370, National Bureau of Economic Research, Inc.
- Estrella, A., & Hardouvelis, G. A. (1991). The term structure as a predictor of real economic activity. *The journal of Finance*, 46(2), 555-576.
- Fix, E. and Hodges, J.L. (1951) Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field.
- Fouliard, J., Howell, M., & Rey, H. (2021). Answering the queen: Machine learning and financial crises (No. w28302). National Bureau of Economic Research.
- Fouliard, Jérémy, Héléne Rey, and Vania Stavrageva (2021) “Is this Time Different? Financial Follies across Centuries,” London Business School and CEPR.
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *icml* (Vol. 96, pp. 148-156).
- Fricke, Daniel (2017). *Financial Crisis Prediction: A Model Comparison* (November 29, 2017).
- Garson, G.D. (1991) Interpreting Neural Network Connection Weights. *AI Expert*, 6, 47-51.
- Goh, A. (1995), Back-propagation Neural Networks for Modeling Complex Systems, *Artificial Intelligence in Engineering* 9, 143–151.
- Gorton, G., & Ordonez, G. (2020). Good booms, bad booms. *Journal of the European Economic Association*, 18(2), 618-665.
- Grimm, M., Jordà, Ò., Schularick, M., & Taylor, A. M. (2023). Loose monetary policy and financial instability (No. w30958). National Bureau of Economic Research.
- Hanke, Steve. (2004). *The Dominican Republic Resolving the Banking Crisis and Restoring Growth*.
- Hardy, Daniel and Pazarbasioglu, Ceyla (1998). Leading Indicators of Banking Crises: Was Asia Different? (June 1998). *IMF Working Paper No. 98/91*.
- Holopainen, Markus and Sarlin, Peter, (2017), Toward robust early-warning models: a horse race, ensembles and model uncertainty, *Quantitative Finance*, 17, issue 12, p. 1933-1963.

- Iacoviello, M., & Navarro, G. (2019). Foreign effects of higher US interest rates. *Journal of International Money and Finance*, 95, 232-250.
- Janiesch Christian & Zscheck Patrick & Heinrich Kai (2021). "Machine learning and deep learning," *Electronic Markets*, Springer;IIM University of St. Gallen, vol. 31(3), pages 685-695, September.
- Jiang Janet Hua (2008). "Banking crises in monetary economies," *Canadian Journal of Economics*, Canadian Economics Association, vol. 41(1), pages 80-104, February.
- Jordà, Ò., Schularick, M., and Taylor, A. M. (2015), Leveraged Bubbles, *Journal of Monetary Economics* 76, S1-S20.
- Jordà Ò., Richter, B., Schularick, M., and Taylor, A.M. (2017), "Macrofinancial History and the New Business Cycle Facts", *NBER Macroeconomics Annual 2016*, Vol. 31, pp. 213-263.
- Joy, M., Rusnák, M., Šmídková, K., & Vašíček, B. (2017). Banking and currency crises: Differential diagnostics for developed countries. *International Journal of Finance & Economics*, 22(1), 44-67.
- Kaminsky, G. L., & Reinhart, C. M. (1999). The Twin Crises: The Causes of Banking and Balance-Of-Payments Problems. *The American Economic Review*, 89(3), 473–500.
- Kaminsky, Graciela, (2006), Currency crises: Are they all the same?, *Journal of International Money and Finance*, 25, issue 3, p. 503-527.
- Knedlik, Tobias, (2013), The European Commission's Scoreboard of Macroeconomic Imbalances: The impact of preferences on an early warning system, *VfS Annual Conference 2013 (Duesseldorf): Competition Policy and Regulation in a Global Economic Order*, Verein für Socialpolitik / German Economic Association.
- Korduban, Pavel (2008). The Jamestown Foundation. Hard Times for Ukrainian Banks, Central Bank Chairman Under Fire, December 17, 2008. Retrieved from <https://jamestown.org/program/hard-times-for-ukrainian-banks-central-bank-chairman-under-fire/>.
- Laeven, Luc & Valencia, Fabian. (2008). Systemic Banking Crises: A New Database. *International Monetary Fund, IMF Working Papers*. 08. 10.5089/9781451870824.001.
- Laeven, L., Valencia, F. (2013). Systemic Banking Crises Database. *IMF Econ Rev* 61, 225–270.
- Laeven Luc & Valencia Fabian (2018). "Systemic Banking Crises Revisited," *IMF Working Papers 2018/206*, International Monetary Fund.
- Laeven Luc & Valencia Fabian (2020). "Systemic Banking Crises Database II," *IMF Economic Review*, Palgrave Macmillan;International Monetary Fund, vol. 68(2), pages 307-361, June.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015 May 28;521(7553):436-44. doi: 10.1038/nature14539. PMID: 26017442.
- Lindgren, Garcia, Saal. (1996). "Bank soundness and macroeconomic policy". Washington, D.C. : International Monetary Fund. ISBN / ISSN: 155775599X.
- Lo Duca Marco, Peltonen Tuomas A. (2013), Assessing systemic risks and predicting systemic events, *Journal of Banking & Finance*, Volume 37, Issue 7, 2013, Pages 2183-2195, ISSN 0378-4266.
- McCulloch, W.S., Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5, 115–133.
- Neunhoeffler, M., & Sternberg, S. (2019). How cross-validation can go wrong and what to do about it. *Political Analysis*, 27(1), 101-106.
- Nguyen Thanh Cong, Castro Vítor, Wood Justine, (2022). A new comprehensive database of financial crises: Identification, frequency, and duration, *Economic Modelling*, Volume 108, 2022, 105770, ISSN 0264-9993.

Obstfeld, Maurice, (2012), Does the Current Account Still Matter?, No 17877, NBER Working Papers, National Bureau of Economic Research, Inc.

Quinlan, J. R. (1986). "Induction of Decision Trees." *Machine Learning*, 1(1), 81-106.

Reinhart Carmen M. & Rogoff Kenneth S. (2008). "This Time is Different: A Panoramic View of Eight Centuries of Financial Crises," NBER Working Papers 13882, National Bureau of Economic Research, Inc.

Ristolainen, Kim, (2018), Predicting Banking Crises with Artificial Neural Networks: The Role of Nonlinearity and Heterogeneity, *Scandinavian Journal of Economics*, 120, issue 1, p. 31-62.

Sarlin, Peter and Peltonen, Tuomas, (2013), Mapping the state of financial stability, *Journal of International Financial Markets, Institutions and Money*, 26, issue C, p. 46-76.

Schularick, M., & Taylor, A. M. (2012). Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870—2008. *The American Economic Review*, 102(2).

Shapley, L. (1953) A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., *Contributions to the Theory of Games II*, Princeton University Press, Princeton, 307-317.

Töölö, Eero (2019). "Predicting systemic financial crises with recurrent neural networks," *Bank of Finland Research Discussion Papers 14/2019*, Bank of Finland.

