University of Warsaw
Faculty of Economic Sciences

Nijat Hasanli
Album N°: 456427

From Data to Decision: Leveraging Machine Learning for Crisis Response

Master degree thesis
Field of the study: Quantitative Finance

The thesis written under the supervision of
Dr Marcin Bielecki
from Faculty of Economic Sciences
WNE UW

Warsaw, June 2024

Declaration of the supervisor

I declare that the following thesis was written under my supervision and I state that it meets all criterias to be submitted for the procedure of academic degree award.

*I declare that my participation in the scientific article, which is a part of this thesis is ~~...~~.%, while the supplement to the thesis was written independently by the graduate (s).

** cross out if not applicable*

Date

05.06.2024

Signature of the Supervisor

*Declaration of the author of the thesis***

Aware of the legal responsibility, I declare that I am the sole author of the following thesis and that the thesis is free from any content that constitutes copyright infringement or has been acquired contrary to applicable laws and regulations.

I also declare that the thesis has never been a subject of degree-awarding procedures in any higher education institution.

Moreover I declare that the attached version of the thesis is identical with the enclosed electronic version.

* I declare that my*** participation in the scientific article, which is part of this thesis is ~~.....~~..% (not less than 60%), while the supplement to the thesis was written by me***.

** cross out if not applicable*

Date

05.06.2024

Signature of the Author

*** each of the co-authors of the thesis submit the statement separately.*
**** in the case of co-author of thesis, substantive and percentage contribution should be declared.*

Summary

This thesis concerns evaluating models in decision-making, particularly for early warning prediction of crises in policy contexts. It emphasizes understanding how selected metrics influence decisions by isolating models' effects from subjective considerations, clarifying their capabilities. This research covers findings on the performance of various modeling methods, highlighting strengths and weaknesses. This comparison is crucial for informed decision-making, noting key variables that consistently emerge as significant predictors across models. This study presents issues in model evaluation for policy contexts, discussing the role of selecting appropriate metrics and parameters. These choices impact decision-making effectiveness. The analysis stresses aligning model evaluation with policymakers' priorities for balanced, effective outcomes.

Key words

machine learning, crisis prediction, early warning, financial stability, economic indicators

Field of the thesis (codes according to the Erasmus program)

Economics (14300)

The title of the thesis in Polish

Od danych do decyzji: wykorzystanie uczenia maszynowego do reagowania kryzysowego

Contents

Abstract

This research examines financial-economic crises using diverse economic indicators across different countries. Distinguishing itself from conventional model comparisons, our study introduces a practical element – a graphical representation illustrating how policymakers' decisions, such as setting thresholds (τ) and importance levels (μ), influence model outcomes. This approach provides valuable insights for policymakers, banks, and governments, enhancing our understanding of model performances. Notably, the thesis compares traditional and novel methods, including the introduction of five boosting methods (Ada, GB, XGB, LGBM, CatB) alongside standard techniques like KNN, SVM, DT, and RF.

Motivation

Firstly, the primary motivation for this research lies in addressing the ongoing and severe impacts of financial-economic crises on global economies. These crises, often triggered by market changes, sector-specific disruptions, or unforeseen external shocks, lead to significant declines in economic activities and widespread financial instability. The consequences extend beyond the financial markets, affecting the livelihoods of individuals, businesses, and entire communities. Understanding the precursors to these crises is crucial for developing effective early warning models (EWMs) that can predict and mitigate their impacts.

Secondly, traditional approaches to early warning systems, while foundational, have their limitations. These methods have relied heavily on signaling techniques and statistical models like logistic regression. However, these conventional models sometimes fall short in accuracy and timeliness, which can result in inadequate preparation and response to impending crises. The advent of machine learning (ML) presents a transformative opportunity to significantly enhance the predictive accuracy of these models. By integrating advanced ML algorithms such as boosting methods (AdaBoost, Gradient Boosting, XGBoost, LightGBM, CatBoost) alongside traditional techniques like KNN, SVM, Decision Trees, and Random Forests, we aim to build more accurate and reliable models.

Thirdly, the potential impact of this research extends beyond academic contributions. It aims to provide policymakers, banks, and governments with robust tools that offer timely insights for proactive decision-making. By introducing a graphical representation of how thresholds (τ) and importance levels (μ) affect model outcomes, this study offers a practical guide to optimizing policy responses. This innovation empowers decision-makers to take preemptive actions, thereby reducing the likelihood and impact of financial crises.

Finally, this research seeks to make a meaningful difference by advancing the field of early warning systems. The goal is not just to enhance predictive accuracy but to provide tangible benefits that safeguard economies and promote financial stability. This endeavor aspires to turn cutting-edge technology into actionable intelligence that can prevent economic disruptions and foster a more resilient global financial system.

Introduction

Financial-economic crises occur when various industries encounter significant issues, often stemming from market changes, sector-specific problems, or unexpected external shocks. These crises entail a notable decline in economic activities, widespread turmoil in money markets, and extensive financial instability [Thakor, 2015]. Excessive indebtedness, speculative investments, banking failures, and unforeseen external disturbances are among the factors contributing to their genesis [Jickling, 2009, Mishkin, 1992]. Understanding these crises requires grasping economic, financial, and psychological elements, including risk-taking tendencies, regulatory deficiencies, and global economic imbalances [Baily et al., 2009]. Models are developed to help anticipate and potentially prevent these crises by identifying precursory signals, enabling preemptive measures, and mitigating risks [Thakor, 2015]. Swift action is crucial for restoring financial systems, protecting investors, and minimizing the profound impact of such crises on economies and communities.

As technological advancements continue to reshape various sectors, artificial intelligence (AI) applications, particularly those employing machine learning techniques, are gaining prominence in economics. This trend extends to the development of early warning models for financial crises, where machine learning algorithms offer promising avenues for improved predictive accuracy. Traditionally, early warning models relied heavily on signaling approaches and discrete statistical methods such as logistic regression [Kaminsky and Reinhart, 1999, Knedlik and Von Schweinitz, 2012]. However, recent years have witnessed a notable shift towards integrating machine learning techniques into this domain, with scholars advocating for their efficacy in enhancing predictive precision [Alessi and Detken, 2018, Tanaka et al., 2016]. Notably, random forests have emerged as a particularly promising tool, surpassing the effectiveness of conventional models [Holopainen and Sarlin, 2017].

In our approach, we adopt a hybrid methodology that combines traditional statistical methods with advanced machine learning algorithms. This includes logistic regression as a foundational analysis, supplemented by regularization techniques such as L1 or L2 regularization for enhanced model robustness. Additionally, we explore the efficacy of diverse machine learning algorithms, ranging from K-Nearest Neighbors and Support Vector Machines to Decision Trees and ensemble methods like AdaBoost and Gradient Boosting. Moreover, we incorporate state-of-the-art algorithms like XGBoost, LightGBM, and CatBoost. Our study utilizes an extensive panel dataset spanning 13 countries over 49 years, focusing on 10 main variables crucial for understanding economic dynamics. To deepen our analysis, we derive an additional set of 10 variables from the main ones, providing a comprehensive view of

the economic landscape. By employing both traditional statistical methods and cutting-edge machine learning approaches on a rich dataset, we aim to contribute to the advancement of early warning models in predicting and addressing economic crises.

Furthermore, we acknowledge the critical role of policymakers in shaping the effectiveness of early warning models. Policymakers must determine thresholds (τ) for signaling potential crises and assign importance (μ) to model-generated alarms based on their understanding of the economic context and risk appetite. These decisions significantly influence the relevance and impact of early warning models in guiding policy responses to financial-economic crises, a topic we will explore in detail in subsequent chapters. In this introductory chapter, we provide an overview of the study's background, objectives, and significance. The subsequent chapters are structured to offer a comprehensive exploration of the research topic.

Literature Review

3.1. Importance of Early Warning Models (EWM)

Early Warning Models (EWM) play a crucial role in identifying and mitigating the risks associated with financial-economic crises. These models leverage a variety of indicators and methodologies to predict potential downturns, enabling policymakers and financial institutions to take preventative measures. The recent advancements in machine learning and data analysis have significantly enhanced the predictive capabilities of EWMs, allowing for more nuanced and timely insights into the complex dynamics preceding crises. Research by [Padhan and Prabheesh, 2019] emphasize the effectiveness of global variables and housing prices as significant predictors of financial crises. Their analysis suggests that these indicators can serve as early warnings, enabling timely interventions to mitigate potential impacts on the economy.

Financial-economic crises, characterized by their sudden onset and far-reaching impacts, underline the necessity for effective early warning systems. Historical events, from the Tulip Mania to the 2008 Global Financial Crisis, highlight common threads of speculative bubbles, unsustainable debt levels, and inadequate regulatory frameworks, showcasing the critical need for vigilance and preparedness. The integration of machine learning algorithms offers a promising avenue for enhancing EWM accuracy, by analyzing vast datasets to identify subtle patterns and predict potential crises. The study by [Babecký et al., 2011] highlights the role of machine learning in improving the efficiency of EWMs. By leveraging advanced analytical techniques, machine learning algorithms can process and analyze large volumes of data, detecting patterns that precede financial-economic crises, thus providing policymakers with critical insights for early intervention.

Early warning systems play a pivotal role in preventing or mitigating the impact of financial-economic crises by providing decision-makers with timely information and insights. Here's why they are essential:

1. Timely Response: Financial economic crises often unfold rapidly, and a delayed response can exacerbate their impact. Early warning systems act as an alarm, allowing policymakers, regulators, and financial institutions to react swiftly and implement measures to stabilize the economy.

2. Risk Mitigation: By identifying emerging risks and vulnerabilities, early warning systems enable proactive risk management. This may involve implementing policies to curb excessive speculation, addressing unsustainable debt levels, or shoring up financial institutions

before they face critical challenges.

3. Preventing Contagion: Financial crises tend to spread across borders, affecting not only the country of origin but also neighboring economies. Early detection and containment can help prevent the contagion effect, limiting the global reach of a crisis.

