ISSN: 2755-0214

Journal of Economics & Management Research

SCIENTIFIC Research and Community

Review Article

Open Access

From Data to Decision: Leveraging Machine Learning for Crisis Response

Nijat Hasanli

University of Warsaw, Faculty of Economic Sciences, Quantitative Finance and Machine Learning Department Ludwika Zamenhofa 10A, 122A, 00-187, Poland

ABSTRACT

This study explores the integration of machine learning (ML) techniques in early warning models (EWMs) for financial crises, emphasizing decision-making in policy contexts. By com paring traditional statistical models such as logistic regression and advanced ML techniques, including boosting methods such as AdaBoost, Gradient Boosting, XGBoost, LightGBM, and CatBoost, the research evaluates predictive accuracy and decision-making efficiency. Utilizing a dataset spanning 13 countries over 49 years, this paper highlights key economic indicators such as Account-to-GDP, Inflation, and Housing Price Cycles as critical predictors. The findings underscore the superior performance of boosting models and provide actionable insights for policymakers on optimizing thresholds (τ) and balancing predictive error through relative preference (μ). Specifically, the analysis demonstrates how varying τ and μ influences model effectiveness, highlighting the trade-offs between Type I and Type II errors. This research contributes to enhancing financial stability through informed crisis anticipation and proactive policy interventions.

***Corresponding author**

Nijat Hasanli, University of Warsaw, Faculty of Economic Sciences, Quantitative Finance and Machine Learning Department Ludwika Zamenhofa 10A, 122A, 00-187, Poland.

Received: December 30, 2024; **Accepted:** January 06, 2025; **Published:** January 13, 2025

Keywords: Machine Learning, Crisis Prediction, Early Warning, Financial Stability, Economic Indicators

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 [1]. Excessive indebtedness, speculative investments, banking failures, and unforeseen external disturbances are among the factors contributing to their genesis [2,3]. Understanding these crises requires grasping economic, financial, and psychological elements, including risk-taking tendencies, regulatory deficiencies, and global economic imbalances [4]. Models are developed to help anticipate and potentially prevent these crises by identifying precursory signals, enabling preemptive measures, and mitigating risks. 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 [5,6]. 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 [7,8]. Notably, random forests have emerged as a particularly promising tool, surpassing the effectiveness of conventional models [9].

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 effective- ness of early warning models. Policymakers must determine thresholds (τ) for signaling poten- tial 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

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 and discrete statistical (probit or logit) innovation has been driven by scholars such [10].

They advocate for enhancing early warning predictive precision using random forest, surpassing the power of conventional logit models and the signaling approach have provided further momentum for the transformation of data analytics by identifying an array of machine learning techniques, 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 [11]. 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 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 [12]. 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, which has since undergone further development and expansion [13,14]. The exploration of methodologies for early warning mechanisms in financial crises has garnered significant attention in economic research. The paper delves into the application of Gradient Boosting [15].

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, proposes an early warning system based on signal extraction, aiming to predict the likelihood of financial stress events [16]. 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, evaluates the performance of Long Short-Term Memory Networks (LSTMs) in economic nowcasting, comparing them to the dynamic factor model [17]. 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, 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, proposes a conceptual framework for deriving early- warning models [18]. 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, provides a critical evaluation [19]. 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.

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.

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 3.1: 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.

Feature Engineering

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.

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 under- standing 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 3.2: 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 bal- ance 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 invest- ment 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, λ = 1, 600
Housing price cycle	Deviation from trend of the real residential property price index	Absolute deviation from trend, λ = 1, 600
Share price cycle	Deviation from trend of the real residential property price index	Absolute deviation from trend, λ = 1, 600

Sources: OECD, World Bank, and own calculations

Dummy Dummy Dummy Dummy Dummy
Absolute deviation from trend, λ = 1, 600 Absolute deviation from trend, λ = 1, 600 Absolute deviation from trend, λ = 1, 600Methodology

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) and further developed in (Laina` et al.).

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 $C_{i,t}$

Table 4.1 Contingency matrix for observed outcome and model prediction.

	0	1
Actual Outcome Pi,t 0	True Negative (TN)	False Positive (FP)
1	False Negative (FN)	True Positive (TP)

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. tConsider $C_{i,t}$ 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 pi,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 $\tau \in (0, 1)$:

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

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 Ci,t and the model's predictions of vulnerability Pi,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):

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

$$T_t(\tau) = \frac{FP(False\ Positive)}{\text{FP(False Positive)} + TN(True\ Negative)}$$

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" (Ua) 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 Ua 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" (Ur), we delve into a perspective that assesses the model's effectiveness about the maximum potential benefit. Rather than isolating the model's utility, Ur quantifies its performance as a portion of the ultimate attainable success:

$$U_r(\mu, \tau) = \frac{U_a}{\min[\mu P1,\ (1 - \mu) P2]}$$

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.

Forecasting Horizon

Early-warning models are designed with the primary objective of issuing timely alerts for po- tential 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, argue that it is more beneficial for it to be predetermined to align with the specific task rather than optimized to fit historical data [20].

Recent contributions in the field of forecasting banking crises have adopted various prediction horizons. Studies, such as, have utilized longer horizons ranging from 5 to 12 quarters, while others, like extended the range to 5 to 16 quarters [21]. 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.

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.

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), XG Boost (XGB), Light GBM (LXGB), and Cat Boost. 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 into their practical implications.