Turing, A.M. (1950) *Computing Machinery and Intelligence*. *Mind*, 59, 433-460.

Werbos Paul J. (1988), Generalization of backpropagation with application to a recurrent gas market model, *Neural Networks*, Volume 1, Issue 4, 1988, Pages 339-356, ISSN 0893-6080.

Wicksell, Knut. (1898). *Geldzins und Güterpreise. Eine Studie über die den Tauschwert des Geldes bestimmenden Ursachen*. Jena: Verlag von Gustav Fischer.

Wright, J. H. (2006). *The yield curve and predicting recessions*.

# APPENDIX

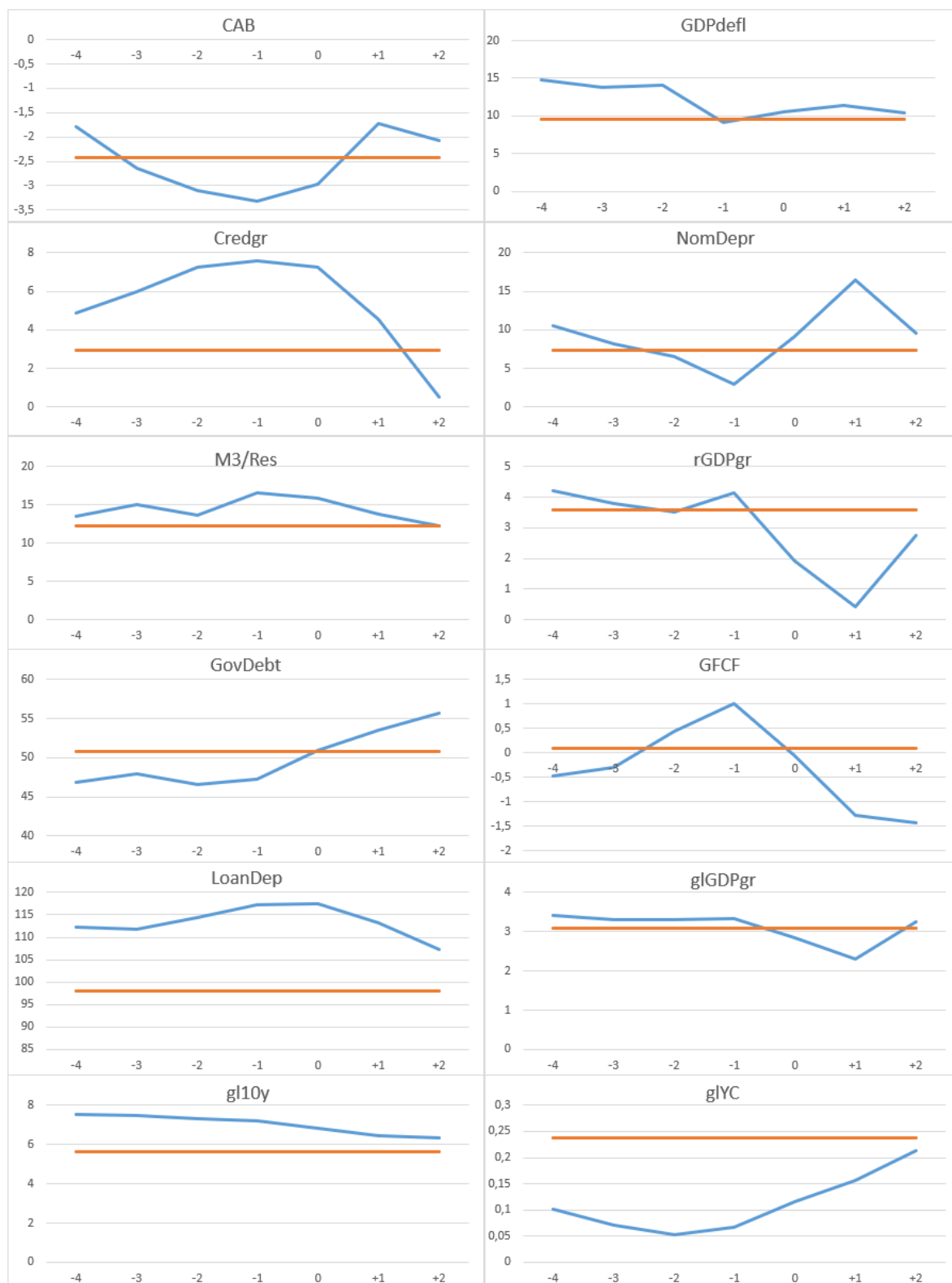


Figure A.1: Signals, overall database 1970-2019



Figure A.2: Signals, developing economies 1970-2019





Figure A.3: Signals, advanced economies 1970-2019

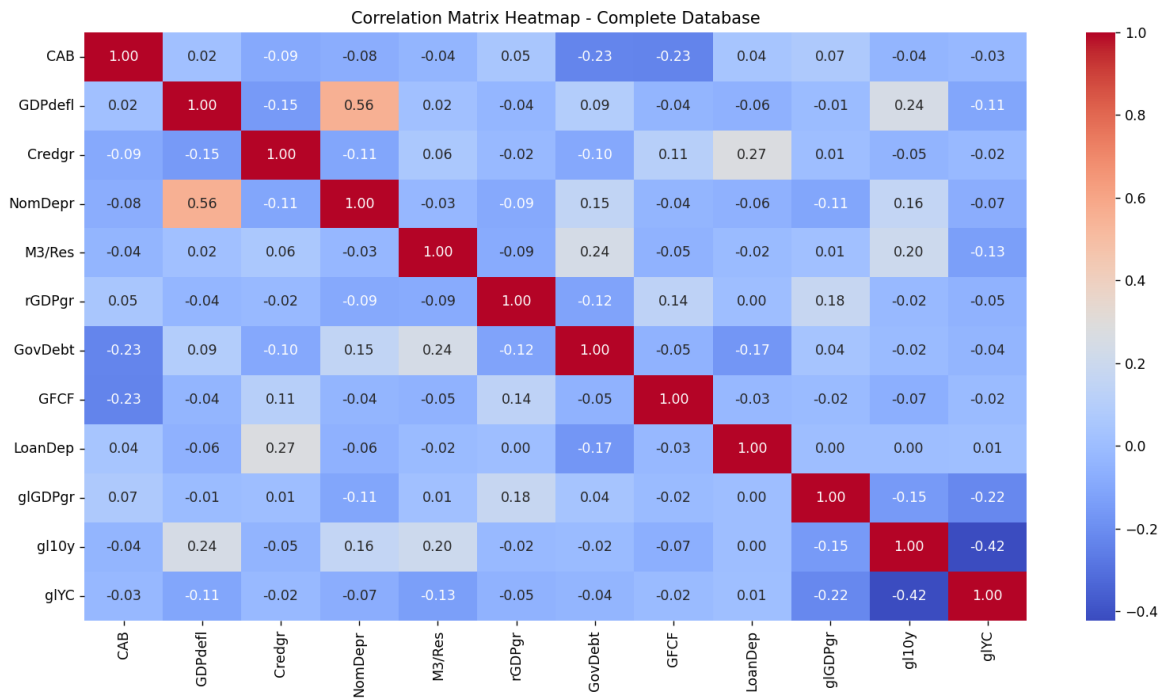


Figure A.4: Correlation matrix, overall database

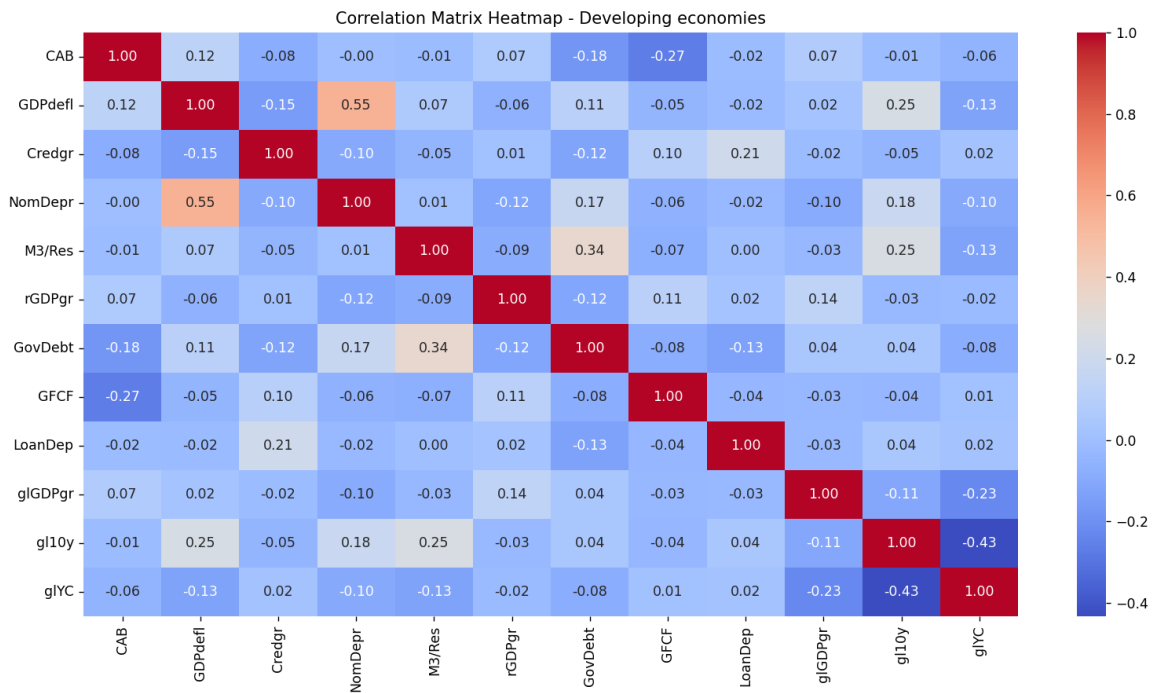
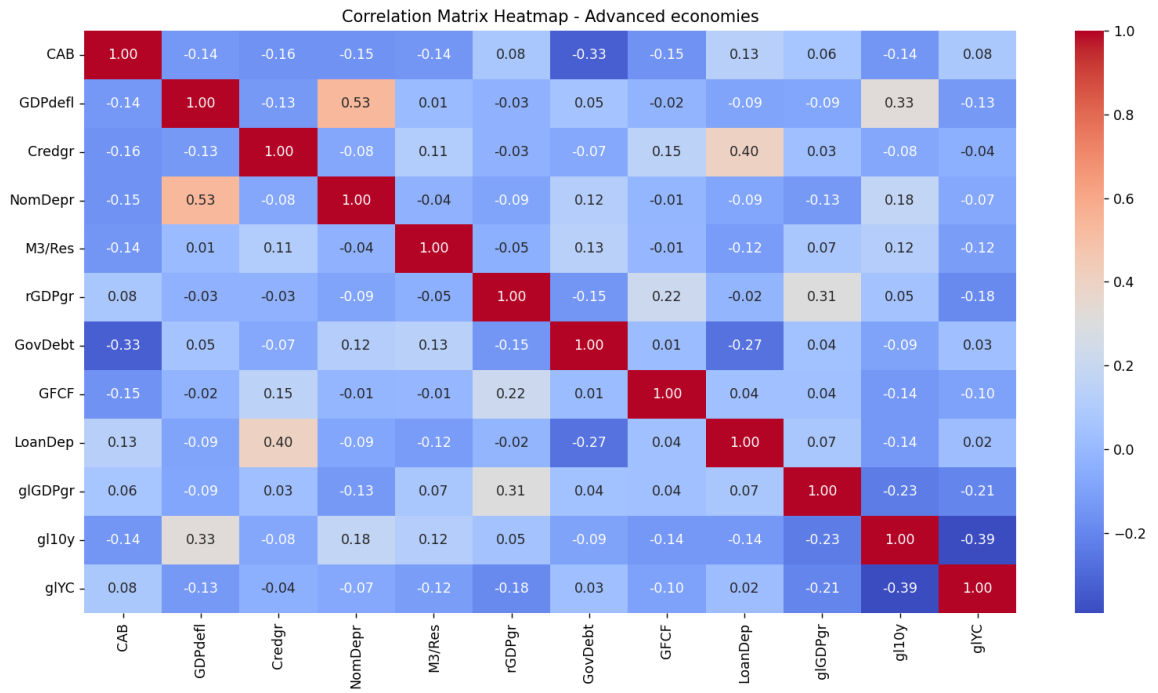


