Institut européen
European Institute

LUISS
School of Government

Joint Master in Global Economic Governance and Public Affairs

Beyond the Black Box: An Explainable AI Approach to Sovereign Downgrade Prediction

Supervised by Michel Henry Bouchet

Arthur DESSARD

2024/2025

Thesis pitch

https://youtu.be/KFqi0rakQpA

Statutory Declaration

I hereby declare that I have composed the present thesis autonomously and without use of any other than the cited sources or means. I have indicated parts that were taken out of published or unpublished work correctly and in a verifiable manner through a quotation. I further assure that I have not presented this thesis to any other institute or university for evaluation and that it has not been published before.

26/06/2025 Dessard Arthur

Acknowledgements

I wish to extend my profound thanks to my thesis supervisor, Professor Bouchet, as well as M. Landi and M. Apoteker. Their intellectual guidance, critical feedback, and unwavering support were instrumental throughout this research journey. Professor Bouchet's expertise in country risk challenged me to deepen my analysis and think more critically about the implications of my findings. M. Landi's and M. Apoteker practical insights in term of modelling and data collection kept me motivated and focused, particularly during the challenging phases of the project. I am immensely grateful for their mentorship, their patience, and their commitment to helping me develop as a researcher.

Abstract

This thesis challenges the opaque and lagging nature of traditional sovereign credit ratings by developing and testing a transparent machine learning framework for risk assessment. An XGBoost model is trained on a 24-year panel dataset for 32 emerging markets, with its predictions rendered fully transparent using the SHAP (SHapley Additive exPlanations) framework. The model significantly outperforms a logistic regression baseline and, while functioning primarily as a highly accurate coincident indicator of downgrades, case studies reveal it can provide early warnings in specific contexts of slow, persistent economic deterioration. The primary contribution is a new paradigm for risk management: a "glass box" Sovereign Risk Index (SRI) that offers a transparent, real-time audit of the specific factors driving a country's risk, presenting a clear advantage over traditional methodologies.

Table of Contents

I. Introduction

A global financial system integrated at an advanced level makes sovereign creditworthiness assessments essential for worldwide capital flow management and national economic policies and investment choices. The "Big Three" Credit Rating Agencies (CRAs) such as Standard & Poor's and Moody's and Fitch issue standardized forward-looking opinions about sovereign entities' ability and willingness to fulfil their financial obligations in full and on time. Their influence extends far beyond a simple letter grade; they function as a critical piece of infrastructure in the architecture of global finance.

These ratings function as fundamental risk assessment tools for international investors to help them determine how to allocate capital across the diverse complex global landscape of countries. When investors receive investment-grade ratings they gain access to a wider and more stable global capital market. Institutional investors such as pension funds insurance companies and mutual funds need to meet minimum quality thresholds to fulfil their investment mandates thus restricting their holdings to these security levels[1]. Achieving and maintaining investment-grade ratings represents a top policy objective for sovereign issuers particularly those in emerging markets (Akter, Min Su and Amankwah, 2021). Strong credit ratings enable governments to save money on borrowing expenses because investors reduce their risk premium requirements thus providing additional fiscal resources for public investments and social programs. A positive credit rating enhances both policy credibility and institutional stability which leads to greater foreign direct investment and economic expansion. The "issuer-pays" model of major CRAs which has raised academic concerns about conflicts of interest drives the necessity for objective and independent risk assessment tools[2].

The major CRAs and their rating processes have faced prolonged academic and public examination following major financial crises because their performance has been questioned. The literature consistently shows three fundamental weaknesses in traditional ratings which highlight the necessity for supplemental or substitute risk assessment methods to address these deficiencies.

[1] This practice, often referred to as "cliff effects," is a major feature of the financial architecture. A downgrade from investment-grade (e.g., from 'BBB-') to speculative-grade (e.g., 'BB+') can trigger forced selling by these investors, exacerbating financial instability.

[2] The 'issuer-pays' model, where the rated entity compensates the CRA for its own assessment, has been a subject of extensive regulatory debate regarding potential conflicts of interest, particularly following the agencies' role in rating structured financial products prior to the 2008 Global Financial Crisis.

The major challenge with current rating systems is that they operate with unclear processes. The final rating a sovereign receives results from both economic modelling and subsequent qualitative analysis where expert committee members decide the outcome. The “black box” system combines difficult-to-quantify factors for rating decisions yet creates an opaque process which makes external observers unable to replicate or fully comprehend the outcome (e.g., Fitch Ratings, 2025). The exact distribution of evaluation factors and the logical basis for committee decisions to modify model results together with the specific rating review triggers remain unclear thus creating uncertainty for sovereign-agency policy discussions.

Additionally, there is the critical problem of timeliness. The CRAs face regular criticism because they demonstrate pro-cyclical tendencies (Bar-Isaac and Shapiro, 2011) and delayed responses by making decisions “behind the curve”. Various studies demonstrate that rating changes primarily occur during crises when market prices have already reflected the worsening creditworthiness of a nation. The agencies received widespread criticism because they failed to predict the extent of the 1997 Asian Financial Crisis (Ferri, Liu and Stiglitz, 1999) while their 2008 Global Financial Crisis participation revealed shortcomings in their risk detection capabilities. Their ratings become ineffective as early warning tools because they lag market sentiment which leads them to merely validate existing crises.

Finally, the ratings themselves are discrete. A sovereign entity maintains its rating at a certain level (e.g., ‘BBB-’) across multiple years without showing either the growing vulnerabilities or the positive changes from structural reforms. A rating system with its “step-like” characteristics delivers an incomplete and delayed view of country risk development because ratings may indicate stability while underlying issues continue to worsen until a sudden major downgrade occurs. A combination of these restrictions demonstrates that we need a new method which would deliver transparency along with detailed analysis and immediate results beyond what traditional approaches provide.

The documented shortcomings of traditional rating systems make it necessary to develop analytical tools that use dynamic approaches and data-driven methods with full transparency. The current breakthroughs in computational statistics and machine learning (ML) give researchers the chance to build analytical instruments. The XGBoost algorithm in this study demonstrates how non-linear ensemble methods in ML models analyse complex patterns across multiple datasets while avoiding traditional econometric model limitations. ML algorithms can automatically identify non-linear relationships between a fiscal deficit and sovereign risk which manifests differently at various debt levels. Through their feature analysis capabilities these

techniques generate ongoing risk measurements which advance letter-grade ratings by providing detailed risk assessments of sovereign entities throughout their risk evolution.
The implementation of sophisticated ML models in critical financial domains encounters opposition because of their obscure decision-making processes. A predictive model that lacks transparency through its decision-making process exchanges one form of opacity with another which hinders its value for risk assessment understanding among policymakers and investors.
The emergence of Explainable AI (XAI) simultaneously addresses the "black box" problem that complex models in finance experience. The field of artificial intelligence focuses on developing methods which explain the decisions made by complex models. Modern artificial intelligence techniques including SHAP (SHapley Additive exPlanations) enable users to access black box decision-making processes through its cooperative game theory foundation (Lundberg and Lee, 2017). The SHAP method enables users to determine the exact feature contributions that led to increased or decreased risk assessments in each prediction. The combination of XGBoost as a high-performance predictive algorithm with SHAP as a robust interpretation framework allows for the development of "glass box" models that deliver both excellent accuracy and complete transparency and auditability. This thesis relies on the technological and conceptual foundation established by dual capabilities.

The traditional rating paradigm faces recognized limitations which lead to an opportunity for modern computational techniques to improve both predictive accuracy and transparency. The research goes beyond confirming machine learning model downgrades to determine if they can create a superior risk assessment framework. The central research question is defined as follows:

"*Can an explainable machine learning model produce a Sovereign Risk Index that is demonstrably more transparent, granular, and timely than the ratings issued by traditional agencies?"*

The research investigates four essential objectives to determine the following question:

- ➢ The research develops an extensive panel data set which includes 23 emerging market economies and establishes a reliable downgrade event variable through their macroeconomic, fiscal and institutional parameters.
- ➢ The study creates and tests an XGBoost model for sovereign downgrade prediction alongside a Logistic Regression baseline model to evaluate their performance through out-of-sample testing.

- The research applies SHAP as an XAI technique to analyse the predictive model's results which enables the identification of sovereign risk factors throughout both the entire sample dataset and individual country analyses.
- The research applies event study methods to evaluate the timing of model signals for determining its performance in real-time monitoring.