4. Preserving Confidence: Financial markets thrive on confidence, and a sudden loss of trust can lead to a rapid unraveling of economic stability. Early warnings provide an opportunity to address underlying issues before they erode confidence in financial systems, helping to maintain stability.

3.2. Existing Literatures

Recent strides have ushered in a paradigm shift in the realm of early warning models, propelled by the integration of machine learning techniques as an innovative alternative to conventional methods. While historically dominated by venerable signaling approaches [Kaminsky and Reinhart, 1999] and discrete statistical (probit or logit) [Duca and Peltonen, 2013], innovation has been driven by scholars such ([Alessi and Detken, 2018], [Tanaka et al., 2016]). They advocate for enhancing early warning predictive precision using random forest, surpassing the power of conventional logit models and the signaling approach [Holopainen and Sarlin, 2017] have provided further momentum for the transformation of data analytics by identifying an array of machine learning techniques [Frankel and Rose, 1996], including artificial neural networks, support vector machines, k-nearest-neighbors, and decision trees, and highlighting their new applications in early warning models. By acknowledging these unique perspectives, we can better understand the evolving landscape of data analytics. In addition to model considerations, leading indicators represent another crucial aspect in creating early warning models. Finding an appropriate leading indicator may not be as straightforward as it might seem, due to the multitude of endogenous and exogenous variables that can trigger crises, with variations in each crisis. Notably, the paper [Babecký et al., 2012] employed the Bayesian averaging method, contributing to the identification of leading indicators, such as Money market interest rate, Commodity prices, Current account (GDP), Government debt (GDP), Stock market index, Global inflation, etc., for crisis incidence. Having an evaluation framework that aligns seamlessly with policymakers' decision-making context is of paramount importance. Specifically, the framework must navigate a policymaker's preferences concerning the balance between committing type I and type II errors, while also considering the practical utility of employing a model compared to its non-utilization. This approach stems from the loss-function concept initially introduced in [Alessi and Detken, 2014], which has since undergone further development and expansion [Lainà et al., 2015]. The exploration of methodologies for early warning mechanisms in financial crises has garnered significant attention in economic research. The paper [Benhamou et al., 2021] delves into the application of Gradient Boosting

Decision Trees (GBDT) to classify financial markets into normal and crisis regimes. Notably, their work demonstrates GBDT's prowess in learning from historical data and outperforming other machine learning methods. The study's predictive strength is underscored by its timely forecasts of the Covid crisis, presenting a promising tool for early detection of potential future crises. In a separate investigation, [Christensen and Li, 2014] proposes an early warning system based on signal extraction, aiming to predict the likelihood of financial stress events. Their approach involves constructing three composite indicators and evaluating their in-sample and out-of-sample performance. The authors find that the weighted composite indicator excels in predicting financial stress events, suggesting potential avenues for refining dynamic components and considering economic status effects in future research. Moving to a different facet of predictive modeling, [Hopp, 2022] evaluates the performance of Long Short-Term Memory Networks (LSTMs) in economic nowcasting, comparing them to the dynamic factor model. Hopp's study reveals the superior predictive results of LSTMs in global merchandise trade exports and services exports. Despite acknowledging drawbacks such as stochastic nature and interpretability challenges, the author advocates for heightened consideration of LSTMs in economic nowcasting, emphasizing their potential for facilitating future research. Taking a broader perspective on early-warning models, [Holopainen and Sarlin, 2017] presents robust models by comparing conventional statistical methods with more recent machine learning approaches. Their exploration includes ensemble approaches to aggregate information and methods for estimating model uncertainty. The study underscores the significance of structured modeling, ensemble methods, and acknowledging uncertainty in early-warning exercises, particularly in predicting the ongoing financial crisis in Europe. Shifting the focus to banking distress prediction, [Lang et al., 2018] proposes a conceptual framework for deriving early-warning models. Their flexible modeling solution combines the loss function approach with regularized logistic regression and cross-validation. The authors illustrate the application of this framework to a dataset of EU banks, offering valuable insights for both micro- and macro-prudential policy analysis. In addressing the question of machine learning's superiority in financial crisis prediction, [Beutel et al., 2018] provides a critical evaluation. Their study, based on the most recent European crises database, challenges the notion that machine learning methods outperform conventional models. The authors caution against overconfidence in machine learning's predictive performance, stressing the need for further improvements before these models can be reliably applied in policy-making. They identify key early warning indicators, including expansions in credit and investment, asset price booms, and external imbalances.

However, it's worth noting that none of the reviewed studies tested the models we use in our research. Our work aims to fill this gap by conducting a thorough evaluation of these models. We seek to contribute by testing a variety of models to understand their strengths and

weaknesses in predicting financial crises.

Data

In the field of economic research, the significance of data cannot be overstated. The quality, accuracy, and relevance of the data employed in a study play a pivotal role in shaping the outcomes and conclusions drawn. The choice of data sources, the precision in data collection methods, and the comprehensive coverage of relevant variables collectively influence the robustness and reliability of the research findings. Data serves as the backbone of any empirical investigation, providing the raw material upon which analytical methodologies and models are applied. The selection of appropriate data sources ensures that the research captures a faithful representation of the economic phenomena under scrutiny. Inaccurate or incomplete data can introduce bias, distort patterns, and compromise the validity of research outcomes. Moreover, the dynamism of economic systems demands a keen awareness of temporal and spatial variations in data. Economic conditions, regulatory landscapes, and market dynamics are subject to continuous evolution, requiring researchers to maintain an acute awareness of the timeliness and relevance of their datasets. The authors acknowledge that the outcomes of studies can be sensitive to the specifics of the data used. Therefore, a rigorous and transparent approach to data selection, validation, and preprocessing is essential. This ensures that the findings derived from the research are not only academically robust but also applicable and insightful for policymakers, practitioners, and stakeholders. As we delve into the data section of this study, it is imperative to recognize the critical role that data quality and appropriateness play in the subsequent analyses. The datasets chosen for this research undergo scrutiny and validation, aligning to produce meaningful insights into the dynamics of financial-economic crises.

The dataset employed in this study encompasses a diverse set of 13 countries, each characterized by its unique economic landscape and corresponding financial indicators. Table 3.1 provides a comprehensive overview of these countries and their associated crises. The temporal scope of our data analysis spans from the first quarter of 1970 (Q1 1970) to the first quarter of 2019 (Q1 2019). It is important to note that we intentionally concluded our data analysis before the onset of the COVID-19 crisis, as this event introduces unprecedented and stochastic elements that may not align with the predictive nature of our modeling.

4.1. Data Frequency

The inclination towards utilizing quarterly data in early warning models for economic crisis prediction is driven by the need to find a balance between capturing meaningful trends and

Table IV.1 Country crisis dates

Country	Start crisis	End crisis	Start crisis	End crisis	Start crisis	End crisis
DE	1974 Q2	1974 Q4	2001 Q1	2003 Q4	2007 Q3	2013 Q2
DK	1987 Q1	1995 Q1	2008 Q1	2013 Q4		
FI	1991 Q3	1996 Q4	-	-	-	-
FR	1991 Q2	1995 Q1	2008 Q2	2009 Q4		
GB	1973 Q4	1975 Q4	1991 Q3	1994 Q2	2007 Q3	2010 Q1
IT	1991 Q3	1997 Q4	2011 Q3	2013 Q4		
JP	1997 Q4	2001 Q4	-	-	-	-
NL	2008 Q1	2013 Q1	-	-	-	-
NO	1988 Q3	1992 Q4	2008 Q3	2009 Q3	-	-
PT	2008 Q4	2015 Q4	-	-	-	-
SE	1991 Q1	1997 Q2	2008 Q3	2010 Q4	-	-
US	2007 Q4	2010 Q4	-	-	-	-
CA	1985 Q2	1987 Q4	2008 Q4	2009 Q4	-	-

Source: MacroHistory Database

mitigating the inherent noise associated with high-frequency data or infrequent observations. By employing quarterly data, the volatility and noise present in economic markets are smoothed out, allowing a more focused examination of significant underlying patterns. This choice is particularly advantageous as key macroeconomic indicators relevant to crisis prediction, such as GDP growth and unemployment rates, are commonly reported quarterly, ensuring consistency and compatibility. The medium-term nature of economic crises aligns well with the quarterly data frequency, effectively capturing dynamic trends without being overly influenced by short-term fluctuations. Additionally, the broader availability of quarterly data across countries and its reduced data collection burden contribute to its practicality for model building. The stability of quarterly data further enhances its suitability for predictive modeling. Finally, the alignment of quarterly data with policy cycles is noteworthy, facilitating synchronization with key policy decisions made every quarter.

4.2. Feature Engineering

4.2.1. Missing data replacement

For variables exhibiting NaN values at the commencement of the data, a twofold strategy was implemented for both tranquil and crisis periods. In tranquil periods, if NaN values were identified during tranquil periods, the missing data points were replaced with the average values observed in the subsequent tranquil period where data was available. Conversely, in the case of NaN values during crisis periods, the missing data points were imputed using the average values from the subsequent crisis period where data was present. This imputation strategy was

applied individually for each country in the dataset, ensuring a tailored and consistent approach throughout. The utilization of averages from adjacent periods helped maintain accuracy while addressing missing values, contributing to the overall reliability and completeness of the dataset for subsequent analysis.

4.2.2. Mathematical Transformation of Variables

In total, our dataset comprises 10 main variables, each representing a key aspect of the economic and financial landscape. Additionally, we have created 10 variables derived from these main indicators, introducing further dimensions to our analysis. The details of these variables and their definitions are outlined in Table 3.2, providing a comprehensive guide to the components shaping our research. The meticulous selection, validation, and construction of these variables contribute to the robustness of our dataset and, consequently, the reliability of our analyses. It's worth mentioning that we initially included share price and housing price in growth variables and dummy variables. However, for the sake of model simplicity and a more focused analysis on primary economic indicators, we decided to exclude them from subsequent analyses. Consequently, we narrowed down our variables to a final set of 18 for in-depth examination.