Figure A.5: Correlation matrix, developing economies



*Figure A.7: Correlation matrix, advanced economies*

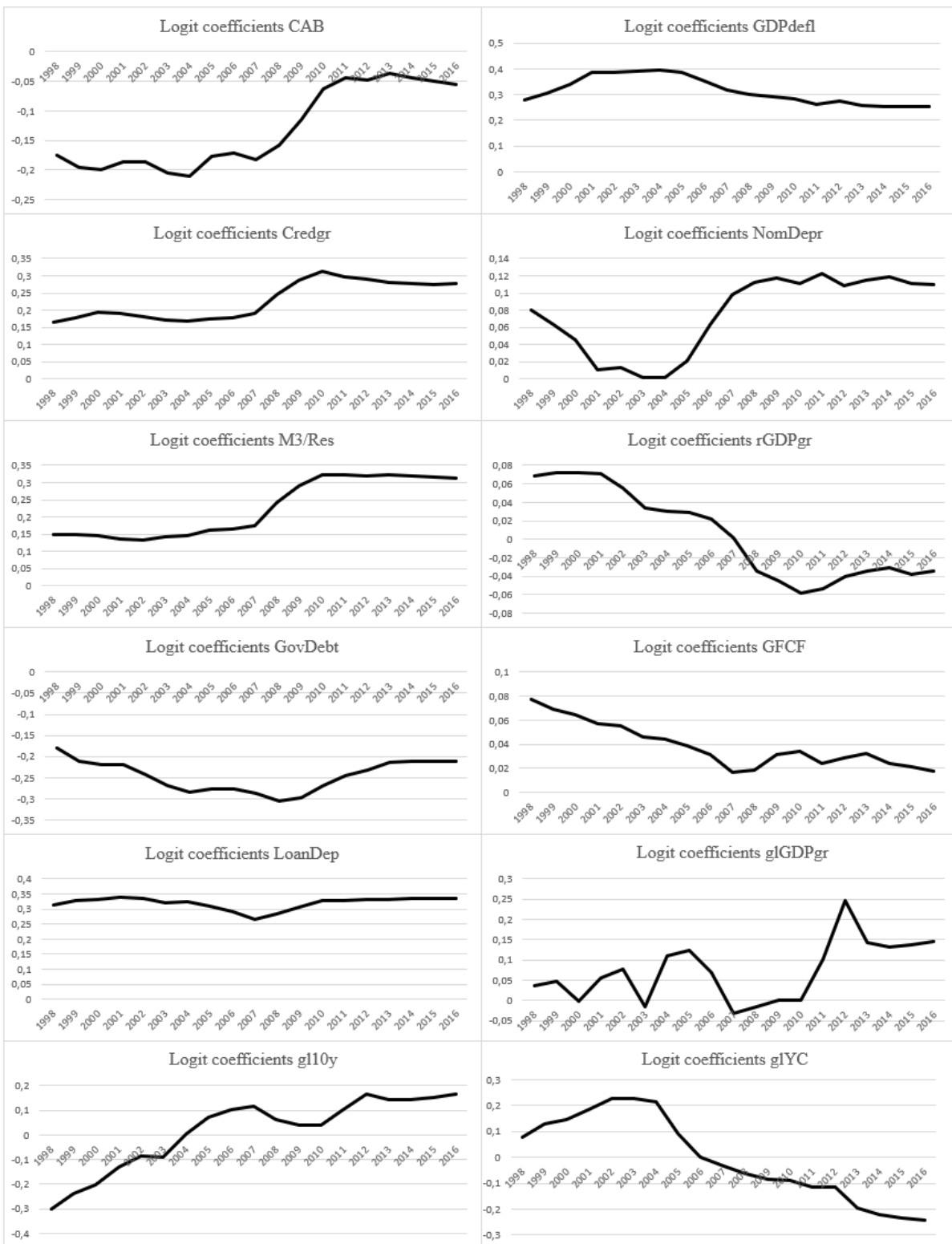


Figure A.8: Logit coefficients over time, overall database

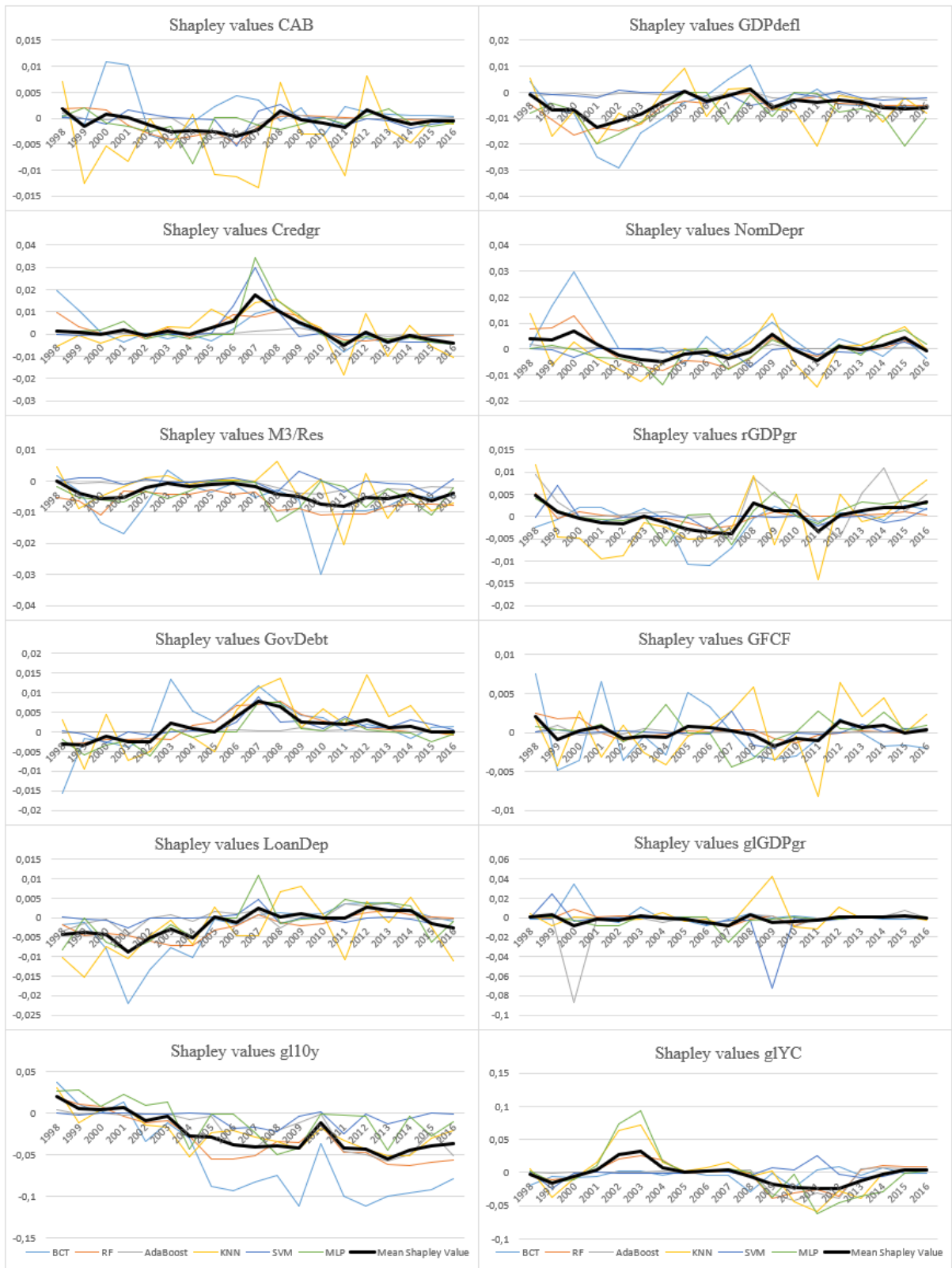


Figure A.9: Shapley Values over time, overall database

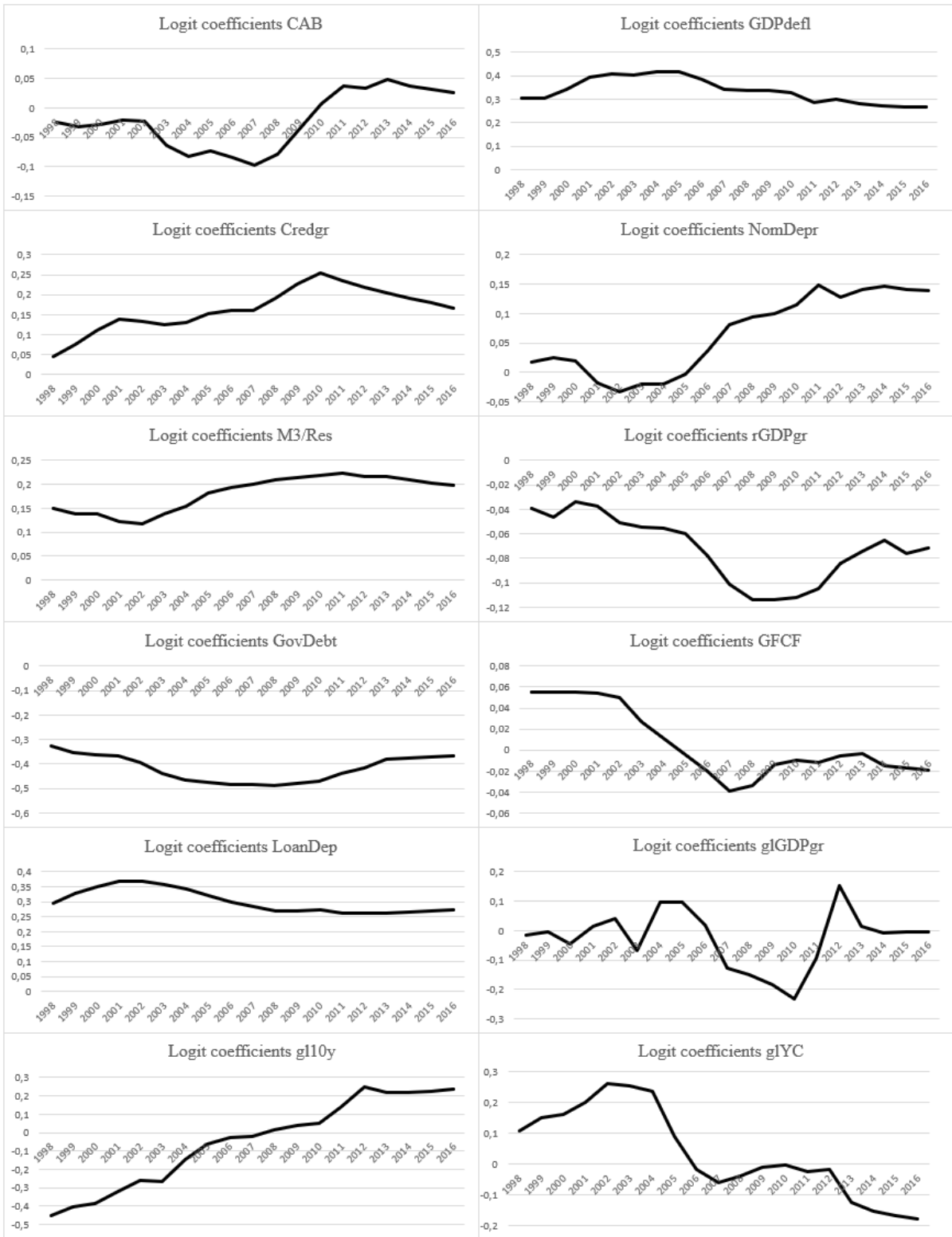


Figure A.10: Logit coefficients over time, developing economies

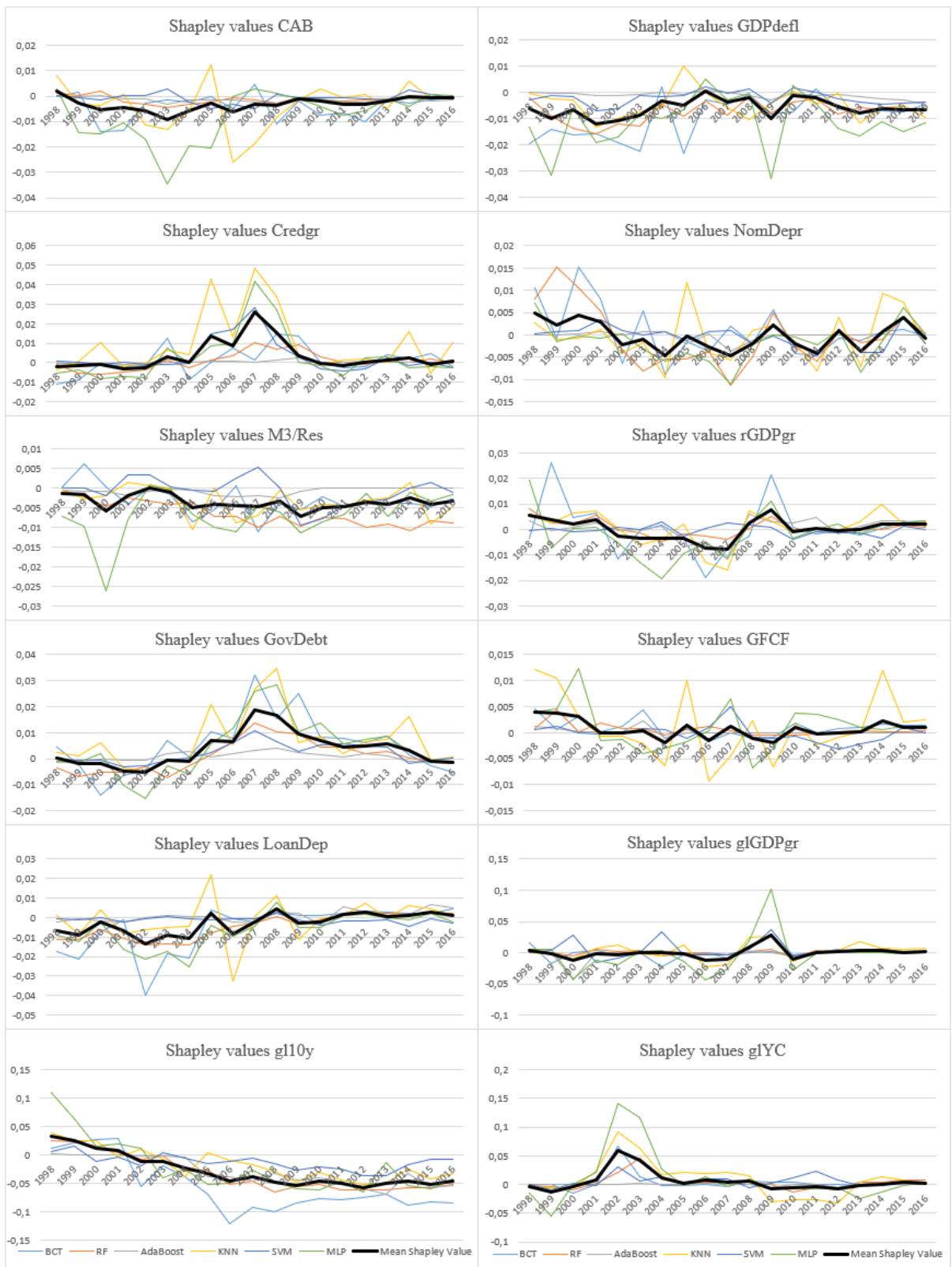