The subsequent sections of this thesis follow a logical order to achieve the research goals. Chapter 2 reviews the relevant academic literature, covering the traditional determinants of sovereign risk, the evolution of predictive models, and the rise of explainable AI, thereby identifying the research gap this study aims to fill. Chapter 3 delivers an extensive description of the data and methodology which explains data collection methods and variable development as well as feature engineering techniques and evaluation and modelling approaches. Chapter 4 presents the core empirical results of the study, including the model performance comparison, the explainability analysis, and the findings of the temporal event study. Chapter 5 provides a summary of the research findings followed by policy and practical implications and limitations of the study together with future research directions.

II. Literature Review

Most research on sovereign credit risk evaluates the factors which affect national rating scores through economic and social and political perspectives. The rating process integrates both quantitative data-based models and qualitative expert assessments which together form the Qualitative Overlay. The distinction between these two components remains essential because the quantitative component provides objective measurements but the qualitative overlay which evaluates future risks and institutional characteristics introduces subjective assessments that remain under academic and regulatory analysis.

a. The Determinants of Sovereign Default: A Review of the Foundational Literature

Numerous academic studies along with institutional publications have worked to establish the fundamental elements that drive sovereign credit risk and default events. The complex nature of sovereign crises prevents researchers from developing a universal model yet multiple primary analysis pillars have received broad theoretical and empirical acceptance. The development of reliable predictive models requires a solid grasp of fundamental concepts which form four interconnected areas including public finance sustainability and external exposure and domestic economic performance and institutional quality.

Fiscal variables serve as the most established and obvious foundation to evaluate a government's capacity to fulfil its outstanding financial obligations. The government faces an intertemporal budget constraint which establishes that debt levels today should match the present value of upcoming primary surpluses. The uncertainty regarding a government's ability and political motivation to produce required surpluses leads to sovereign risk formation. The Government Debt-to-GDP ratio stands as the principal tool for evaluating this metric. Reinhart & Rogoff (2009) present in their seminal research on financial crisis history that high levels of public debt serve as a robust indicator for default risks. The authors introduce debt intolerance as a concept which explains why emerging markets experience financial crises at debt-to-GDP ratios lower than those of advanced economies because their institutions are weaker and their economic histories are less stable. The analysis requires assessment of both debt amount and its distribution throughout the portfolio. The concept of "original sin" as described by Eichengreen, Hausmann, & Panizza (2005) shows that foreign currency-denominated debt poses major risks because it leads to destructive balance sheet mismatches when the domestic currency weakens. Short-term maturity debt composition creates a significant rollover risk problem because governments need to continuously access market financing to service their outstanding obligations. Assessing debt trajectory depends on both stock variables and flow variables of fiscal nature. The Fiscal Balance (% of GDP) and primary balance (% of GDP) show the present fiscal policy orientation as well as the amount of new debt formation. Sustained primary deficits indicate an unworkable financial situation which will end in either debt default or excessive inflation unless corrected.

The assessment of external vulnerabilities stands as the second essential pillar which stems from first-generation crisis models developed by Krugman (1979). The models of the time showed that policy inconsistencies between fixed exchange rates and fiscal expansion would inevitably result in speculative attacks which would deplete foreign reserves. The critical indicator of Foreign Exchange Reserves exists as two separate measurement tools which include import coverage duration and external debt obligations under the Guidotti-Greenspan rule[3]. The reserves function as an essential defence mechanism to support debt service payments

[3] The Guidotti-Greenspan rule is an influential rule of thumb in international finance which suggests that a country should hold foreign exchange reserves equal to its total short-term external debt (debt with a remaining maturity of one year or less). This ensures it can withstand a sudden stop of capital flows for a full year without needing to access capital markets.

and currency defence against foreign capital flow interruptions known as “sudden stops”. A persistent Current Account Balance (% of GDP) serves as a fundamental indicator of external risk because it shows structural foreign capital needs to bridge the investment and savings gap between nations. A nation becomes highly exposed to global risk mood shifts when it depends on volatile “hot money” portfolio flows instead of stable Foreign Direct Investment (FDI) for financing its needs. The balance of payments for many emerging markets and commodity exporters depends heavily on the Terms of Trade which represents the export-import price ratio because external shocks can rapidly affect their payment balances.

The performance of the domestic macroeconomy functions as a fundamental factor which determines both revenue generation and social stability maintenance for sovereign nations. The most essential economic indicator in this category is Real GDP Growth rate. An expanding economy strengthens taxation capabilities while making the Debt-to-GDP ratio more favourable through base expansion and simultaneously supports fiscal consolidation efforts. When an economy experiences a recession, its finances suffer because tax income decreases while public spending increases thus leading to major public financial stress. The degree and fluctuation of inflation serve a dual-purpose function. Most observers interpret high and volatile inflation rates as indicators of unstable macroeconomic management and poor policy practices even though moderate expected inflation helps reduce the value of debt denominated in local currency. The increase of government debt through seigniorage financing (printing money) leads to central bank credibility deterioration and causes capital flight.

A substantial collection of literature demonstrates that economic factors by themselves fail to determine sovereign risk properly. Economic policy depends entirely on the quality of a country’s institutional framework and political structure for its establishment and preservation. North (1990) developed the concept that institutions function as basic rules which determine economic incentives. The willingness of governments to fulfil debt obligations directly depends on the factors of Rule of Law and Control of Corruption within the framework of sovereign risk. The respect of property rights along with contract enforcement by a government makes it more likely to fulfil its sovereign financial commitments. Butler and Fauver (2006) show that institutional quality independently affects credit ratings while other macroeconomic variables are fully controlled. Political Stability represents an essential factor because social unrest and government instability together with civil conflicts produce major doubts about economic policy continuity and future debt payment obligations. The Worldwide Governance Indicators (Kaufmann, Kraay and Mastruzzi, 2010) developed by the World Bank serve as the primary quantitative metrics to measure fundamental yet slowly evolving determinants of sovereign creditworthiness.

b. The Three Generations of Crisis Models

A comprehensive understanding of modern sovereign risk modelling approaches requires analysis within the historical development of economic financial crisis theory. The academic research about this subject consists of three well-defined crisis model generations. Each crisis generation based on historical events developed separate understandings of sovereign distress through different variable and mechanism analysis.

Krugman (1979) established first-generation models to explain currency crises that occurred in Latin America by showing how policy inconsistencies naturally lead to financial crises. The first-generation model features a government that maintains a fixed exchange rate system yet continues running fiscal deficits which it supports through domestic credit expansion through monetary printing. The central bank must constantly purchase foreign exchange to support the currency peg, so the monetary expansion creates a gradual depletion of its reserve assets. Rational speculators can determine when the central bank will run out of foreign exchange reserves by analysing the unsustainable financial path. Speculators initiate attacks on the currency before the central bank exhausts its reserves which forces an abandonment of the exchange rate peg. According to this generation crises emerge directly from worsening macroeconomic conditions. The main variables of interest include the fiscal deficit as well as the rate of domestic credit growth together with the level of foreign exchange reserves.

Second-generation models emerged due to European and Mexican currency crises during the early 1990s because these events happened in nations with good economic fundamentals. The research of Obstfeld (1996) and other authors incorporated self-fulfilling prophecies and multiple equilibria into their models. This framework shows that a nation's economic fundamentals exist between absolute strength and complete weakness. A nation can exist within a "zone of vulnerability" which allows potential crises yet does not guarantee their occurrence. The outcome depends on what investors expect to happen. A shift in investor sentiment towards currency devaluation triggers them to sell their assets which creates downward market pressure on the currency. To protect the currency, peg the government must increase interest rates until the levels become both economically and politically distressing thus causing an economic depression. When governments face this trade-off, it becomes optimal for them to exit the peg and devalue which confirms the market's initial negative expectations. The main contribution of this generation is that market sentiment changes along with investor expectations can lead to economic crises which demonstrate both the unstable nature of capital flows and the possibility of abrupt “stops” without fiscal or external account deterioration.

The Asian Financial Crisis from 1997-98 demonstrated that corporate and financial sector problems rather than fiscal waste led to third-generation models (Krugman, 1999). The analytical approach of these models focused on analysing financial instability and balance sheet problems that exist within the domestic economy. The literature demonstrates that severe currency and maturity mismatches create significant economic risks. The term “currency mismatch” describes situations when firms or banks carry foreign currency debt while their local currency assets and revenues exist in their portfolio. A currency depreciation creates a real term explosion of debt that results in numerous business failures. Maturity mismatches in banking refer to the practice of using short-term funding to support long-term assets. The sudden disappearance of short-term funding creates a liquidity crisis for financial institutions. According to this generation’s findings a weak financial system alongside poor corporate governance practices can cause a crisis through variables such as credit booms and asset bubbles and financial sector debt structures.