One of the crucial parts of feature engineering part is a scaling is a crucial preprocessing step in machine learning, playing a vital role in ensuring that different features of a dataset are on a comparable scale. The significance of scaling lies in its ability to prevent certain features from disproportionately influencing the learning algorithm, particularly those that have larger magnitudes or variances. Without scaling, models that rely on distance metrics, such as k-nearest neighbors or support vector machines, may be biased towards features with larger numerical ranges. Additionally, scaling aids gradient-based optimization algorithms, like those used in neural networks, in converging faster and more efficiently. When dealing with numerical variables in diverse ranges, applying an appropriate scaler becomes imperative. In cases where the dataset includes outliers, opting for RobustScaler over StandardScaler becomes advantageous as it ensures a more robust and accurate representation of the data's underlying patterns, thereby enhancing the overall performance of machine learning models.

In Appendix 1, we present supplementary visualizations aimed at enhancing our understanding of the variables employed in our models. These figures depict the correlations among all variables utilized in our models. In contrast to the comprehensive overview in the first figure, the second figure specifically highlights variable pairs with an absolute correlation exceeding 0.5. This selective focus provides a nuanced exploration of inter-variable relationships, emphasizing pairs with stronger correlations for a more targeted analysis.

Table IV.2 A list of indicators

Variable name	Definition	Transformation and additional information
GDP growth	Real gross domestic product	1-year growth rate
House price growth	Real residential property price index	1-year growth rate
Inflation	Real consumer price index	1-year growth rate
Unemployment	people of working age who are without work, are available for work, and have taken specific steps to find work	1-year growth rate
Share price growth	Real stock price index	1-year growth rate
Account to GDP	Nominal current account balance and nominal GDP	Ratio
Credit to GDP	Nominal total credit to the private non-financial sector and nominal GDP	Ratio
Gross fixed capital formation to GDP	Formerly gross domestic investment to GDP	Ratio
10Y government bond rate	Real long-term government bond rate	Level
3M money market rate	Short-term borrowing rate	Level
Credit to GDP and 3MMMR dummy	higher than mean = 1 otherwise = 0	Dummy
Inflation and 10YGB	higher than mean = 1 otherwise = 0	Dummy
GDP and Unemployment	higher than mean = 1 otherwise = 0	Dummy
Credit to GDP and Unemployment	higher than mean = 1 otherwise = 0	Dummy
GDP and Share price	higher than mean = 1 otherwise = 0	Dummy
Credit to GDP cycle	Nominal bank credit to the private non-financial sector and nominal GDP	Absolute deviation from trend, $\lambda = 1,600$
Housing price cycle	Deviation from trend of the real residential property price index	Absolute deviation from trend, $\lambda = 1,600$
Share price cycle	Deviation from trend of the real residential property price index	Absolute deviation from trend, $\lambda = 1,600$

Sources: OECD, World Bank, and own calculations

Methodology

5.1. Concept of Early Warning Mechanism and Metrics

Early warning models (EWMs) are designed to address classification problems, particularly those associated with events characterized by low probabilities but potentially high impacts. In assessing these models, it is crucial to establish evaluation criteria that account for the unique nature of the underlying concern events. Creating an evaluation framework that seamlessly aligns with the decision-making context of policymakers is of utmost importance.

This framework should consider policymakers' preferences regarding the trade-off between committing type I and type II errors. Simultaneously, it should weigh the practical utility of employing a model against not utilizing it. This approach is rooted in the concept of a loss function introduced in [Alessi and Detken, 2014] and further developed in [Lainà et al., 2015].

Type I error, also known as a false positive, occurs when a test incorrectly identifies something as significant when it is not. On the other hand, Type II error, or false negative, happens when a test incorrectly identifies something as not significant when it is. Understanding and addressing these errors are essential in developing an evaluation framework that meets the specific needs of policymakers in the realm of early warning models.

		Model Prediction $C_{i,t}$	
		0	**1**
Actual Outcome $P_{i,t}$	**0**	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Table V.1 Contingency matrix for observed outcome and model prediction.

Table 4.1 shows the Contingency matrix, a table that displays the frequency distribution of two categorical variables. It allows one to compare the observed distribution of the two variables against the expected distribution based on a specific hypothesis.

Consider $C_{i,t}$ as a binary state variable that indicates whether a crisis occurred ($C_{i,t} = 1$) or did not occur ($C_{i,t} = 0$) for a specific entity i at time t. Let $p_{i,t}$ denote the estimated likelihood of entity i being in a state of vulnerability during period t. This probability is subsequently translated into a binary indicator by applying a threshold value τ, which falls within the interval of $\tau \in [0, 1]$:

$$z_{i,t} = \begin{cases} 1, & \text{if } z_{i,t} > \tau \\ 0, & \text{otherwise} \end{cases}$$

The relationship between $C_{i,t}$ and $P_{i,t}$ is as follows:

- When $C_{i,t} = 1$ (crisis occurred if a crisis occurred at time t for entity i), then the corresponding $P_{i,t}$ value indicates whether the model correctly predicted vulnerability (1) or not (0).
- When $C_{i,t} = 0$ (no crisis occurred), if there was no crisis at time t for entity i, the $P_{i,t}$ value generated by the model still indicates whether the vulnerability was predicted (1) or not (0). In this case, a $P_{i,t} = 1$ corresponds to a "false alarm," implying that the model predicted vulnerability even when there was no actual crisis.

In this context, the comparison between the actual crisis occurrences $C_{i,t}$ and the model's predictions of vulnerability $P_{i,t}$ allows for an assessment of the model's accuracy in predicting crises. This relationship serves as the basis for evaluating the model's performance against historical crisis data. In the realm of predicting economic crises, evaluating model performance requires an approach that mirrors the real-world decisions of policymakers. This entails considering two types of errors: missing pre-crisis periods (type I error) and issuing false alarms during tranquil times (type II error):

$$T_1(\tau) = \frac{FN(\text{False Negative})}{FN(\text{False Negative}) + TP(\text{True Positive})}$$
$$T_2(\tau) = \frac{FP(\text{False Positive})}{FP(\text{False Positive}) + TN(\text{True Negative})}$$

These errors are weighted based on the policymaker's preferences ($\mu \in [0, 1]$) to gauge their impact on decision-making. However, the evaluation process extends beyond preferences. It also factors in the inherent class imbalances between crisis and non-crisis periods. By accounting for the unconditional probabilities of these periods, the evaluation becomes more realistic.

$$P_1 = \Pr(C_{(i,t)} = 1)$$
$$P_2 = \Pr(C_{(i,t)} = 0)$$

For instance, a few false alarms might have a lesser impact when they're a small fraction of numerous non-crisis instances. In essence, the evaluation framework combines preferences, class imbalances, and the actual distribution of crisis and non-crisis periods. This holistic

approach ensures that early-warning models align with the nuanced decision-making needs of policymakers, making it pivotal in creating effective prediction systems.

Based on these components, the loss function of policymakers can be written as follows:

$$L(\mu, \tau) = \mu P_1 T_1(\tau) + (1 - \mu) P_2 T_2(\tau)$$

Depending on this equation, we can go further and describe the absolute usefulness (U_a):

$$U_a(\mu, \tau) = \min[\mu P_1, (1 - \mu) P_2] - L(\mu, \tau)$$

The concept of "absolute usefulness" (U_a) for a model can be understood by comparing the policymaker's loss when using the model to the loss when the model is not used. It's like measuring how much benefit the model brings in terms of reducing errors. For example, if using the model results in fewer missed crises and false alarms compared to not using the model, then the model is considered more useful, and U_a would be higher. It's a way to quantify how well the model improves decision-making in predicting crises. Switching our focus to "relative usefulness" (U_r), we delve into a perspective that assesses the model's effectiveness about the maximum potential benefit. Rather than isolating the model's utility, U_r quantifies its performance as a portion of the ultimate attainable success:

$$U_r(\mu, \tau) = \frac{U_a}{\min[\mu P_1, (1 - \mu) P_2]}$$

In addition to the aforementioned metrics, the contingency matrix can be leveraged to calculate various other quantitative assessments. One such tool is the F1-score, a metric that balances precision, the ratio of true positive predictions to the total number of positive predictions, and recall, the ratio of true positive predictions to the total number of actual positive instances. It is used to evaluate the performance of classification models, particularly when the classes are imbalanced. The F1-score is particularly useful in binary classification tasks, providing a single value that combines both false positives and false negatives.

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score ranges between 0 and 1, with higher values indicating better performance in achieving a balance between precision and recall. It is a valuable measure when the class distribution is imbalanced or when both false positives and false negatives are of concern.

To ensure that our models accurately capture patterns from the data and that their predictions are reliable, we employed a robust testing method. After obtaining initial results, we randomly altered the dependent variable and retrained all existing models using this modified dataset. If, under these conditions, the model performance deteriorates significantly, it confirms that the original results were indeed reasonable. This methodology not only validates the

accuracy of the results but also affirms the relevance of the variables identified by the models as the most contributory. Thus, we can confidently assert that the model outcomes and the selected variables are both valid and significant.

5.2. Forecasting Horizon

Early-warning models are designed with the primary objective of issuing timely alerts for potential distress events. The specific timeframe for making these predictions can vary depending on the context. Although it is possible to treat the forecast horizon as a flexible parameter, [Bussiere and Fratzscher, 2006] argue that it is more beneficial for it to be predetermined to align with the specific task rather than optimized to fit historical data.

Recent contributions in the field of forecasting banking crises have adopted various prediction horizons. Studies, such as [Behn et al., 2013], have utilized longer horizons ranging from 5 to 12 quarters, while others, like [Alessi and Detken, 2014], extended the range to 5 to 16 quarters. This suggests that objectives related to macroprudential considerations often necessitate extended prediction horizons. However, there is a lack of consensus on the ideal timeframe, as evidenced by studies commonly testing multiple prediction horizons ([Behn et al., 2013]; [Lainà et al., 2015]).

In our work, the modeling window is ranged from 3 to 12 quarters. This extended timeframe enables the model to discern patterns and signals that may precede a crisis, allowing for a more comprehensive understanding of the factors at play. The modeling process concludes with the introduction of ones for the final two quarters within the window. By incorporating information from the two quarters immediately before the crisis, the model enhances its capability to serve as an early warning system, affording decision-makers a valuable lead time of two quarters to implement preventive measures and mitigate the potential impact of the impending financial-economic crisis.