Table 4.2: Comparison of Employed Methods: Benefits and Drawbacks Method Benefits 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	Efficient for nonlinear problems,	Can overfit, complex to explain, not inherently probabilistic
Machine (SVM)	computationally efficient	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

Results

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 corre- sponding values chosen for each model. These optimized settings represent the configurations that yielded the highest F1 scores during the tuning process.

Table 5.1 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.

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 5.2: 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.

Why Boosting Models Outperform Others?

- **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.
- **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.
- **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.
- **Advanced Optimization:** Models like XGBoost and LightGBM utilize advanced opti- mization 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.

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 Useful- ness 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 5.3: 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 [22-33].

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.

References

1. 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: 155-205.
2. Jickling M (2009) Causes of the financial crisis. https://www.researchgate.net/publication/37150993_Causes_of_the_Financial_Crisis.
3. Mishkin F S (1992) Anatomy of a financial crisis. Journal of evolutionary Economics 2: 115-130.
4. Baily M N, Litan R E, Johnson M S (2009). The origins of the financial crisis. https://www.brookings.edu/wp-content/uploads/2016/06/11_origins_crisis_baily_litan.pdf.
5. Kaminsky G L, Reinhart C M (1999) The twin crises: the causes of banking and balance-of-payments problems. American economic review 89: 473-500.
6. Knedlik T, Von Schweinitz G (2012) Macroeco- nomic imbalances as indicators for debt crises in europe. JCMS: Journal of Common Market Studies, 50: 726-745.
7. Alessi L, Detken C (2018) Identifying excessive credit growth and leverage. Journal of Financial Stability 35: 215-225.
8. Tanaka K, Kinkyo T, Hamori S (2016) Random forests-based early warning system for bank failures. Economics Letters 148: 118-121.
9. Holopainen M, Sarlin P (2017) Toward robust early- warning models: A horse race, ensembles and model uncertainty. Quantitative Finance 17: 1933-1963.
10. Duca M L, Peltonen T A (2013) Assessing systemic risks and predicting systemic events. Journal of Banking & Finance 37: 2183-2195.
11. Frankel J A, Rose A K (1996) Currency crashes in emerging markets: An empirical treatment. Journal of international Economics 41: 351-366.
12. Babecky` J, Havranek T, Mateju J, Rusna´k M, Smidkova K et al., (2012) Banking, debt and currency crises: early warning indicators for developed countries. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2162901.
13. Alessi L, Detken C (2014) On policymakers' loss functions and the evaluation of early warning systems: Comment. Economics Letters 124: 338-340.
14. Laina` P, Nyholm J, Sarlin P (2015) Leading indicators of systemic banking crises: Finland in a panel of EU countries. Review of Financial Economics 24: 1-835.
15. Benhamou E, Ohana JJ, Saltiel D, Guez B (2021) Detecting crisis event with gradient boosting decision trees. https://hal.science/hal-03320297v1/document.
16. Christensen I, Li F (2014) Predicting financial stress events: A signal extraction approach. Journal of Financial Stability 14: 54-65.
17. Hopp D (2022) Economic nowcasting with long short-term memory artificial neural networks (lstm). Journal of Official Statistics 38: 847-873.
18. Lang J H, Peltonen TA, Sarlin P (2018) A framework for early-warning modeling with an application to banks. https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2182.en.pdf.
19. Beutel J, List S, Von Schweinitz G (2018) An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3312608.
20. Bussiere M, Fratzscher M (2006) Towards a new early warning system of financial crises. journal of International Money and Finance 25: 953-973.
21. Behn M, Detken C, Peltonen T A, Schudel W (2013) Setting countercyclical capital buffers based on early warning models: would it work? https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1604.pdf.
22. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46: 175-185.
23. Babecky` J, Havranek T, Mateˇj J, Rusna´k M, Sˇm´ıdkova´ K et al., (2011). early warning indicators of economic crises: evidence from a panel of 40 developed countries. https://www.cnb.cz/export/sites/cnb/en/economic-research/.galleries/research_publications/cnb_wp/cnbwp_2011_08.pdf.
24. Barrell R, Davis E P, Karim D, Liadze I (2010) Bank regulation, property prices and early warning systems for banking crises in oecd countries. Journal of Banking & Finance 34: 2255-2264.
25. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://dl.acm.org/doi/10.1145/2939672.2939785.
26. Cortes C, Vapnik V (1995) Support-vector networks. Machine learning 20: 273-297.
27. Dorogush AV, Ershov V, Gulin A (2018). Catboost: unbiased boosting with categorical features. In Proceedings of the 32nd International Conference on Neural Information Processing Systems. https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf.
28. Freund Y, Schapire R E (1996) Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning (ICML-96), https://dl.acm.org/doi/10.5555/3091696.3091715.
29. Friedman J H (2001) Greedy function approximation: A gradient boosting machine. Annals of Statistics 29: 1189-1232.
30. Fuertes AM, Kalotychou E (2006) Early warning systems for sovereign debt crises: The role of heterogeneity. Computational statistics & data analysis 51: 1420-1441.
31. Ke G, Meng Q, Finley T, Wang T, Chen W et al., (2017) Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
32. Padhan R, Prabheesh K (2019) Effectiveness of early warning models: A critical review and new agenda for future direction. Buletin Ekonomi Moneter Dan Perbankan 22: 457-484.
33. Tong L, Tong G (2022) A novel financial risk early warning strategy based on decision tree algorithm. Scientific Programming. https://onlinelibrary.wiley.com/doi/10.1155/2022/4648427.