Our empirical framework covers all aspects of financial crises by choosing variables that relate to traditional economic and external factors and proxies for institutional quality and financial stability.

c. Econometric and Machine Learning Approaches to Predicting Crises

Theoretical definitions of risk drivers have been parallelled by an empirical literature which employs these variables to develop predictive models of sovereign crises and defaults and rating changes. This field of study has undergone substantial development during the last thirty years because of better statistical approaches and enhanced computing capabilities. The development progresses from conventional parametric econometric methods toward modern non-parametric machine learning techniques.

Research in the first wave focused on building Early Warning Systems (EWS) during the 1990s emerging market crisis period by utilizing discrete choice models. Kaminsky, Lizondo, and

Reinhart (1998) established the foundational “signals” method for this field of study[4]. Each economic indicator received monitoring status according to their framework until an indicator passed a specific threshold to generate a “signal”. The number of signals was then used to assess the probability of a crisis. The field adopted Probit and Logit models as its standard tools for research after Kaminsky et al. (1998) because these models were adopted by staff members at the International Monetary Fund and the World Bank. The core set of predictive variables that researchers at institutions such as the International Monetary Fund and the World Bank identified as statistically significant emerged from these studies led by Frankel and Rose (1996) and Berg and Pattillo (1999). The authors confirmed through their research that real exchange rate overvaluation together with rapid domestic credit growth and current account deficits and low foreign exchange reserves functioned as dependable crisis indicators. The main benefit of these models stems from their ability to provide straightforward measurements because logistic regression coefficients reveal the exact impact of each variable on crisis probabilities. The primary restriction of these models stems from their linear relationship assumption which does not capture the intricate non-linear economic system dynamics.

The second phase of research uses Machine Learning (ML) algorithms as a solution to overcome the limitation of linear prediction. Research within this literature stream evaluates the predictive power of traditional econometric methods by testing them against non-parametric models which include Support Vector Machines (SVMs), Random Forests and the contemporary Gradient Boosting Machines (XGBoost). Studies by Alessi & Detken (2011) and Bluwstein et al. (2023) among others show that Machine Learning models consistently outperform econometric models when evaluating out-of-sample predictive performance. Their ability to automatically detect complex non-linearities and interaction effects—for example, the idea that a fiscal deficit might only become dangerous when government debt is already high—gives them a distinct advantage in forecasting.

The superior predictive capabilities of these approaches have historically required users to forfeit their ability to interpret model results. The coefficients from logistic regression are straightforward to understand but decision-making processes behind complex ensemble decision trees such as Random Forests or XGBoost models remain extremely challenging to interpret. The “black box” issue functions as a primary obstacle which prevents these models

[4] In the Kaminsky, Lizondo, and Reinhart (KLR) signals approach, an indicator is said to ‘signal’ a crisis if it crosses a pre-defined percentile threshold of its historical distribution. For example, a signal might be triggered if FX reserves fall below the 10th percentile of their historical range.

from being used in operational policy and investment contexts where understanding the underlying causes of risk signals holds equal or greater importance than the signals themselves. A model which fails to provide explanations about its decision-making process remains hard to trust and both difficult to audit and inappropriate for policy recommendations. Research in Explainable AI emerges as a solution to address the well-known trade-off between model transparency and prediction accuracy.

d. The Rise of Explainable AI (XAI) in Finance

The “black box” problem stands as a major obstacle that hinders the practical use of advanced machine learning models in financial domains that require high stakes. These models deliver superior prediction capabilities, but their lack of transparency leads to major problems regarding trustworthiness and regulatory compliance. The practical use of a prediction requires understanding its underlying logic especially when dealing with sovereign risk because this field affects policy and investment decisions. Operationally and ethically problematic recommendations emerge from models that fail to explain their decision-making processes.

The requirement for transparency stems from multiple underlying factors. Financial regulators worldwide have strengthened their focus on model risk management thus requiring institutions to demonstrate their ability to validate and audit their algorithmic systems. The EU’s GDPR along with similar regulations demonstrate a growing worldwide acceptance that decision-making systems need to provide explanations to users. End-users including policymakers and investment committees require model interpretability for building trust in practical applications. Models which explain their decision-making process gain better chances of getting adopted by existing decision-making frameworks. Explainability functions as an essential diagnostic instrument that helps users evaluate and fix their models. A model produces unexpected predictions only through internal logic inspection to identify whether it has discovered a new relationship or trains on artificial correlations in the data.

The current research landscape shows a challenging relationship between achieving model performance and maintaining interpretability. A choice exists between using a transparent logistic regression model that yields lower predictive results and a high-performance gradient boosting machine that lacks transparency. The new field of Explainable AI (XAI) has emerged to address this exact challenge. XAI aims to create multiple techniques which provide clear and consistent and understandable explanations for complex model outputs to convert them from “black boxes” into “glass boxes”.

The XAI landscape features SHAP (SHapley Additive exPlanations) as a highly effective and theoretically sound method which Lundberg and Lee (2017) introduced. SHAP uses Shapley values from cooperative game theory to create a single framework for model prediction interpretation. The main benefit of SHAP exceeds previous feature importance measures because it produces explanations which remain both globally coherent and locally precise. SHAP determines the exact marginal effect of each feature on a prediction by showing how individual feature values influence the prediction compared to the baseline value of the dataset average. This method enables users to conduct a detailed risk audit across every observation in the dataset. The local explanations generated by SHAP can be combined to establish a solid measure of worldwide feature importance which shows the primary drivers behind model decisions. These new techniques enable users to access complete predictive capabilities of advanced machine learning models without compromising transparency requirements for the first time.

e. Research Gap and Contribution

The academic literature offers detailed information that explains the sovereign risk landscape in its entirety. The research community has established a strong agreement regarding the fundamental macroeconomic elements alongside fiscal variables and institutional factors which impact creditworthiness. The academic community has developed a solid body of work that demonstrates how machine learning approaches deliver better out-of-sample predictive results than traditional econometric models. The research indicates that predictive improvement often occurs through black box solutions which reduce model interpretability thus hindering their adoption for policy and investment choices.

Research on Explainable AI (XAI) has started to address this challenge yet a significant gap persists in available literature. The development of a high-performance machine learning model combined with an XAI framework to produce a Sovereign Risk Index which functions as an auditable and transparent replacement for traditional agency ratings remains an unexplored research area. Most empirical studies assessing model performance focus on accuracy measurements rather than signal timeliness and the critical difference between coincident and early warning indicators.

This thesis seeks to bridge the essential knowledge gap that exists where predictive modelling meets explainability and temporal analysis. Our research makes four essential contributions to the field. The research establishes a balance between high accuracy and full transparency by combining XGBoost with SHAP interpretation methods to create a transparent model. The paper implements a complete temporal evaluation approach which combines aggregate event analysis with detailed case examinations to evaluate the model's signal speed. Our model features engineered dynamic features which specifically measure both economic trend patterns and volatility characteristics. The research develops a risk assessment methodology through integrated elements which presents an alternative to static and opaque agency ratings.

III. Data and Methodology

In this Chapter, we provide a detailed description of the variables, feature engineering techniques, and the modelling and evaluation framework employed in this research. The objective is to construct a robust and methodologically sound process for testing whether an explainable machine learning model can provide timely and transparent signals of sovereign credit downgrades. The chapter is organized as follows: Section 3.1 describes the data sources and the sample of countries. Section 3.2 details the process of constructing the dependent variable, our downgrade event flag. Section 3.3 outlines the independent variables selected for the analysis. Section 3.4 describes the feature engineering process used to create dynamic predictors. Section 3.5 presents the modelling framework, and finally, Section 3.6 explains the framework used to evaluate model performance, interpretability, and timeliness.

a. Data Sourcing and Sample

The empirical part of this thesis uses a proprietary database which was constructed solely for this research. The study utilizes a sample of 32 emerging market economies across a time span of 24 years from 2000 to 2024. The selected countries were chosen to give a diverse yet equivalent representation of countries in different geographic regions including Latin America, Emerging Europe, Asia and Africa. The analysed period is particularly advantageous for assessing sovereign risk because it includes multiple global economic cycles including the commodity super-cycle in early 2000s and the Global Financial Crisis in 2008 as well as the subsequent "Taper Tantrum" in 2013 and the oil price shock in 2014 and the COVID-19 pandemic that caused the recent global economic disturbance. The study benefits from multiple stable and unstable economic periods which creates a challenging environment for testing predictive model robustness.