5.3. Used Models

In our analytical exploration, we strategically employed a diverse array of machine learning models, ranging from classic logistic regression to its regularized variants with L1 and L2 regularization. The objective was twofold: firstly, to draw insightful comparisons between traditional approaches and contemporary methodologies, and secondly, to delve into the nuances within newer techniques. For this purpose, we enlisted k-nearest neighbors (KNN), support vector machines (SVM), decision trees (DT), random forests (RF), AdaBoost, Gradient Boosting (GB), XGBoost (XGB), LightGBM (LXGB), and CatBoost. By examining the performance of these models, we aimed to discern the strengths and weaknesses of each, shedding light on the evolution of machine learning techniques and offering valuable insights

Table V.2 Comparison of Employed Methods: Benefits and Drawbacks

Method	Benefits	Drawbacks
Logit	Probabilistic foundations, highly interpretable	Fixed functional form limits flexibility
KNN	Simple and intuitive methodology	Suffers from the curse of dimensionality
Decision Trees	Automatically selects important variables, intuitive	Prone to instability across different samples
Random Forest (RF)	Enhances stability and accuracy over decision trees	Risk of overfitting, complexity in interpretation
Support Vector Machine (SVM)	Efficient for nonlinear problems, computationally efficient	Can overfit, complex to explain, not inherently probabilistic
AdaBoost	Combines multiple weak learners effectively, resists overfitting	Vulnerable to noisy data and outliers, longer training times
Gradient Boosting	High accuracy, handles complex data interactions well	Can overfit if not tuned, computationally demanding
XGBoost	Optimized gradient boosting with high efficiency	Similar to GB; can overfit, requires careful tuning
LightGBM	Rapid processing, efficient with large data sets	Less interpretable, limited support for categorical data
CatBoost	Excellently manages categorical data, resists overfitting well	Comparatively slower with large data sets, needs extensive tuning

into their practical implications.

5.3.1. Logistic Regression

Logistic regression is a statistical method used for modeling the probability of binary outcomes. In the context of financial crises, it serves as an effective tool for constructing early warning models. In logistic regression modeling for financial crises, ([Frankel and Rose, 1996]; [Fuertes and Kalotychou, 2006]; [Barrell et al., 2010]) have made significant contributions. ([Frankel and Rose, 1996]; [Barrell et al., 2010]) have explored logistic regression for currency crises, while [Fuertes and Kalotychou, 2006] focus on debt crises. These studies collectively demonstrate the versatility of logistic regression in understanding and forecasting diverse financial challenges

Binary response variable Y represents the occurrence (1) or non-occurrence (0) of a financial crisis. Let $X_1, X_2, \ldots, X_n$ be predictor variables. The logistic regression model is defined as:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

where $P(Y=1)$ is the probability of a crisis occurring, and $\beta_0, \beta_1, \ldots, \beta_n$ are coefficients estimated from historical data.

The logistic regression model estimates the log odds of a financial crisis based on predictor variables. By fitting historical data, the model learns the relationship between these variables and crisis occurrence. Thresholds can be set to trigger warnings when the estimated probability surpasses a predefined level. In logistic regression, L1 and L2 regularization are techniques used to prevent overfitting by adding penalty terms to the objective function. This helps in creating more robust models, especially when dealing with limited data and complex relationships.

L1 Regularization (Lasso)

L1 regularization adds the absolute values of the coefficients as penalty terms:

$$\text{Objective Function with L1:} \quad J(\beta) = -\sum_{i=1}^{N} \left(y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\right) + \lambda \sum_{j=1}^{p} |\beta_j|$$

where β represents the coefficients, $\hat{p}_i$ is the predicted probability, and λ controls the strength of the regularization.

L2 Regularization (Ridge)

L2 regularization adds the squared values of the coefficients as penalty terms:

$$\text{Objective Function with L2:} \quad J(\beta) = -\sum_{i=1}^{N} (y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)) + \lambda \sum_{j=1}^{p} \beta_j^2$$

where the terms have the same meanings as in L1 regularization.

In financial crisis prediction, L1 regularization aids in feature selection by driving certain coefficients to zero. This promotes sparsity in the model, highlighting the most relevant indicators for crisis prediction. L2 regularization helps prevent overfitting by penalizing large coefficients. In the context of financial crises, this promotes stability in the model and reduces sensitivity to noise in the data.

5.3.2. K-Nearest Neighbor (KNN)

K-Nearest Neighbors, as proposed in [Altman, 1992], is a machine learning algorithm employed for classification tasks, making it suitable for identifying potential financial crises. The algorithm operates based on the proximity of data points in a feature space.

Consider a dataset with predictor variables $X_1, X_2, \ldots, X_n$ and a binary response variable Y indicating the occurrence (1) or non-occurrence (0) of a financial crisis. The KNN algorithm classifies a new data point by finding its k-nearest neighbors in the feature space.

The distance between two data points, i and j, can be calculated using a distance metric, commonly the Euclidean distance:

$$\text{Distance}(i, j) = \sqrt{\sum_{p=1}^{n} (X_{i,p} - X_{j,p})^2}$$

The probability of a financial crisis occurring ($P(Y = 1)$) for a new data point is determined by the majority class among its k-nearest neighbors:

$$P(Y = 1) = \frac{1}{k} \sum_{i \in \text{neighbors}} Y_i$$

where Y_i is the class label of the i-th neighbor.

Additionally, KNN allows for different weighting schemes. One common approach is uniform weighting, where each neighbor's contribution is equally weighted in the calculation. Alternatively, distance-based weighting can be used, giving more influence to closer neighbors.

The application of KNN involves selecting an appropriate value for k, the number of

neighbors considered. Historical data is used to train the model, and when faced with a new data point, the algorithm identifies its k-nearest neighbors based on the chosen distance metric and weighting scheme. The majority class among these neighbors determines the probability and classification of a financial crisis.

5.3.3. Support Vector Machine (SVM)

Support Vector Machines are powerful machine learning algorithms used for classification tasks, making them suitable for predicting financial crises. SVM aims to find the optimal hyperplane that best separates different classes in a feature space. This approach was introduced in [Cortes and Vapnik, 1995] .

Consider a dataset with predictor variables $X_1, X_2, \ldots, X_n$ and a binary response variable Y indicating the occurrence (1) or non-occurrence (0) of a financial crisis. The SVM algorithm seeks to find a hyperplane represented by the equation:

$$\sum_{i=1}^{n} \alpha_i Y_i \langle X_i, X \rangle + b = 0$$

where Y_i is the class label, X_i is a support vector in the training set, $\langle X_i, X \rangle$ denotes the dot product, and α_i are the coefficients determined during training.

The SVM algorithm classifies a new data point by evaluating which side of the hyperplane it falls on. The decision function is given by:

$$f(X) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i Y_i \langle X_i, X \rangle + b\right)$$

where sgn is the sign function. The output of this function determines the predicted class of a financial crisis.

5.3.4. Decision Tree (DT)

Decision Trees, used in [Tong and Tong, 2022], is powerful machine learning algorithms used for classification tasks, providing a clear and interpretable structure. In the context of predicting financial crises, Decision Trees can serve as effective early warning models.

Consider a dataset with N data points, each characterized by predictor variables $X_1, X_2, \ldots, X_n$ and a binary response variable Y indicating the occurrence (1) or non-occurrence (0) of a financial crisis.

A Decision Tree recursively partitions the data based on feature values. Let j represent the selected feature index, t be the threshold, and R denote a region in the feature space. The decision function for a binary classification task can be expressed as:

$$f(X) = \begin{cases} \text{left}, & \text{if } X_j \leq t \\ \text{right}, & \text{otherwise} \end{cases}$$

The recursive splitting process continues until a predefined stopping criterion is met or the tree reaches its maximum depth.

At each node, the algorithm selects the optimal feature j and threshold t to maximize the information gain or minimize impurity. Common impurity measures include Gini impurity and entropy:

$$\text{Gini}(R) = 1 - \sum_{k=1}^{K} p_k^2$$

$$\text{Entropy}(R) = -\sum_{k=1}^{K} p_k \log(p_k)$$

where K is the number of classes, and p_k is the proportion of data points in class k within region R.

To classify a new data point X, it traverses the Decision Tree based on the feature values, ultimately reaching a leaf node. The class label assigned to the leaf node represents the predicted outcome, serving as an early warning for the potential occurrence of a financial crisis.

5.3.5. Random Forest (RF)

Random Forests [Alessi and Detken, 2014], have been exclusively employed in early-warning exercises within, is an ensemble learning algorithm that leverages multiple decision trees to improve prediction accuracy and robustness. It excels in capturing complex relationships in data, making it a powerful tool for predicting financial crises.

Consider a dataset with predictor variables $X_1, X_2, \ldots, X_n$ and a binary response variable Y indicating the occurrence (1) or non-occurrence (0) of a financial crisis. Let D be the dataset with N observations.

The algorithm builds T decision trees, each trained on a bootstrap sample D_i of size N, drawn with replacement from D. Additionally, at each split, a random subset of features m is considered.

The decision tree t is trained to minimize the impurity criterion, such as Gini impurity or entropy, resulting in a series of binary splits based on feature values.

The prediction for a new data point $\mathbf{x}$ is obtained by aggregating the individual tree predictions:

$$\hat{Y}(\mathbf{x}) = \text{mode}\left(\hat{Y}_1(\mathbf{x}), \hat{Y}_2(\mathbf{x}), \ldots, \hat{Y}_T(\mathbf{x})\right)$$

where $\hat{Y}_t(\mathbf{x})$ is the prediction of the t-th tree.

Random Forest mitigates overfitting by building diverse trees on different data subsets. The randomness introduced in feature selection and bootstrapping contributes to the model's generalization ability.

To classify a new data point, the majority vote among the ensemble determines the predicted class. The aggregation process results in a stable and robust prediction, making Random Forest a formidable choice for early warning systems in financial contexts.