Figure A.11: Shapley Values over time, developing economies

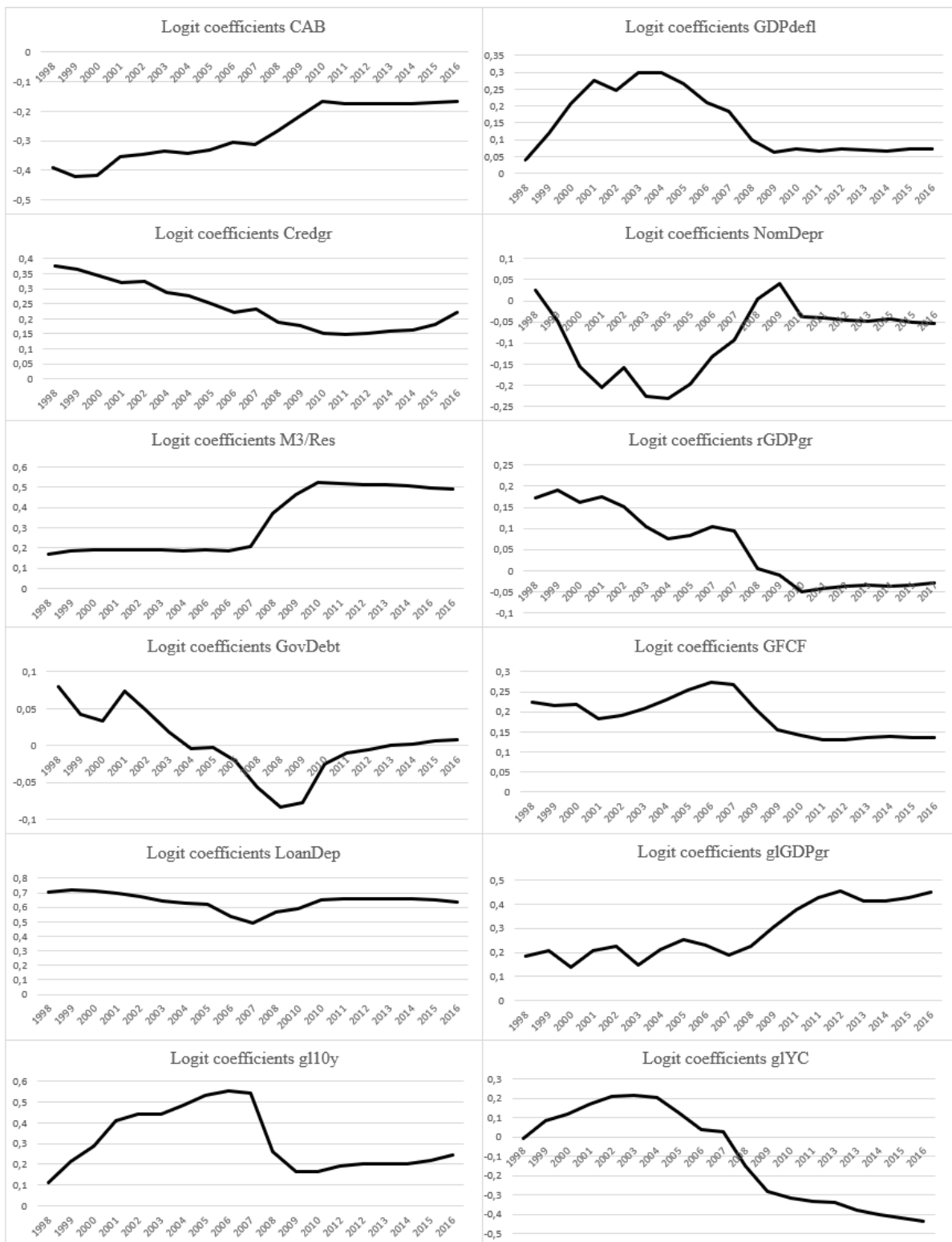


Figure A.12: Logit coefficients over time, advanced economies



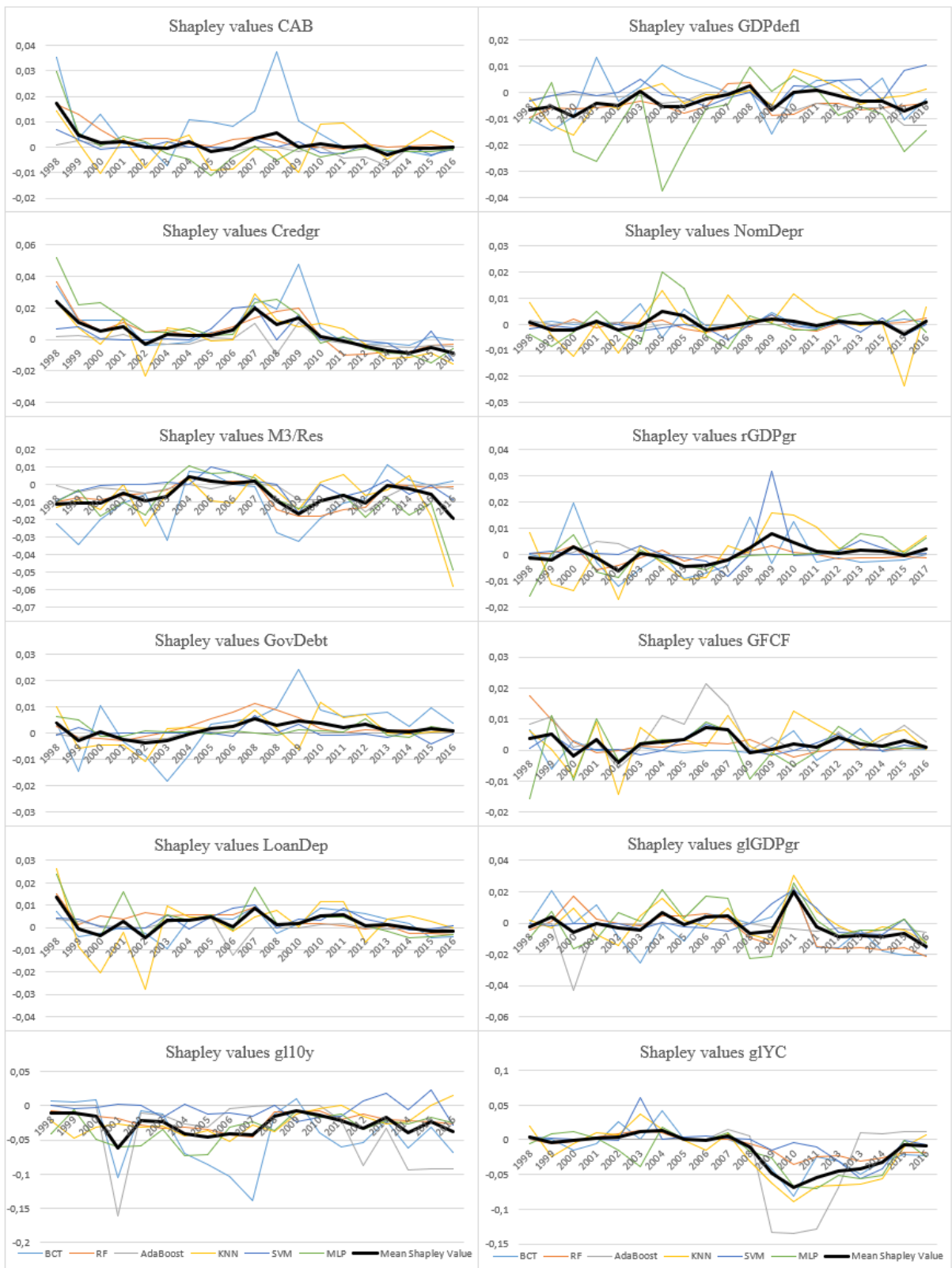


Figure A.13: Shapley Values over time, advanced economies

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,6541	0,1337	0,295	0,07	0,57	0,14	0,74	0,55	0,23
BCT	0,5294	0,2394	0,136	0,07	0,61	0,12	0,51	0,62	0,19