For model development, two different types of raw information were acquired. The first type includes independent variables which function as model predictors. A total of 15 quantitative indicators describing each country's economic and institutional features were obtained from reputable publicly available databases. This data was obtained from the World Development indicators (WDI) and the Global Economic Monitor (GEM) of the World Bank Database; and from the World Economic Outlook (WEO) of the International Monetary Fund. These institutional sources were chosen intentionally because they offer standardized data across countries that scholars commonly use in their research thus ensuring that our research remains transparent and replicable. The selected variables were chosen because they are significant according to both theoretical and empirical research about sovereign default as discussed in the previous chapter.

The second part of data consists of unprocessed data that will be used to generate our dependent variable. A detailed historical archive of all sovereign rating modifications for each nation in our research sample. Data collection was performed from the three main Credit Rating Agencies (CRAs) which are Standard & Poor's, Moody's and Fitch. The researchers documented both the exact alphanumeric rating along with the exact date when the rating action became effective. Our research depends heavily on this detailed historical data because it enables us to detect individual downgrade events that our predictive model needs to forecast. Our goal is to develop a more comprehensive and robust creditworthiness measure by including ratings from all three major agencies instead of using a single source.

The complete list of the 23 countries used in the final analysis is shown in Annex I. The following sections of this chapter explain the complete process that transforms raw data into the analytical dataset used for empirical analysis.

b. Dependent Variable: Defining a Downgrade Event

The construction of the dependent variable for this study requires a systematic approach to transform the discrete, alphanumeric rating actions from three separate agencies into a single, unified, binary indicator suitable for a classification model. This process is fundamental to the analysis, as the definition of the target event directly shapes the problem that the machine learning models are trained to solve. To ensure consistency, comparability, and a conservative risk focus, this process involved three key stages: linearization of the rating scales, unification of the agency ratings, and binarization into a downgrade event flag.

The S&P Moody's and Fitch rating scales received linear mapping to develop a standardized 21-point numeric scale[5] which allows quantitative comparison and mathematical aggregation.

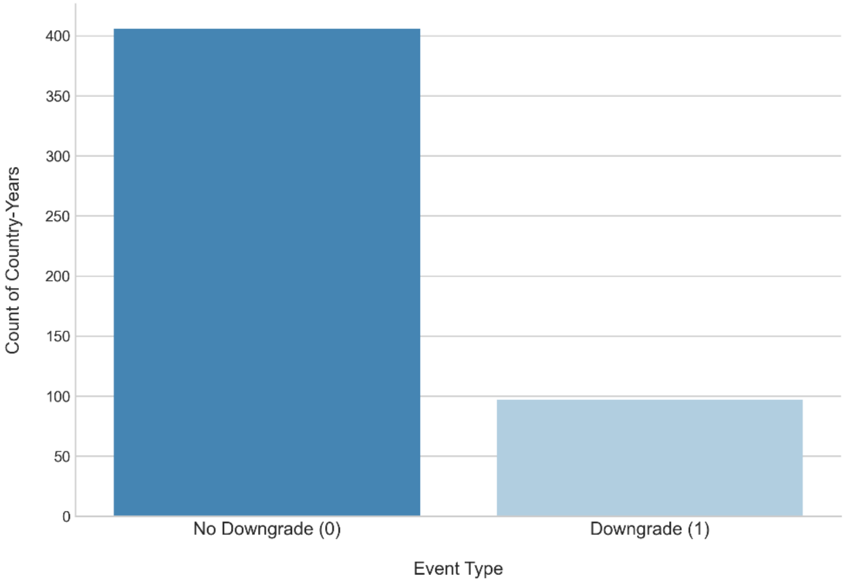
[5] For example, a rating of 'AAA' corresponds to 21 on the scale, 'AA+' to 20, and so on, down to 'D' or 'SD' at the bottom of the scale. This quantitative mapping is a common practice in the empirical credit risk literature to allow for the comparison of different rating systems.

The credit quality indicator increases with rising numbers on this scale so that AAA stands at 21 and AA+ stands at 20 while lower numbers represent increased default risks. The academic literature uses linearization techniques to study credit ratings because they exist as categorical data points.

Second, to create a single, definitive rating for each country-year from our three sources, we established a rule to handle the frequent discrepancies in ratings between the agencies. We adopted a methodologically conservative approach by defining a country's unified annual rating as the minimum (i.e., the most pessimistic) rating assigned by any of the three agencies that was effective on the final day of the calendar year. This choice is grounded in the principle that financial markets and risk managers often react to the most negative available public information, making a "worst-case" indicator the most prudent and relevant choice for a risk-focused model. This ensures our dependent variable is highly sensitive to any signal of deteriorating creditworthiness from any major agency.

Finally, from this unified, numeric time series, we defined our binary dependent variable. As the objective of this thesis is to build an Early Warning System (EWS), our focus is on the dynamic event of a downgrade rather than on predicting a static rating level. The downgrade_event variable was therefore created to take the value of 1 if a country's unified rating in year t is strictly lower than its unified rating in the preceding year t-1, and 0 otherwise. This is our dependent variable in the binary classification problem. After this full data processing pipeline was completed, the final dataset used for modelling contains 97 such downgrade events across our sample period. The specific timing and distribution of these events are detailed further in Annex I.

Figure 1: Distribution of Downgrade Events in Final Model Sample



Source: Author's Calculations

c. Independent Variables and Feature Engineering

The predictive power of any model relies on the selection and preparation of independent variables. The process of developing features required multiple stages which integrated theoretical importance with predictive strength and methodological standards. Our initial step involved choosing numerous core predictor variables that stem from economic literature findings. Our feature engineering process developed advanced dynamic indicators that extracted essential time-series characteristics from sovereign risk.

i. Core Predictor Variables and Data Handling

The foundation of our model is a set of 15 core independent variables, selected based on their established significance in the literature on the determinants of sovereign default, as reviewed in Chapter 2. Here is a detailed description for each of these core variables in Table 1.

Table 1: Core Independent Variables - Definitions, Rationale, and Expected Impact

Variable Name	Description & Rationale	Source
Institutional & Structural		
Political Stability	Index measuring perceptions of political instability and/or politically motivated violence. Higher stability is crucial for policy predictability and investor confidence.	World Bank WGI
Rule of Law	Index reflecting the quality of contract enforcement, property rights, the police, and the courts. Strong rule of law reduces uncertainty and is key to a government's perceived willingness to honour its obligations.	World Bank WGI
Control of Corruption	Index measuring perceptions of the extent to which public power is exercised for private gain. Low corruption signals institutional strength and fiscal discipline.	World Bank WGI
Government Effectiveness	Index measuring the quality of public services and the credibility of the government's commitment to policies. Higher effectiveness suggests a greater capacity to manage the economy.	World Bank WGI
Regulatory Quality	Index measuring perceptions of the government's ability to formulate and implement sound policies that permit and promote private sector development.	World Bank WGI
Voice and Accountability	Index measuring perceptions of a country's citizens' ability to participate in selecting their government, as well as freedom of expression and association. A proxy for institutional maturity.	World Bank WGI
GDP per capita (PPP)	Gross Domestic Product per person, adjusted for purchasing power parity. A primary proxy for a country's level of economic development, wealth, and capacity to absorb shocks.	World Bank/IMF
Macroeconomic Performance		
Real GDP Growth (%)	The annual percentage change in inflation-adjusted GDP. A primary indicator of economic health and the capacity to generate revenue to service debt. A recession is a major predictor of distress.	World Bank/IMF
Inflation (CPI, %)	The annual percentage change in the Consumer Price Index. High and volatile inflation can signal economic instability, poor monetary policy management, and can erode debt sustainability.	World Bank/IMF
Fiscal Health		
Government Debt (% of GDP)	The gross stock of general government debt as a percentage of GDP. The primary measure of a sovereign's accumulated debt burden and its overall fiscal vulnerability.	IMF
General Government Revenue (% of GDP)	Measures the total revenue collected by the government as a share of the economy. It is a key indicator of the government's capacity to service its debt.	IMF
Fiscal Balance (% of GDP)	The government's budget surplus or deficit as a percentage of GDP (also known as Net Lending/Borrowing). A persistent deficit indicates an increasing debt trajectory.	IMF
Interest Payments (% of Revenue)	The share of government revenue dedicated to servicing interest payments. A high ratio indicates a constrained fiscal position and high debt service risk.	IMF
External Vulnerabilities		
Current Account Balance (% of GDP)	Measures the net flow of transactions with the rest of the world. A persistent deficit indicates a dependence on foreign financing, increasing vulnerability to a "sudden stop."	IMF
Domestic Credit to Private Sector (% of GDP)	Measures the stock of credit provided to the private sector. While financial deepening is positive, a very rapid increase can signal an unsustainable credit boom and financial fragility.	World Bank