5.3.6. Boosting Models

Boosting methods in Early Warning Models (EWM) are powerful tools for predicting financial crises. These algorithms, like AdaBoost, Gradient Boosting, XGBoost, LightGBM, and CatBoost, work by combining multiple weak models to create a robust predictor. They sequentially learn from data, focusing on correcting errors, and adapt well to complex financial patterns. Implementation involves training weak models, often decision trees, and combining their predictions. With their adaptability and ability to handle various data types, boosting methods contribute to effective EWMs, providing valuable insights for proactive risk management in the financial sector.

AdaBoost

AdaBoost [Freund and Schapire, 1996] combines weak learners h_t with different weights w_t to create a strong learner:

$$F(x) = \sum_{t=1}^{T} w_t \cdot h_t(x)$$

Weights are adjusted based on the accuracy of weak learners, with misclassified instances gaining higher weights. The final prediction is determined by a weighted majority vote.

Gradient Boosting (GB)

Gradient Boosting [Friedman, 2001] builds an ensemble of decision trees sequentially. Given a loss function $L(y, F(x))$, where y is the true label and $F(x)$ is the current prediction, each tree corrects the residuals:

$$F_t(x) = F_{t-1}(x) + \gamma \cdot h_t(x), \quad \text{where } h_t = \arg\min_h \sum_{i=1}^{N} L(y_i, F_{t-1}(x_i) + h(x_i))$$

The learning rate γ controls the contribution of each tree.

XGBoost (XGB)

XGBoost [Chen and Guestrin, 2016] extends GB by adding a regularized objective function:

$$\text{Obj} = \sum_{i=1}^{N} L(y_i, F_{t-1}(x_i) + h(x_i)) + \sum_{k=1}^{K} \Omega(f_k)$$

where $\Omega(f_k)$ is the regularization term. XGBoost optimizes this objective function using second-order Taylor expansion for faster convergence.

LightGBM

LightGBM [Ke et al., 2017] is a histogram-based gradient boosting framework. It constructs histograms for continuous features and finds the best splits. The objective function is defined as:

$$\text{Obj} = \sum_{i=1}^{N} L(y_i, F_{t-1}(x_i) + h(x_i)) + \sum_{k=1}^{K} \Omega(f_k)$$

LightGBM uses a leaf-wise tree growth strategy, choosing the leaf with the max delta loss.

CatBoost

CatBoost [Dorogush et al., 2018] efficiently handles categorical features using an efficient algorithm for processing them during training. The objective function includes a regularization term:

$$\text{Obj} = \sum_{i=1}^{N} L(y_i, F_{t-1}(x_i) + h(x_i)) + \sum_{k=1}^{K} \Omega(f_k)$$

CatBoost automatically handles categorical variables, reducing the need for preprocessing.

Results

6.1. Best Hyperparameters and Key Predictive Variables

To optimize the performance of our predictive models for crisis detection, we conducted an extensive hyperparameter tuning process. We utilized the Grid Search method to systematically explore various combinations of hyperparameters, aiming to maximize the F1 score, a critical metric for assessing model performance in imbalanced binary classification tasks.

Table 5.1 provides a summary of the hyperparameters used in Grid Search and the corresponding values chosen for each model. These optimized settings represent the configurations that yielded the highest F1 scores during the tuning process.

Table VI.1 Best Hyperparameters for F1 Score

Model	Hyperparameters	Values
Logistic Regression	Penalty, C	l1, 10
K-Nearest Neighbors	n_neighbors, weights, p	3, distance, 1
Support Vector Machine	C, kernel, gamma	10, rbf, 0.1
Decision Tree	Criterion, max_depth, min_samples_split, min_samples_leaf	gini, 30, 5, 1
Random Forest	n_estimators, criterion, max_depth, min_samples_split, min_samples_leaf	100, entropy, 30, 2, 1
AdaBoost	n_estimators, learning_rate	150, 0.2
Gradient Boosting	n_estimators, learning_rate, max_depth, min_samples_split, min_samples_leaf	150, 0.2, 5, 10, 1
XGBoost	n_estimators, learning_rate, max_depth, min_child_weight, gamma	200, 0.2, 7, 1, 0
LightGBM	n_estimators, learning_rate, max_depth, min_child_samples, subsample	200, 0.2, 5, 30, 0.8
CatBoost	iterations, learning_rate, depth, l2_leaf_reg	300, 0.2, 7, 1

In predictive modeling, the selection and importance of variables are paramount, as they directly influence the model's ability to make accurate forecasts. Variables serve as the foundational elements through which models capture the complexities of real-world phenomena. Identifying and understanding the key variables is essential for enhancing

model robustness, tailoring interventions, and ultimately improving decision-making processes. Effective models hinge on the precise inclusion of variables that are significantly correlated with the outcome of interest, ensuring that the predictive insights are both relevant and actionable. The significance of these variables is particularly evident in financial modeling, where economic indicators can delineate trends and potential market shifts.

Across the range of models assessed, **Account to GDP**, **Inflation**, and **Housing Price Cycle** consistently emerge as top-ranking variables out of 18 variables, with at least one or a combination of two or even all three being among the top three predictors. While they may not always appear together, their recurring presence underscores their significant influence on predictive outcomes. This consistent pattern highlights their fundamental importance across diverse models, reaffirming their critical role in shaping predictive accuracy and capturing underlying data dynamics.

The Account to GDP ratio, which measures a country's current account balance relative to its GDP, serves as a critical macroeconomic indicator. It reflects the nation's economic engagement with the rest of the world, where a surplus indicates net lending and a deficit indicates net borrowing. This metric is essential for understanding the impact of trade balances and international investment flows on overall economic health. Inflation, another key predictor, directly affects purchasing power, savings, and investment decisions by altering the value of money, influencing both consumer behavior and financial stability. Meanwhile, the Housing Price Cycle, by reflecting fluctuations in real estate markets, impacts consumer wealth, construction activity, and mortgage markets, and can lead to significant economic shifts, as demonstrated during the 2008 financial crisis. Together, these variables provide a comprehensive view of economic dynamics, making them indispensable in enhancing the accuracy of models.

6.2. Model Results

As we mentioned in the "Introduction", there are two components of the forecasting process, τ and μ, that need to be decided by policymakers. In our case, we use 0.2 and 0.8 respectively. The results may change when we change these values, to show how the forecast is changing in Appendix 2 for each model we have 2 graphs, Error Rates vs. Threshold and Loss, Absolute Usefulness, and Relative Usefulness vs. Relative Preference. In these graphs we show how they change when we change τ and μ, we show three error rates (Overall, for Crisis, and non-crisis) and in the second graph, we show Loss, Absolute Usefulness, and Relative Usefulness. It is important to note that the graphical representation illustrating Loss, Absolute Usefulness, and Relative Usefulness against Relative Preference is sensitive to variations not only in the parameter μ but also in the threshold parameter τ. When the threshold τ is adjusted, for instance, transitioning from 0.2 to 0.3, while keeping the μ value constant at 0.8, the outcomes displayed

in the figure will exhibit variations. In other words, altering the threshold parameter τ introduces changes in the depicted results, highlighting the joint influence of both μ and τ on the observed metrics.

In the presented Threshold Analysis Table (Table 5.2), we evaluate the performance of models in predicting financial-economic crises. The models, including Logit, Lasso, KNN, SVM, DT (Decision Tree), RF (Random Forest), AdaB (AdaBoost), GB (Gradient Boosting), XGB (Extreme Gradient Boosting), LGBM (Light Gradient Boosting Machine), and CatB (CatBoost), are assessed based on multiple metrics.

Table VI.2 Threshold Analysis Table Transposed

Model	**F1-score**	**Loss**	**Absolute Usefulness**	**Relative Usefulness**
Logistic Regression	0.3210	0.0558	0.0104	0.1570
Lasso Regression	0.3529	0.1835	-.01173	-0.7733
K-Nearest Neighbors	0.7350	0.0119	0.0542	0.8198
Support Vector Machine	0.7010	0.0215	0.0446	0.6744
Decision Tree	0.7126	0.0235	0.0427	0.6453
Random Forest	0.8515	0.0058	0.0604	0.9128
AdaBoost	0.6207	0.0250	0.0412	0.6221
Gradient Boosting	**0.9412**	**0.0054**	**0.0608**	**0.9186**
Extreme Gradient Boosting	0.8667	0.0092	0.0569	0.8605
Light Gradient Boosting Machine	0.9302	0.0058	0.0604	0.9128
CatBoost	0.8736	0.0100	0.0562	0.8488

Notably, GB stands out with the highest F1-score of 0.9412, closely followed by LGBM at 0.9302. These results suggest that GB and LGBM demonstrate superior performance in terms of precision and recall. The Loss metric, representing the logistic loss function, emphasizes the accuracy of the models in predicting class probabilities. GB again excels with the lowest loss value of 0.0054, indicating its robust predictive capabilities. Absolute Usefulness and Relative Usefulness metrics gauge the contribution of each model to the overall predictive power. Here, the models are ranked based on their impact, with GB, LGBM, and RF emerging as top performers. These metrics provide insights into each model's practical significance and relative importance in contributing to the overall predictive accuracy.

6.2.1. Why Boosting Models Outperform Others?

1. **Sequential Learning:** Boosting models build trees sequentially, each one focusing on the mistakes of the previous one. This iterative refinement enables the model to progressively improve its accuracy. Unlike bagging methods like Random Forest, which build trees independently, boosting's sequential approach allows it to zero in on difficult-to-predict

instances, enhancing overall performance.

2. **Weighted Errors:** In each iteration, boosting models assign higher weights to the instances that were incorrectly predicted. This means that the model pays more attention to these challenging cases, continually adjusting to reduce the overall error. This adaptive weighting mechanism ensures that the model is robust against a variety of data distributions and anomalies.

3. **Regularization Techniques:** Boosting models incorporate regularization methods to prevent overfitting. Techniques such as learning rate adjustment, tree pruning, and subsampling help to balance the model's complexity and its ability to generalize. Regularization is particularly important when dealing with noisy data or small datasets, as it ensures that the model does not become overly complex and tailored to the training data.