RF	0,6514	0,0893	0,344	0,09	0,58	0,15	0,78	0,56	0,25
AdaBoost	0,6808	0,1514	0,351	0,38	0,63	0,16	0,73	0,62	0,26
KNN	0,6981	0,0834	0,331	0,07	0,56	0,14	0,80	0,53	0,24
SVM	0,6561	0,1185	0,274	0,07	0,50	0,13	0,80	0,47	0,22
MLP	0,7053	0,1337	0,394	0,05	0,58	0,16	0,84	0,56	0,26

Table A.14: Performance score, cross-validation, developing economies

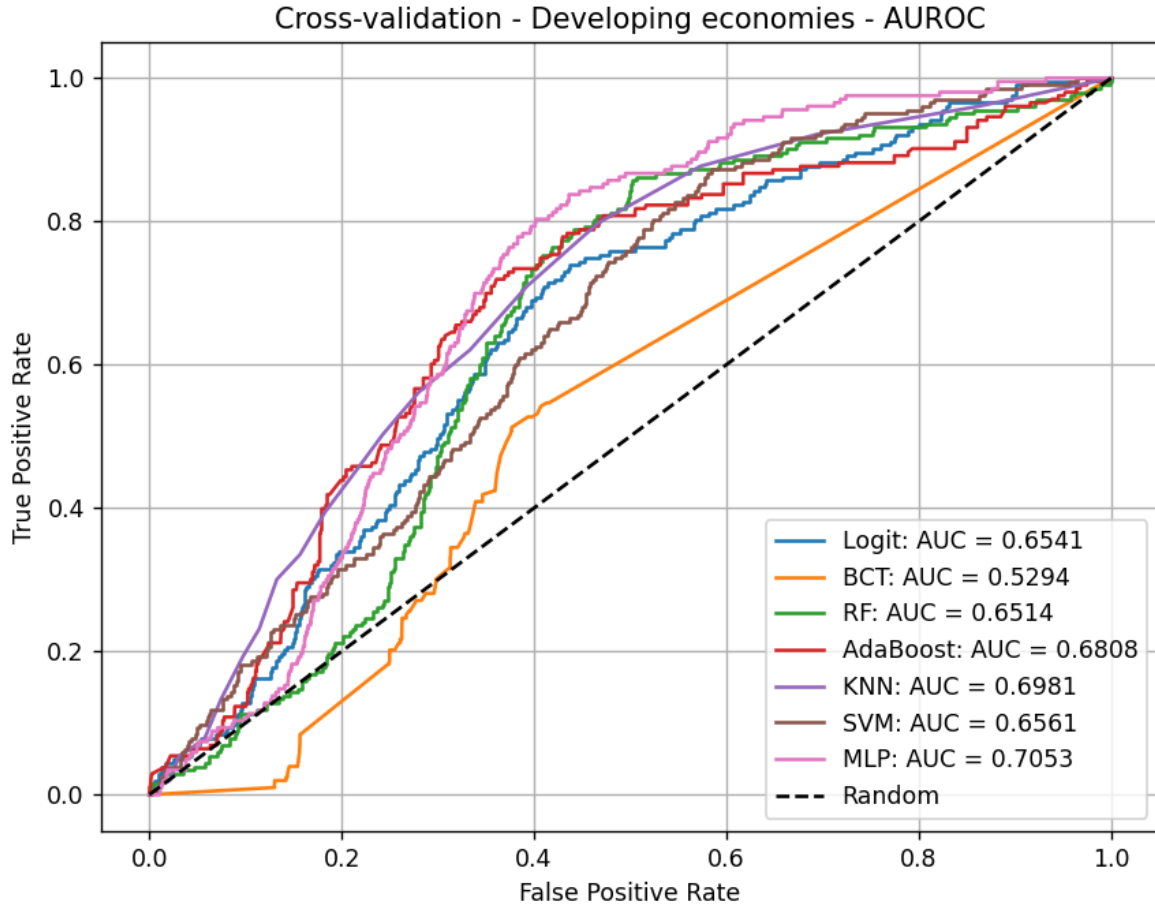


Figure A.15: AUROC, cross-validation, developing economies

Method	AUROC	Brier	Usefulness	Threshold	Accuracy	Precision	Recall	Specificity	F1-Score
Logit	0,7446	0,0898	0,387	0,09	0,64	0,18	0,76	0,63	0,29
BCT	0,6627	0,1951	0,338	0,01	0,59	0,16	0,76	0,57	0,27
RF	0,7894	0,0823	0,420	0,09	0,59	0,18	0,86	0,56	0,29
AdaBoost	0,7783	0,1340	0,421	0,41	0,77	0,24	0,64	0,78	0,35
KNN	0,7592	0,0823	0,384	0,12	0,65	0,18	0,75	0,64	0,29
SVM	0,7098	0,0959	0,320	0,13	0,66	0,17	0,66	0,66	0,28
MLP	0,7952	0,0897	0,451	0,12	0,76	0,25	0,68	0,77	0,36