Constructing a complete panel dataset for 23 emerging market economies over a 24-year period inevitably presents challenges with missing observations for certain variables in certain years. Rather than excluding countries or variables—which could risk sample selection bias or omitting important information—we adopted a rigorous imputation strategy. We employed country-specific linear interpolation to fill sporadic gaps. This method respects the time-series nature of the data by estimating a missing value based on the last available prior data point and the next available future data point for that specific country, a technique methodologically sounder than using a simple cross-sectional mean. Any remaining missing data at the very beginning or end of a country's time series was handled with a backfill/forward-fill approach. The summary statistics for the core variables after this imputation process are presented in Table 2.

Table 2: Descriptive Statistics of Core Independent Variables

Variable	Observations	Mean	Std. dv.	Min.	Max.
Political Stability	503	33.23	20.14	1.01	91.01
Rule of Law	503	47.08	17.26	10.43	88.04
Control of Corruption	503	44.54	18.28	8.47	91.22
Government Effectiveness	503	52.47	15.07	14.63	85.44
Regulatory Quality	503	53.59	16.92	7.66	92.72
Voice and Accountability	503	48.44	19.24	2.35	89.42
GDP per capita (PPP)	503	15241.26	11633.10	2090.86	64162.54
Real GDP Growth (%)	503	3.65	3.85	-28.76	14.05
Inflation (CPI, %)	503	6.34	7.16	-2.09	96.10
Government Debt (% of GDP)	503	48.67	21.90	1.54	115.54
General Government Revenue (% of GDP)	503	23.84	7.65	8.27	44.81
Fiscal Balance (% of GDP)	503	-3.53	3.71	-19.84	11.84
Interest Payments (% of Revenue)	503	13.41	11.39	0.34	79.87
Current Account Balance (% of GDP)	503	-0.74	5.50	-13.70	28.12
Domestic Credit to Private Sector (% of GDP)	503	52.27	33.92	7.13	164.08

Source: Author's Calculations

ii. Dynamic Feature Engineering

While the core variables provide a static snapshot, sovereign risk is inherently dynamic. We hypothesized that the trajectory and stability of these fundamentals are more powerful predictors of future downgrades than their static levels alone. To provide our model with this crucial dynamic context, we undertook a two-step feature engineering process.

The first and most critical step was to ensure the analytical soundness of our model by preventing "lookahead bias". To build a true out-of-sample forecasting model, it is essential that the model's predictions for any given year t are based solely on information that would have been available at the end of the preceding year, t-1. To enforce this principle rigorously across our panel dataset, every independent variable was lagged by one period. This was implemented in Python by first grouping the data by country and then applying a .shift(1) operation to each variable's time series. The result of this procedure is that all features used for model training are t-1 values.

The second step was to explicitly model the concepts of trend and volatility. Using a three-year window, which is long enough to establish a trend but short enough to be responsive to recent changes, we calculated rolling statistics for key lagged variables. The 3-Year Rolling Mean was calculated to smooth out short-term, single-year volatility and provide the model with a clearer signal of a variable's recent trajectory. This helps the model distinguish between a temporary shock and a persistent deterioration. Simultaneously, we calculated the 3-Year Rolling Standard Deviation to provide the model with a direct and quantitative measure of a variable's stability. High volatility in core metrics like GDP growth can itself be a powerful signal of underlying economic instability, even if the average level appears benign. By creating these dynamic features, our final dataset for the models includes not just a snapshot of the economy in the prior year, but also a richer description of its stability and direction of travel.

d. The Modelling Framework

This section outlines the framework used to train and test our predictive models. A rigorous out-of-sample testing methodology is first described, which is designed to provide a robust assessment of the models' true forecasting performance. Following this, we specify the two classification algorithms employed in this study: a traditional Logistic Regression model to serve as a baseline (Berg and Pattillo, 1999), and an advanced Gradient Boosting model (XGBoost) to test the capabilities of a more complex, non-linear approach (Chen and Guestrin, 2016).

A critical component of evaluating any predictive model is a robust out-of-sample testing procedure. A simple random sampling for splitting the data into training and testing sets is inappropriate for panel data with a time-series dimension, as it would lead to "lookahead bias" — whereby information from the future could be used to train a model that is then tested on the past. This would produce an unrealistically optimistic and invalid measure of the model's true predictive power. To avoid this, we employ a strict temporal split. The dataset receives its time-based division at a fixed point so all information before 2019 goes into the training set and the subsequent years go into the test set. This approach duplicates real-world forecasting because the model uses all available historical data until a certain date to generate predictions for completely new unobserved data during the following years. The evaluation process becomes an authentic test of model generalization for future events.

Two distinct models were developed to address our research question. In line with much of the existing academic literature on credit risk and crisis prediction, we first estimate a Logistic Regression model to serve as a performance benchmark. The logistic regression models the probability of a binary outcome—in our case, the downgrade_event—by fitting a linear equation to a logistic function. While less complex than modern machine learning algorithms, its high degree of interpretability makes it a standard baseline. A crucial challenge in downgrade prediction is the inherent class imbalance of the dataset; downgrade events are rare compared to non-events. To address this issue, the logistic regression was implemented using the class_weight='balanced' parameter. This setting adjusts the weights in the loss function to be inversely proportional to class frequencies, in that way placing a higher penalty on misclassifying the minority class (downgrades) and preventing the model from defaulting to predicting the majority class.

To test the capabilities of a state-of-the-art machine learning approach, we then developed a model using the XGBoost algorithm. XGBoost is a highly effective and computationally efficient implementation of gradient boosted decision trees, renowned for its superior performance on structured, tabular data. As an ensemble method, it does not rely on a single model but rather combines the predictive power of hundreds of individual decision trees which are built sequentially, where each new tree is explicitly trained to correct the residual errors of the preceding ones. This process allows XGBoost to capture highly complex, non-linear relationships and feature interactions within the data that a linear model cannot. To manage class imbalance in this model, we utilized the scale_pos_weight parameter, which scales the gradient for the positive class by a factor equal to the ratio of negative to positive samples, ensuring that the model pays significantly more attention to correctly predicting the rare downgrade events during the training process.

e. The Evaluation Framework

To provide a comprehensive assessment of our models, we employ a multi-faceted evaluation framework designed to measure three distinct dimensions of performance: i) predictive accuracy, ii) model interpretability, and iii) the timeliness of its signals. This framework allows us to move beyond a simple measure of accuracy and to directly address the core components of our research question. Given the imbalanced nature of our dataset, where downgrade events are rare, overall accuracy is an insufficient and often misleading metric, as a model could achieve a high score simply by always predicting the majority class (no downgrade). To provide a more nuanced and informative assessment of performance, we therefore rely on three key metrics derived from the confusion matrix. Precision, defined as the ratio of true positives to all positive predictions, measures the reliability of the model's signals; in the context of this study, it answers the question, "Of all the countries the model flagged as being at risk of a downgrade, what percentage were actually downgraded?" Recall, or sensitivity, is defined as the ratio of true positives to all actual positive cases and measures the model's ability to identify all relevant events. It answers the question, "Of all the sovereign downgrades that truly occurred, what percentage did our model successfully identify?" which is a critical metric for evaluating the completeness of an Early Warning System. Finally, the F1-Score, the harmonic mean of Precision and Recall, provides a single, balanced measure of a model's overall performance, which is particularly useful for assessing its effectiveness on imbalanced datasets.