4. **Advanced Optimization:** Models like XGBoost and LightGBM utilize advanced optimization techniques, including parallel processing and efficient memory usage, to handle large datasets more effectively. These models also offer flexible hyperparameter tuning options, allowing for fine-grained control over the learning process, which enhances model performance and efficiency.

6.3. Assessing Model Stability with Altered Data

As discussed in the last paragraph of the "Concept of Early Warning Mechanism and Metrics" section, we implemented a validation process to test the robustness of our models. The analysis of the provided Table 5.3 reveals a noteworthy discrepancy between the F1-scores and the Absolute and Relative Usefulness metrics. Despite the high F1-scores (but lower compared to the actual results), the Absolute and Relative Usefulness values appear disproportionately low, even reaching negative figures. This disparity underscores a crucial insight: while F1 metrics offer valuable insights, they may not always provide a comprehensive understanding of model performance. Therefore, a holistic evaluation incorporating Absolute and Relative Usefulness metrics is imperative for robust conclusions.

The observed inconsistency between the F1-score and the Absolute and Relative Usefulness metrics in Table 5.3 can be attributed to several factors that influence how these metrics evaluate model performance. Firstly, the F1-score primarily measures the balance between precision and recall, offering a view of model accuracy in terms of the harmonic mean of these two metrics. It is highly sensitive to class imbalance but does not account for the cost or benefit of each type of classification error, which can be critical in certain applications.

On the other hand, Absolute and Relative Usefulness metrics are designed to assess the practical impact of model predictions in real-world scenarios. These metrics consider the

Table VI.3 Threshold Analysis Table Transposed

Model	F1-score	Loss	Absolute Usefulness	Relative Usefulness
Logistic Regression	0.0465	0.0015	-0.0011	-2.8988
Lasso Regression	0.1227	0.0004	0.0000	0.0000
K-Nearest Neighbors	0.7342	0.0002	0.0001	0.3788
Support Vector Machine	0.5778	0.0007	-0.0003	-0.7771
Decision Tree	0.7579	0.0003	-0.0004	-0.9405
Random Forest	0.7579	0.0003	0.0000	0.1030
AdaBoost	0.3261	0.0009	-0.0005	-1.3238
Gradient Boosting	0.7561	0.005	-0.0001	-0.3149
Extreme Gradient Boosting	0.8276	0.0003	0.0000	0.1199
Light Gradient Boosting Machine	0.7937	0.0004	0.0000	-0.0671
CatBoost	0.6761	0.0005	-0.0001	-0.2032

implications of false positives and false negatives, effectively weighing the utility of model outputs in a more contextual manner. For example, a model like Decision Tree showing high F1-scores suggests good balance in precision and recall, yet its negative Usefulness scores indicate that the cost of its errors might outweigh the benefits, possibly due to misclassifying critical instances that have high consequence in practical applications. This discrepancy highlights the limitation of relying solely on traditional accuracy measures like F1-scores for evaluating model performance, especially in complex scenarios where decision-making costs are significant. It underscores the importance of incorporating utility-based metrics into model evaluation to capture the broader impact of model predictions, ensuring that the models not only predict accurately but also contribute positively to the intended outcomes.

Consequently, we can confidently assert that the results obtained from the original data are indeed reliable. Moreover, the variables selected by our most high-performance models are pivotal contributors to the overall predictive accuracy. This underscores the importance of considering multiple evaluation metrics to ensure the integrity and reliability of analytical outcomes.

Limitations

While this study strives to contribute valuable insights to the field, it is essential to acknowledge and address certain limitations that may impact the scope and generalizability of the findings. Two primary constraints, namely data limitations and computational constraints, are crucial to consider.

7.1. Data Limitations

In economic research, the quality of findings depends significantly on the underlying data. This study, relying on existing datasets, faces several notable limitations. One critical issue is the presence of missing values (NaNs). Understanding the origins of missing data is crucial, as it can impact the analysis and lead to incorrect conclusions. While the method discussed in "Missing data replacement" can improve model performance, the improvements may be artificial if the imputation does not accurately reflect the true underlying data distribution. Consequently, interpretations and conclusions should be approached cautiously, acknowledging potential biases or limitations in the dataset. Additionally, datasets often span extended periods and diverse geographic regions, introducing heterogeneity. Aggregating data without considering these differences can mask important patterns and lead to misleading conclusions.

The accuracy and reliability of the data sources can also vary. Economic data collected from surveys, administrative records, or financial reports may suffer from response biases, sampling errors, or clerical inaccuracies, affecting the robustness of the findings. Another concern is data granularity. Economic datasets might be aggregated at a high level, such as national or regional statistics, which can obscure finer details crucial for specific analyses. The lack of granular data can limit the scope and depth of the research. Moreover, changes in data collection methodologies over time can introduce inconsistencies, complicating the comparison of data across different periods. A comprehensive understanding of these data limitations is essential for a nuanced interpretation of the research outcomes. Addressing these limitations can enhance the robustness and credibility of the findings, leading to more accurate and insightful conclusions in economic research.

7.2. Computational Constraints

The demand for substantial computational resources is a significant aspect of modern research methodologies and analyses. Despite the increasing availability of powerful computing infrastructure, the complexity of certain models and simulations often exceeds our computational

capabilities, resulting in prolonged processing times that hinder the exploration of intricate research questions and the execution of resource-intensive algorithms. Running complex machine learning models such as Gradient Boosting, XGBoost, LightGBM, and CatBoost, especially with extensive hyperparameter tuning via Grid Search, can be particularly time-consuming. Each Grid Search operation involves evaluating numerous combinations of hyperparameters, and for models like Gradient Boosting, this process can take several hours or even days, depending on the dataset size and the number of parameters being tuned.

Ensemble methods like Random Forest and boosting techniques are also computationally intensive due to their iterative nature. Each model in the ensemble requires separate training, and aggregating these models further increases the computational load. In our study, running a single iteration of Random Forest or boosting methods on a moderately large dataset can take several hours, even on modern high-performance computing systems. Support Vector Machines (SVMs) with large datasets and complex kernels can be particularly slow, often taking several hours to complete training due to the quadratic programming involved in finding the optimal hyperplane that separates the classes.

The need for substantial computational resources becomes a bottleneck, especially when working with high-dimensional data or conducting extensive model validation and testing. These constraints highlight the necessity for continued advancements in technology, including more efficient algorithms and more powerful hardware, to overcome such limitations. While the significance of data and computational power in contemporary research is undeniable, the constraints outlined above emphasize the need for a nuanced interpretation of study outcomes. Future endeavors should aim to address these limitations through strategic data collection approaches, advancements in computing technology, and a commitment to refining methodologies for more comprehensive and accurate results. This includes leveraging cloud computing resources, optimizing algorithms for better performance, and employing more efficient data handling and preprocessing techniques to mitigate the impact of these constraints on research productivity and innovation.

Conclusion

In conclusion, the evaluation of models, particularly in contexts where decision-making involves policy considerations, it is essential to recognize the nuanced dependencies underlying performance metrics. While F1 scores are commonly employed to assess a model's precision-recall trade-off, it is crucial to acknowledge their sensitivity to threshold adjustments. However, when assessing the absolute and relative usefulness of models, one must consider not only threshold variations but also the preferences of policy makers. The interplay between threshold and relative preference introduces an additional layer of complexity in decision-making. Consequently, as a robust conclusion, it becomes imperative to compare models primarily based on F1 scores, thereby isolating the effect of models from the relative preferences of policy makers and ensuring a more independent evaluation of their performance. This approach enables a clearer understanding of the models' intrinsic capabilities without being unduly influenced by subjective policy considerations. Based on the F1-score results, the remarkable performance of boosting methods, attributed to their adept iterative learning process, is prominently evident in our evaluation. Boosting almost outperforms all other methods, highlighting its effectiveness in capturing intricate patterns within the data. As we delve into ensemble method, Random Forest emerges as the optimal choice subsequent to the success of boosting models. The ensemble approach, harnessing the collective intelligence of multiple decision trees, proves highly effective in navigating complex relationships within the data. The amalgamation of diverse weak learners, each contributing to different facets of the data, positions Random Forest as the preeminent model in our study, building on the commendable performance of boosting methods.

Furthermore, our study emphasizes the crucial significance of key variables such as Account to GDP, Inflation, and Housing Price Cycle in predicting financial crises, consistently ranking high across diverse models. Additionally, while F1 scores provide valuable insights, our analysis underscores the necessity of incorporating Absolute and Relative Usefulness metrics for a more comprehensive evaluation. This highlights the importance of a holistic approach to model assessment, particularly in policy contexts where decision-making is pivotal.

Moving to the policy decisions, the determination of the threshold τ and relative preference μ remains pivotal in the landscape of financial crisis prediction. As highlighted throughout our analysis, these parameters wield a direct influence on the trade-off between Type I and Type II errors. The judicious selection of the threshold is instrumental in achieving a delicate balance between accurately identifying financial crises and mitigating the risk of false alarms. Simultaneously, the relative preference parameter μ contributes to the nuanced

weighting of costs associated with these errors. The tailored choice of τ and μ must align with the specific priorities, risk appetite, and objectives of decision-makers. The intricate evaluation of the costs and implications of false positives and false negatives serves as a guiding compass for steering predictive models toward outcomes that harmonize with stakeholders' risk management goals in the context of financial-economic crisis prediction.