Table A.16: Performance score, cross-validation, advanced economies

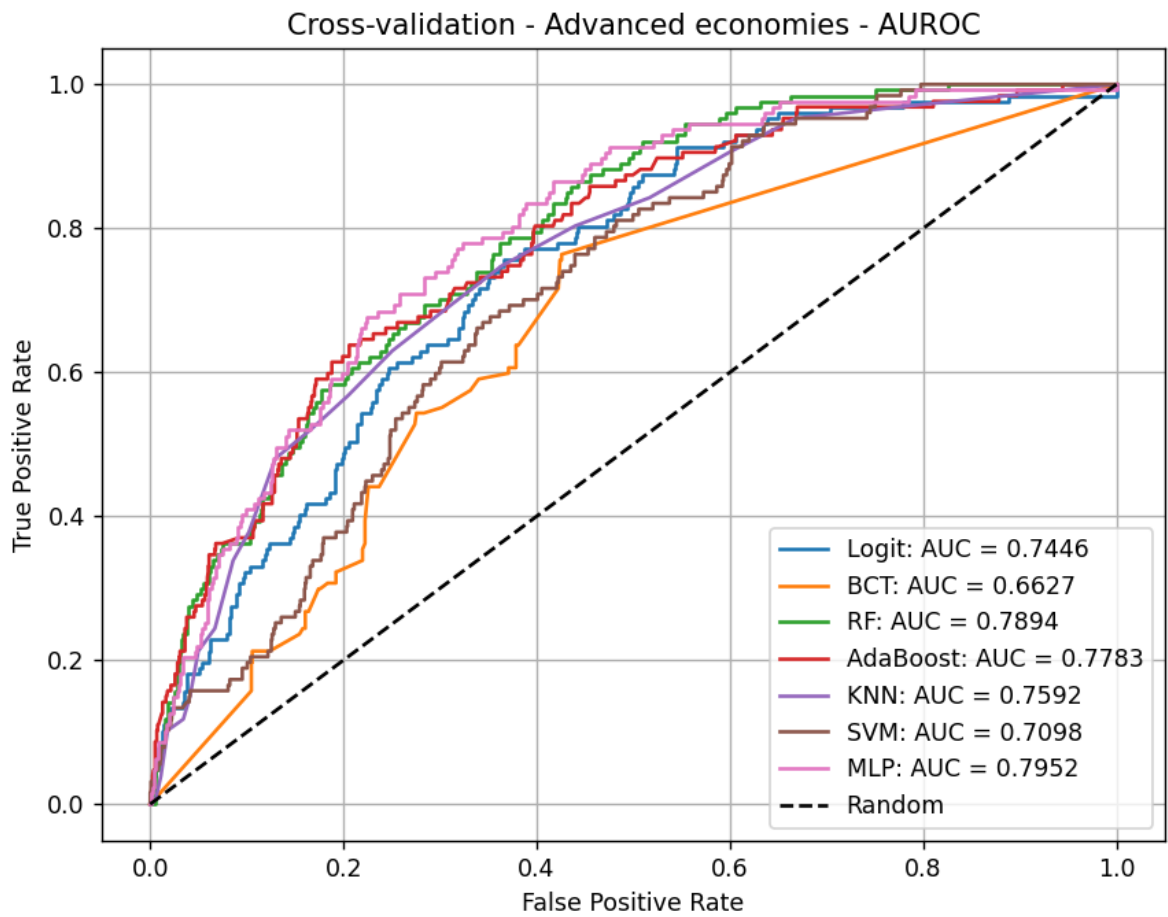


Figure A.17: AUROC, cross-validation, advanced economies

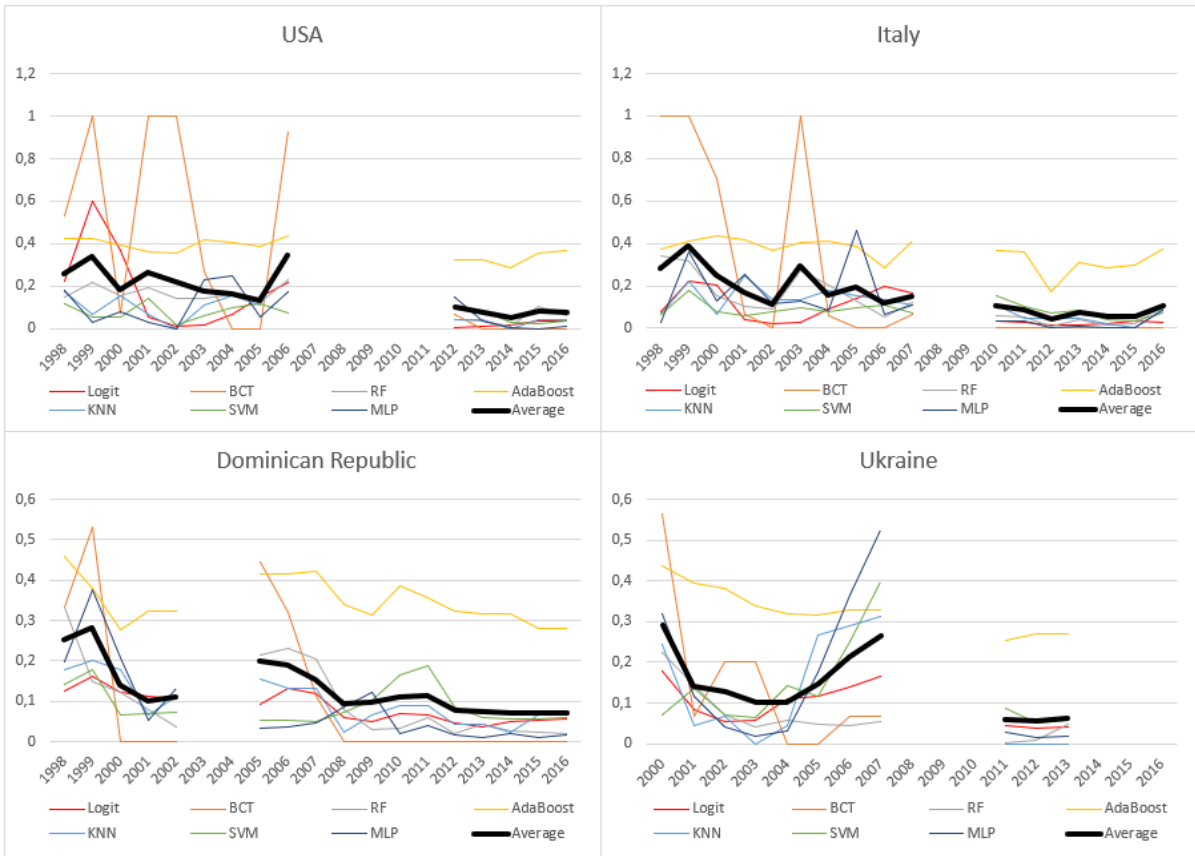


Figure A.18: Predicted crisis probabilities by country with cross-validation, database split in low and high-income countries