Beyond predictive accuracy, a central objective of this thesis is to ensure model transparency and address the "black box" problem inherent in complex machine learning models. To this end, we employ a state-of-the-art technique from the field of Explainable AI (XAI), SHAP (SHapley Additive exPlanations). As introduced by Lundberg and Lee (2017), SHAP is a methodology grounded in cooperative game theory that computes the marginal contribution of each feature to a specific prediction, assigning a "SHAP value" that quantifies how much that feature's value pushed the model's output away from a baseline average. This allows for a complete "risk audit" of any given prediction. We leverage this capability in two ways: first, through global explanations, where we use summary plots of SHAP values to identify the most important features on average across all predictions, providing a general understanding of the key drivers of sovereign risk according to our model. Second, through local explanations, where we use waterfall plots to decompose individual predictions for our case studies, revealing the specific factors that led the model to its conclusion for a particular country in a particular year.

Finally, to directly test the "early warning" hypothesis, we utilize an event study methodology (MacKinlay, 1997) on the out-of-sample test set. This framework allows us to systematically analyse the model's behaviour in the periods immediately preceding and following a downgrade. The process involves identifying all downgrade events in the test set and aligning them at a common event time, t=0. We then calculate and plot the average predicted risk probability generated by our model for the full window surrounding these events, typically from three years prior (t-3) to one year after (t+1). This aggregate analysis allows for a clear, visual assessment of the model's timeliness. A clear, rising trend in the average predicted risk in the periods before t=0 would provide evidence for a general early warning capability, while a sharp spike only at t=0 would indicate a strong, but purely coincident, indicator.

IV. Empirical Results and Discussion

This chapter presents the empirical findings of the study, bridging the methodological framework outlined previously to a detailed analysis of the results. The chapter is structured to systematically build our argument. We begin in Section 4.1 by conducting a rigorous out-of-sample evaluation of the predictive performance of our two models—the Logistic Regression baseline and the advanced XGBoost classifier. This comparative analysis is crucial for establishing the statistical validity and superiority of the more complex machine learning approach. In Section 4.2, we move from prediction to explanation, employing the SHAP framework to interpret the internal logic of the XGBoost model and identify the key global drivers of sovereign risk. In Section 4.3, we directly address the central research question of this thesis by conducting a detailed temporal analysis of the model's predictions, assessing its capabilities as both a coincident and an early warning system through aggregate event studies and specific case studies. Finally, Section 4.4 provides a broader discussion, synthesizing these empirical findings into a cohesive and nuanced answer to our research question.

a. Model Performance: A Comparative Analysis

Our empirical analysis starts by evaluating the out-of-sample predictive accuracy between traditional econometric Logistic Regression models and advanced non-linear XGBoost classifiers. The models received pre-2019 training data before being tested on the 2019 to 2024 test data set. The evaluation method uses a strict temporal split to verify the models' generalization and forecasting capabilities in real-world scenarios. Our evaluation focuses on Precision, Recall and F1-Score for the positive class (downgrade events) because these metrics provide better performance insights than accuracy alone due to the imbalanced nature of our dataset.

The out-of-sample classification results for both models appear in Table 3. The results demonstrate an obvious trade-off between these two approaches which reveals their distinct advantages and disadvantages.

Table 3: Out-of-Sample Predictive Performance for Downgrade Events

Metric	Logistic Regression	XGBoost
Precision	0.24	0.43
Recall	0.65	0.50
F1-Score	0.35	0.46
Overall Accuracy	53%	77%

The Logistic Regression model, which serves as our academic baseline, achieved a high Recall of 0.65. This result indicates that the model, when calibrated with balanced class weights, is highly sensitive to the minority class and successfully identified 65% of all the true sovereign downgrade events that occurred within the test period. From the perspective of building an early warning system, this high sensitivity is a desirable characteristic, as it implies the model is unlikely to miss many emerging crises. However, this high recall came at a significant cost to precision, which stood at only 0.24. This low precision means that when the logistic model signalled a potential downgrade, it was correct in only one in four instances. The practical implication of this is a high rate of false alarms that can undermine the credibility of a warning system and lead to "alarm fatigue" among policymakers and investors.

On the other hand, The XGBoost model demonstrates a more balanced and efficient performance profile. While its recall of 0.50 is lower than that of the baseline — meaning it correctly identified 50% of the true downgrade events — its precision of 0.43 is nearly double that of the logistic regression. This represents a substantial improvement in the reliability of the model's signals. A user of the XGBoost model can have significantly more confidence that a positive signal represents a genuine risk of a downgrade. This improvement in signal quality is crucial for any practical application. The superiority of this balanced approach is captured by the F1-Score, the harmonic mean of Precision and Recall. The XGBoost model achieves an F1-Score of 0.46, representing a significant improvement over the baseline's score of 0.35. Furthermore, its overall accuracy across all predictions was 77%, far superior to the baseline's 53%.

The confusion matrices presented in Figure 3 and Figure 4 provide a granular, visual representation of these results. The matrix for the Logistic Regression model clearly shows many false positives (54) relative to its true positives (17), visually confirming its low precision.

In contrast, the matrix for the XGBoost model shows a much more favourable ratio, with a significantly lower number of false positives (17) for its 13 true positives.

Figure 2 : Confusion Matrix (Logistic Regression Baseline)

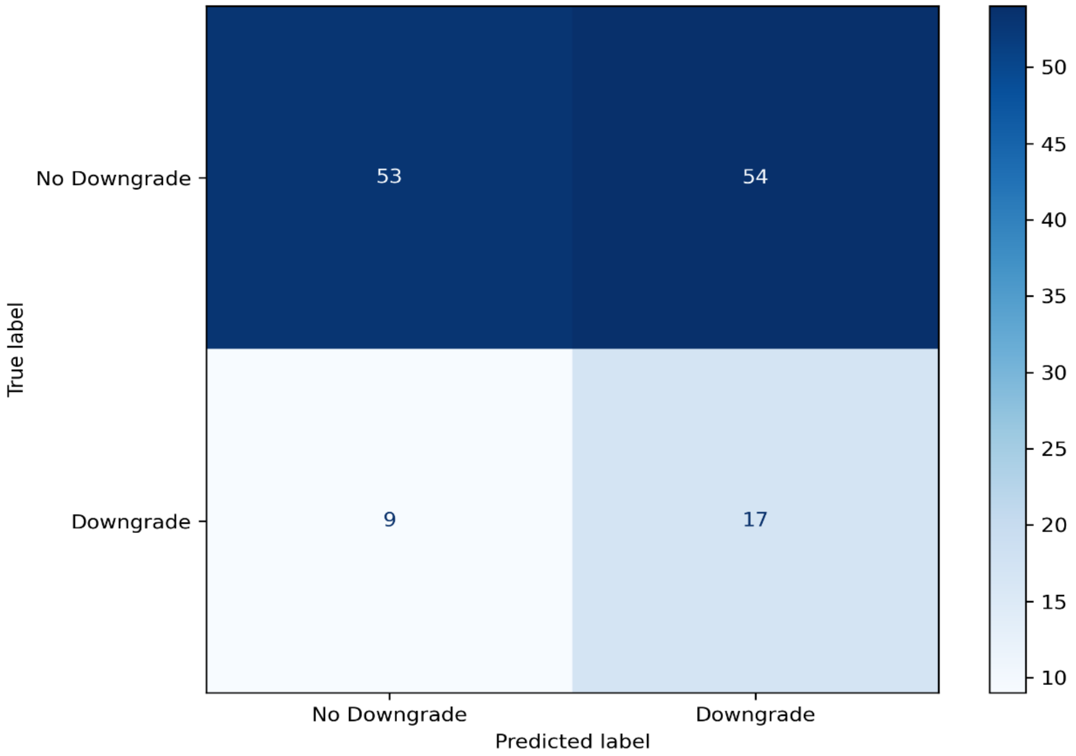
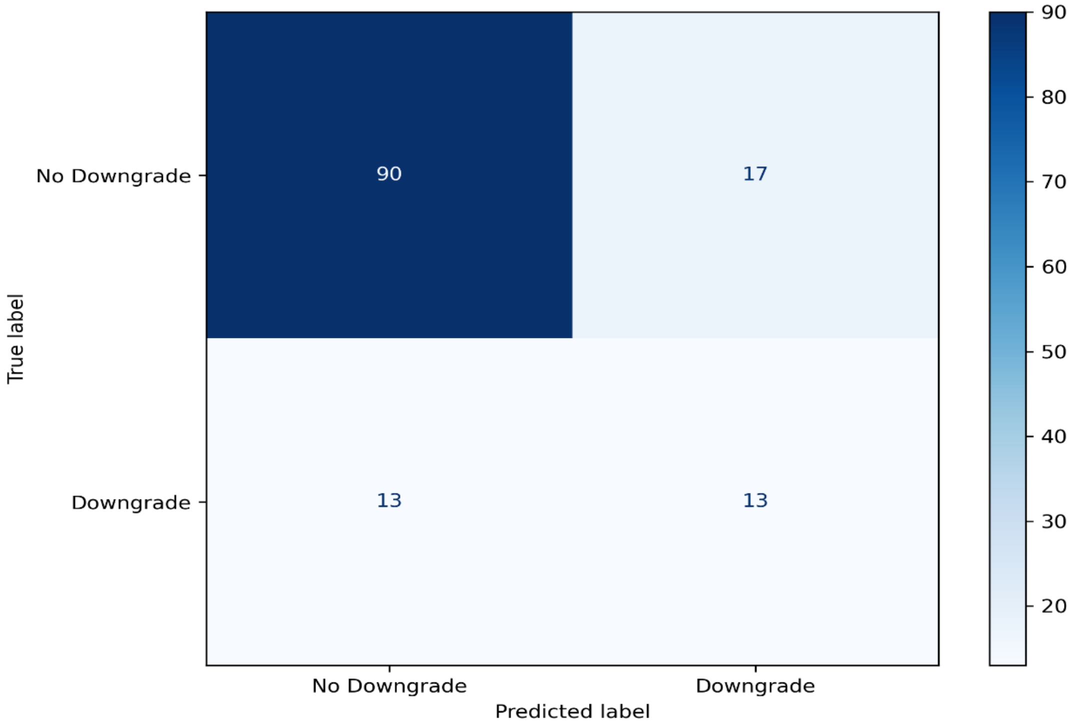