Appendix 1

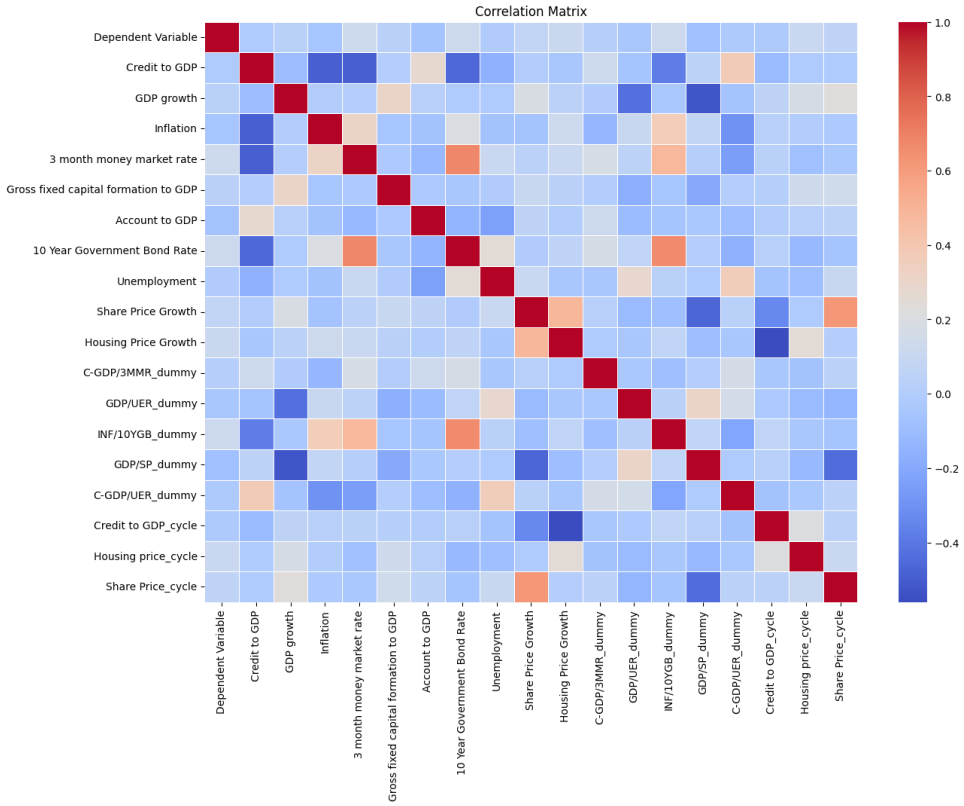


Figure IX.1 Correlation Matrix

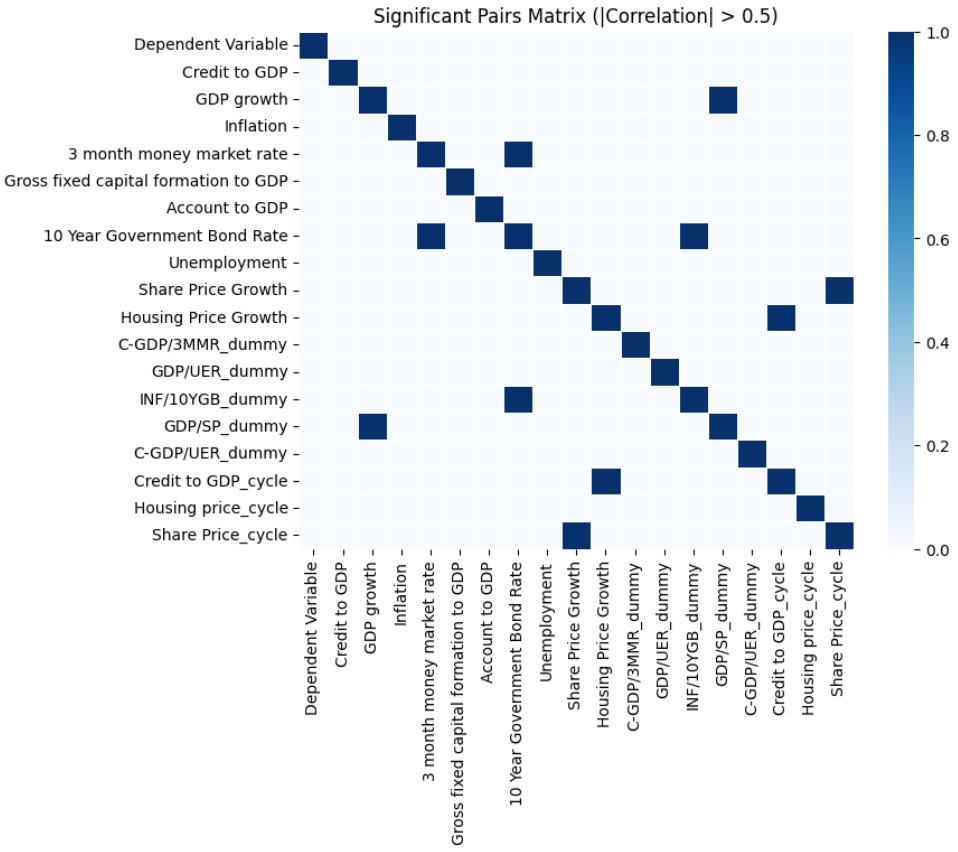
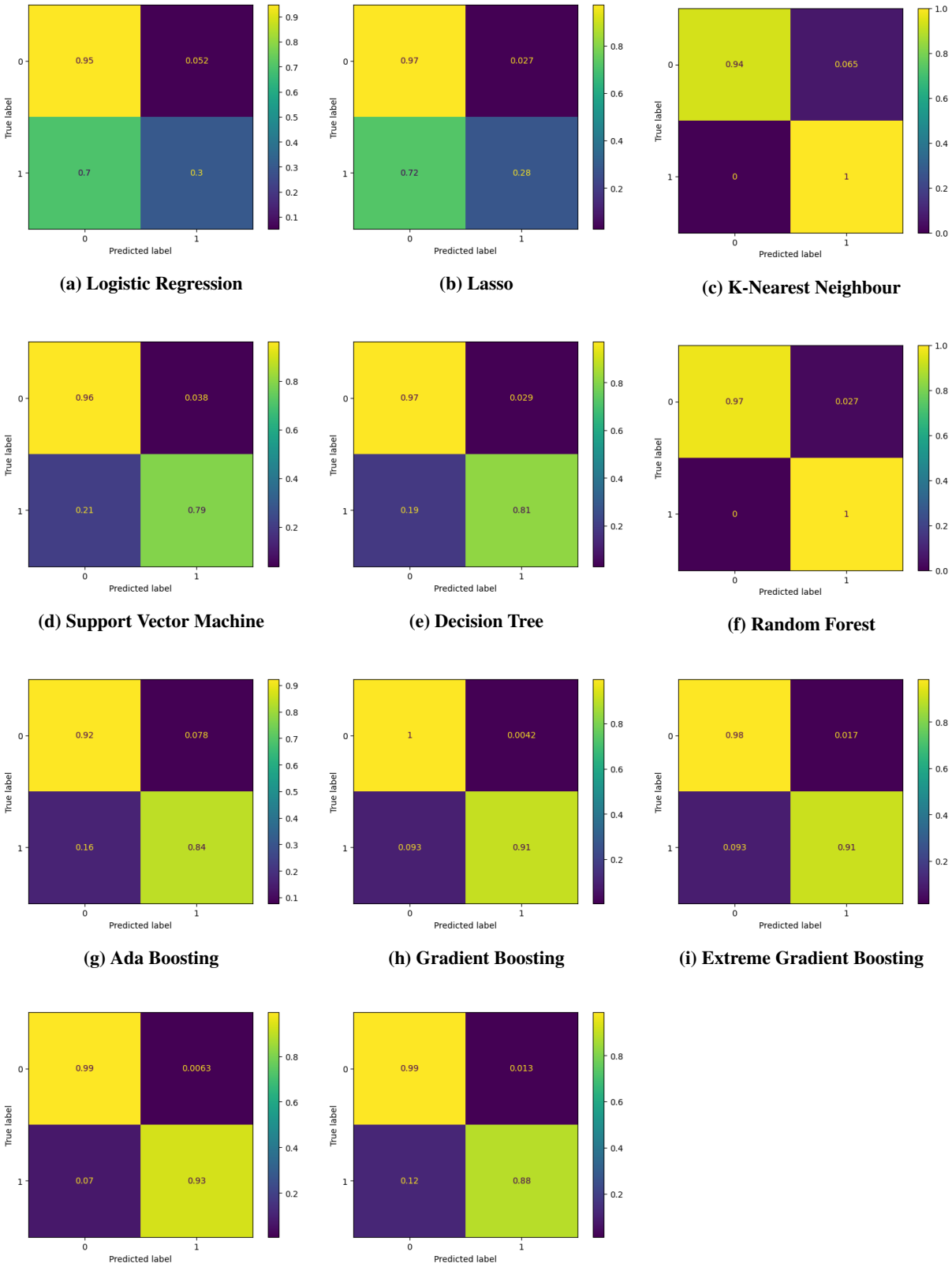


Figure IX.2 Varibles with absolute correlation value more than 0.5

Appendix 2



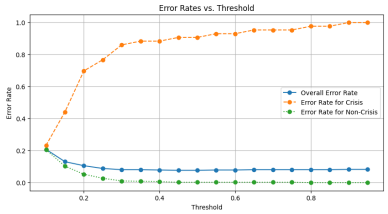
(a) Logistic Regression **(b) Lasso** **(c) K-Nearest Neighbour**

(d) Support Vector Machine **(e) Decision Tree** **(f) Random Forest**

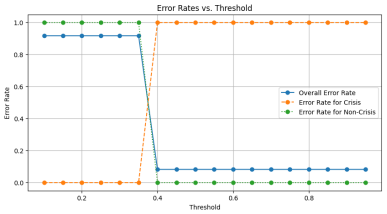
(g) Ada Boosting **(h) Gradient Boosting** **(i) Extreme Gradient Boosting**

(j) Light X-Gradient Boosting Machine **(k) Cat Boosting**

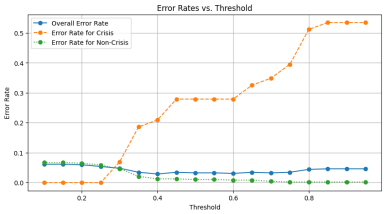
Figure X.1 Confusion Matrices for Different Models