Figure 3 : Confusion Matrix (XGBoost Model)



In conclusion, while the simpler logistic model offers higher sensitivity, the XGBoost model is demonstrably superior for the purposes of this study. It provides a more robust and reliable balance between identifying potential crises and avoiding the destabilizing noise of excessive false alarms. Its higher F1-Score confirms that it is the more effective overall classifier. Therefore, the subsequent interpretability and temporal analyses in this thesis will focus exclusively on the trained XGBoost model to unpack the drivers of sovereign risk and assess the timeliness of its more reliable signals.

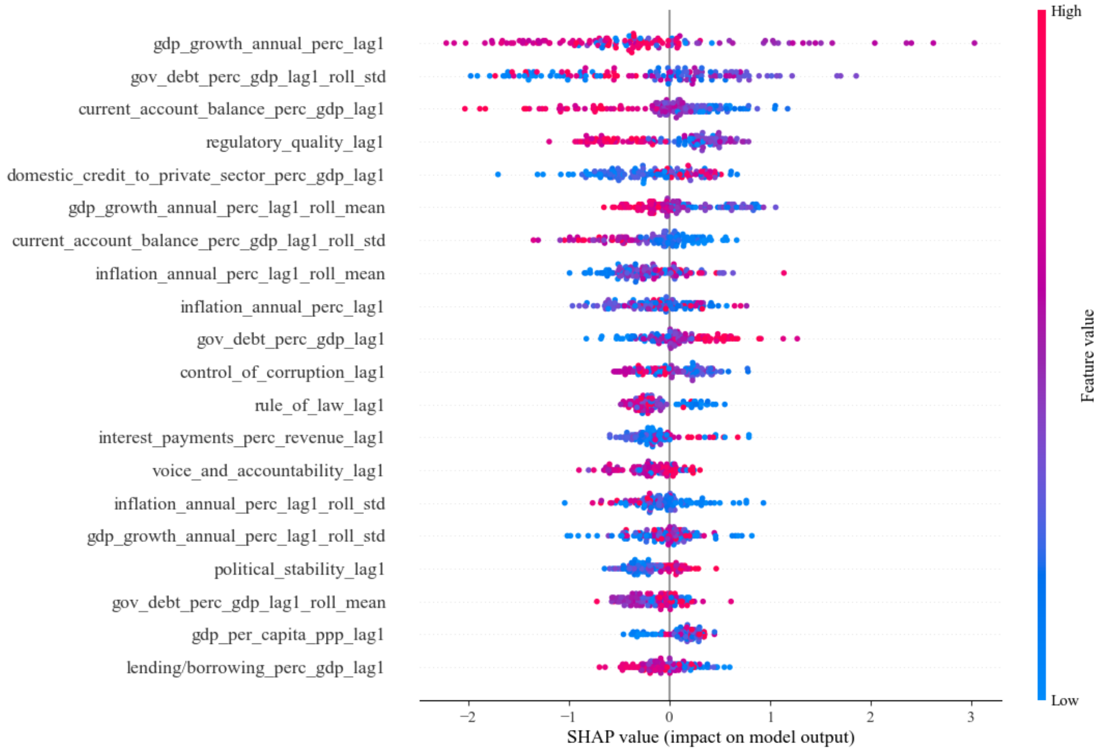
b. Unpacking the "Black Box": Global Risk Drivers

Having established the superior predictive performance of the XGBoost model in the preceding section, we now turn to interpreting its internal logic. A central objective of this thesis is to move beyond a simple assessment of predictive accuracy and to create a transparent "glass box" model whose decision-making process can be audited and understood. A model's ability to generate accurate predictions is of limited practical use if its underlying reasoning is opaque. To address this "black box" problem, we employ the SHAP (SHapley Additive exPlanations) framework[6], a state-of-the-art technique from the field of Explainable AI, to understand the drivers of our model's predictions. This section presents the global feature importance analysis, revealing the factors that, on average, have the greatest impact on the model's assessment of sovereign downgrade risk across the entire out-of-sample test set.

The SHAP summary plot, presented in Figure 5, aggregates the SHAP values for every feature across all predictions in the test set. In this visualization, features are ranked vertically by their mean absolute SHAP value, which represents their overall importance to the model's output. For each feature, the plot displays a scatter of individual SHAP values for each country-year observation. The horizontal position of each point represents that feature's impact on the model's prediction for that specific observation; a positive SHAP value increases the predicted probability of a downgrade, while a negative SHAP value decreases it. The color of each point indicates the feature's value for that observation, with red representing high values and blue representing low values. This allows for a rich interpretation of not only which features are important, but also the directionality of their effects.

[6] As detailed in Chapter 3, the SHAP framework is based on the work of Lundberg & Lee (2017). The SHAP value for each feature represents its marginal contribution to the final model output, which is on a log-odds scale.

Figure 4: Global Feature Importance (SHAP Summary Plot)



A detailed analysis of the plot reveals a set of key insights into the model's decision-making process. The results show that the model has learned a sophisticated and economically sensible set of relationships from the data, strongly aligning with established economic theory and providing confidence in its analytical foundations.

The most influential feature in the model, as evidenced by its top ranking, is the gdp_growth_annual_perc_lag1. The plot shows a clear and powerful relationship: low values for this feature (indicated by blue dots) are clustered on the right side of the plot, corresponding to large, positive SHAP values. Conversely, high values for GDP growth (red dots) are clustered on the left, corresponding to negative SHAP values. This indicates that the model has learned that a contraction or significant slowdown in the real economy in the prior year is the strongest single predictor of a potential downgrade. This aligns perfectly with economic intuition, as weak growth erodes a government's tax base, worsens debt-to-GDP dynamics, and can signal deeper structural problems, all of which impair sovereign debt servicing capacity.

Notably, the analysis reveals that the dynamic features we engineered are among the most important predictors, confirming the value of our feature engineering process. For instance, the gov_debt_perc_gdp_lag1_roll_std ranks highly. The plot for this feature shows that high values (red dots), representing high volatility in a country's debt accumulation path, are associated with positive SHAP values, increasing the predicted risk of a downgrade. This finding suggests the model has learned that it is not just the level of debt that matters, but also its stability and the predictability of fiscal policy. High volatility can signal policy uncertainty or an uncontrolled fiscal situation, which rightly increases perceived risk.

Third, the model assigns high importance to institutional quality. Indicators from the World Bank Governance Indicators suite, such as regulatory_quality_lag1 and rule_of_law_lag1, are prominent in the feature ranking. For these variables, the plot consistently shows that low scores (blue dots) have positive SHAP values, pushing the model's prediction towards a downgrade. This demonstrates that the model has learned the crucial, foundational role that strong institutions play in underpinning sovereign creditworthiness. A stable, predictable, and fair regulatory and legal environment is seen by the model as a significant mitigating factor, even in the face of macroeconomic headwinds.