(a) Logistic Regression

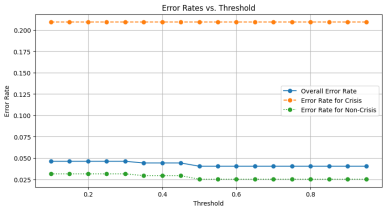


(b) Lasso

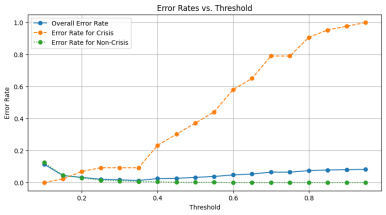


(c) K-Nearest Neighbour

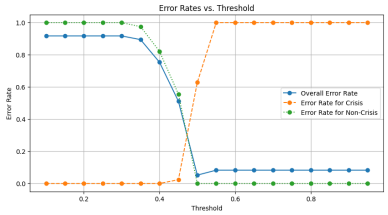
(d) Support Vector Machine



(e) Decision Tree



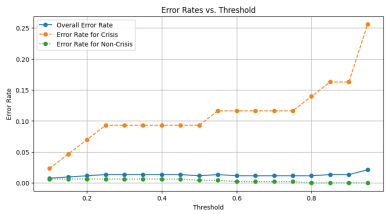
(f) Random Forest



(g) Ada Boosting

(h) Gradient Boosting

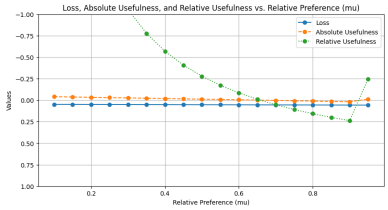
(i) Extreme Gradient Boosting



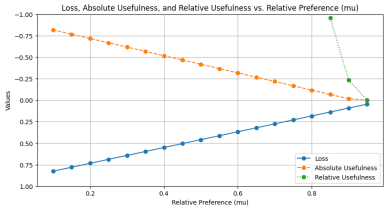
(j) Light X-Gradient Boosting Machine

(k) Cat Boosting

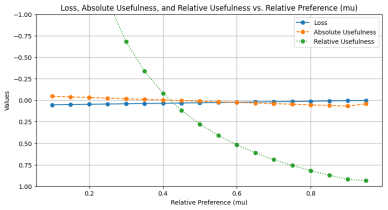
Figure X.2 Error Rate vs. Threshold for Different Models



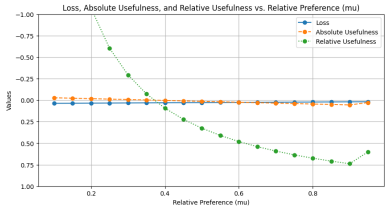
(a) Logistic Regression



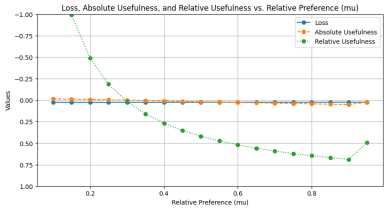
(b) Lasso



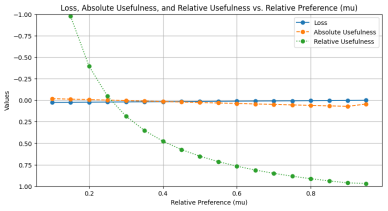
(c) K-Nearest Neighbour



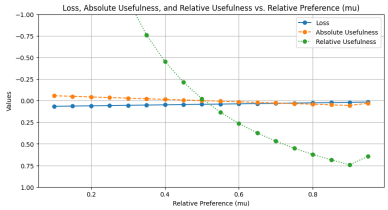
(d) Support Vector Machine



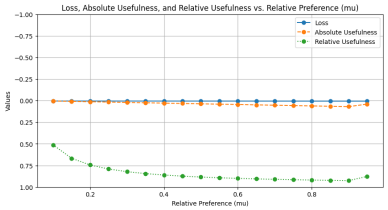
(e) Decision Tree



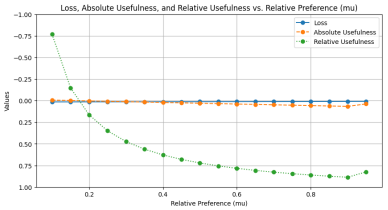
(f) Random Forest



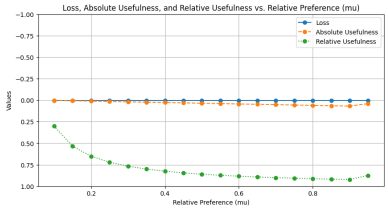
(g) Ada Boosting



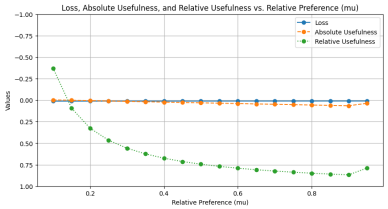
(h) Gradient Boosting



(i) Extreme Gradient Boosting



(j) Light X-Gradient Boosting Machine



(k) Cat Boosting

Figure X.3 Loss, Absolute Usefulness, and Relative Usefulness vs. Relative Preference for Different Models

Bibliography

[Alessi and Detken, 2014] Alessi, L. and Detken, C. (2014). On policymakers' loss functions and the evaluation of early warning systems: Comment. *Economics Letters*, 124(3):338–340.

[Alessi and Detken, 2018] Alessi, L. and Detken, C. (2018). Identifying excessive credit growth and leverage. *Journal of Financial Stability*, 35:215–225.

[Altman, 1992] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

[Babeckỳ et al., 2011] Babeckỳ, J., Havranek, T., Matěj, J., Rusnák, M., Šmídková, K., and Vašíček, B. (2011). early warning indicators of economic crises: evidence from a panel of 40 developed countries. *CNB WP*, 8:2011.

[Babeckỳ et al., 2012] Babeckỳ, J., Havranek, T., Mateju, J., Rusnák, M., Smidkova, K., and Vasicek, B. (2012). Banking, debt and currency crises: early warning indicators for developed countries.

[Baily et al., 2009] Baily, M. N., Litan, R. E., and Johnson, M. S. (2009). The origins of the financial crisis.

[Barrell et al., 2010] Barrell, R., Davis, E. P., Karim, D., and Liadze, I. (2010). Bank regulation, property prices and early warning systems for banking crises in oecd countries. *Journal of Banking & Finance*, 34(9):2255–2264.

[Behn et al., 2013] Behn, M., Detken, C., Peltonen, T. A., and Schudel, W. (2013). Setting countercyclical capital buffers based on early warning models: would it work?

[Benhamou et al., 2021] Benhamou, E., Ohana, J. J., Saltiel, D., and Guez, B. (2021). Detecting crisis event with gradient boosting decision trees.

[Beutel et al., 2018] Beutel, J., List, S., and Von Schweinitz, G. (2018). An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions?

[Bussiere and Fratzscher, 2006] Bussiere, M. and Fratzscher, M. (2006). Towards a new early warning system of financial crises. *journal of International Money and Finance*, 25(6):953–973.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

[Christensen and Li, 2014] Christensen, I. and Li, F. (2014). Predicting financial stress events: A signal extraction approach. *Journal of Financial Stability*, 14:54–65.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

[Dorogush et al., 2018] Dorogush, A. V., Ershov, V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6638–6648.

[Duca and Peltonen, 2013] Duca, M. L. and Peltonen, T. A. (2013). Assessing systemic risks and predicting systemic events. *Journal of Banking & Finance*, 37(7):2183–2195.

[Frankel and Rose, 1996] Frankel, J. A. and Rose, A. K. (1996). Currency crashes in emerging markets: An empirical treatment. *Journal of international Economics*, 41(3-4):351–366.

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning (ICML-96)*, pages 148–156.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

[Fuertes and Kalotychou, 2006] Fuertes, A.-M. and Kalotychou, E. (2006). Early warning systems for sovereign debt crises: The role of heterogeneity. *Computational statistics & data analysis*, 51(2):1420–1441.

[Holopainen and Sarlin, 2017] Holopainen, M. and Sarlin, P. (2017). Toward robust early-warning models: A horse race, ensembles and model uncertainty. *Quantitative Finance*, 17(12):1933–1963.

[Hopp, 2022] Hopp, D. (2022). Economic nowcasting with long short-term memory artificial neural networks (lstm). *Journal of Official Statistics*, 38(3):847–873.

[Jickling, 2009] Jickling, M. (2009). Causes of the financial crisis.

[Kaminsky and Reinhart, 1999] Kaminsky, G. L. and Reinhart, C. M. (1999). The twin crises: the causes of banking and balance-of-payments problems. *American economic review*, 89(3):473–500.

[Ke et al., 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30.

[Knedlik and Von Schweinitz, 2012] Knedlik, T. and Von Schweinitz, G. (2012). Macroeconomic imbalances as indicators for debt crises in europe. *JCMS: Journal of Common Market Studies*, 50(5):726–745.

[Lainà et al., 2015] Lainà, P., Nyholm, J., and Sarlin, P. (2015). Leading indicators of systemic banking crises: Finland in a panel of eu countries. *Review of Financial Economics*, 24:18–35.

[Lang et al., 2018] Lang, J. H., Peltonen, T. A., and Sarlin, P. (2018). A framework for early-warning modeling with an application to banks.

[Mishkin, 1992] Mishkin, F. S. (1992). Anatomy of a financial crisis. *Journal of evolutionary Economics*, 2:115–130.

[Padhan and Prabheesh, 2019] Padhan, R. and Prabheesh, K. (2019). Effectiveness of early warning models: A critical review and new agenda for future direction. *Buletin Ekonomi Moneter Dan Perbankan*, 22(4):457–484.

[Tanaka et al., 2016] Tanaka, K., Kinkyo, T., and Hamori, S. (2016). Random forests-based early warning system for bank failures. *Economics Letters*, 148:118–121.

[Thakor, 2015] Thakor, A. V. (2015). The financial crisis of 2007–2009: Why did it happen and what did we learn? *The Review of Corporate Finance Studies*, 4(2):155–205.

[Tong and Tong, 2022] Tong, L. and Tong, G. (2022). A novel financial risk early warning strategy based on decision tree algorithm. *Scientific Programming*, 2022:1–10.