In summary, the SHAP analysis provides compelling evidence that the XGBoost model is not an unexplainable "black box". It has learned to act like a sound economic analyst, placing the most weight on headline economic performance, the stability of fiscal policy, and the quality of governing institutions. This transparency in its decision-making process increases our confidence in its predictions and its utility as a tool for a deeper, more nuanced analysis of sovereign risk.

c. Answering the "Early Warning" Question: A Temporal Analysis of Model Signals

Having established the predictive accuracy and interpretability of the XGBoost model, we now address the central research question of this thesis: the timeliness of its signals. A key claim for the utility of predictive models in sovereign risk is their ability to function as an Early Warning System (EWS), providing actionable signals prior to an adverse event. In this section, we test this hypothesis rigorously. We first conduct a systematic analysis of the model's average signal behaviour around all downgrade events in the out-of-sample test set. We then proceed with a granular case study to explore the context-dependent nature of the model's timeliness.

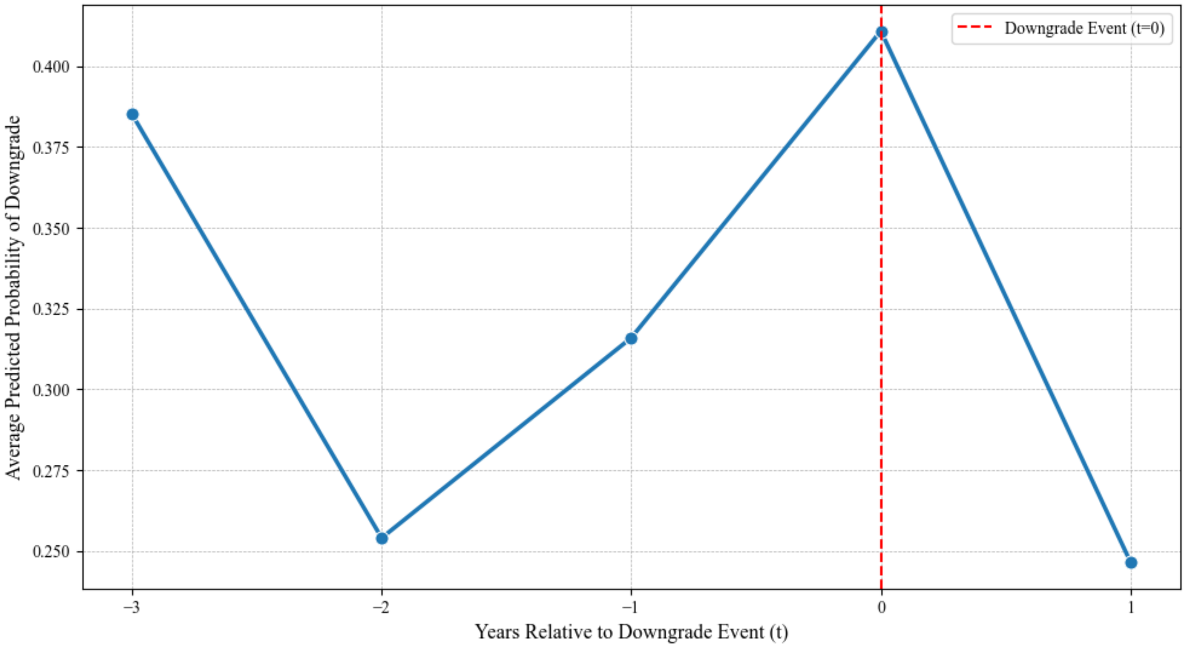
i. Aggregate Analysis: A Systematic Test of the Early Warning Hypothesis

To test the EWS hypothesis in a systematic and generalizable manner, we employ an event study methodology as detailed in Chapter 3. This approach allows us to average the model's behaviour across all downgrade events, revealing the general pattern of its predictive signals in the years immediately preceding and following a rating change. We identify all t=0 downgrade events in our out-of-sample test set and plot the average predicted downgrade probability generated by our XGBoost model for the full window surrounding these events, from three years prior (t-3) to one year after (t+1). A clear, monotonic increase in the average predicted risk in the periods before t=0 would provide strong evidence for a general early warning capability.

The results of this aggregate analysis are presented in Figure 5. The plot reveals a clear and telling pattern. In the three years prior to a downgrade (t-3), the average predicted risk is elevated at approximately 38%, suggesting that countries that are ultimately downgraded are, on average, already perceived as structurally riskier by the model than a typical stable country. However, in the immediate run-up to the event, we do not observe a clear warning signal. The average predicted probability declines at t-2 to 26% before rising modestly at t-1 to 32%. The most significant and unambiguous change occurs at t=0, the year of the downgrade itself, where the average risk probability sharply increases to its peak of 41%. In the year following the event (t+1), the average predicted risk recedes.

This finding provides strong evidence that, on average, the model functions as a highly accurate coincident indicator rather than a consistent, forward-looking early warning system. It reliably distinguishes between event and non-event years, but it does not, in aggregate, provide a clear and actionable warning signal one or two years in advance. This suggests that for many "sudden-stop" style crises, which may be triggered by sharp, unexpected shocks, the underlying annual economic data used in this study only reflects the full extent of the damage in the year the shock occurs, thus limiting the predictive horizon of any model built upon it.

Figure 5: Average Predicted Risk Around Downgrade Events

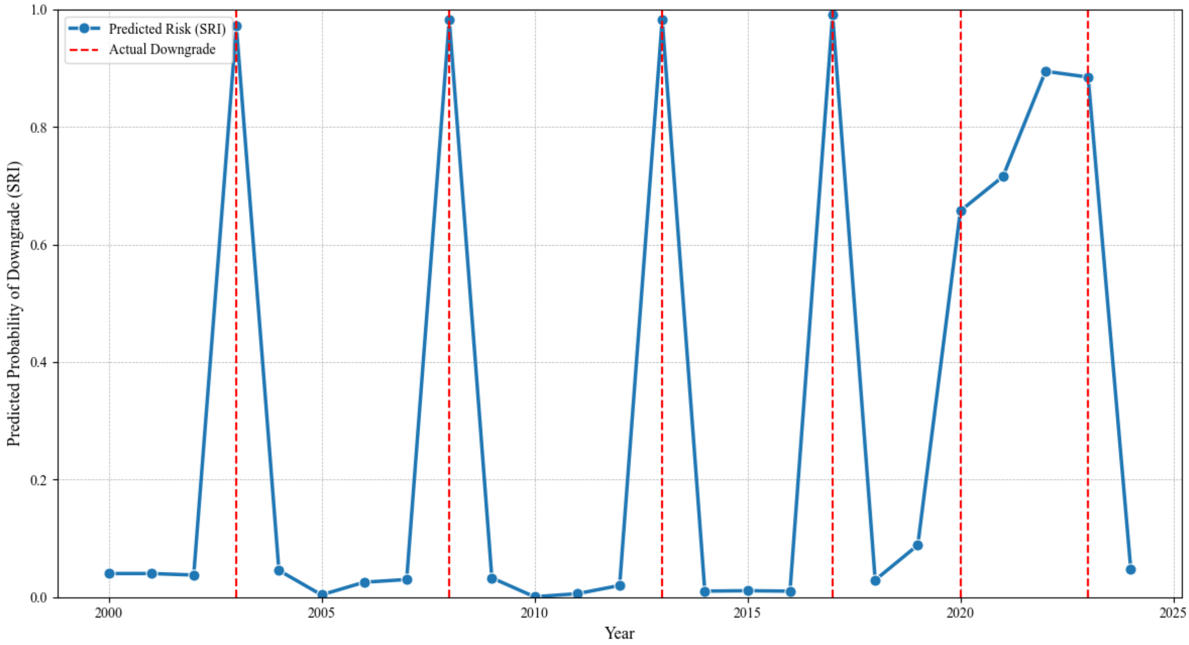


ii. Decomposing the Average: Case Study Evidence of Context-Dependent Timeliness

While the aggregate analysis provides a powerful general conclusion, it may mask important heterogeneity in sovereign crises and the model's response to them. Averages can obscure different patterns for different types of events. To investigate the specific contexts in which the model might provide early signals, we now turn to a granular case study analysis. We focus on the case of Chile, a country that experienced multiple downgrade events, including several within our out-of-sample test period, offering a rich timeline for analysis.

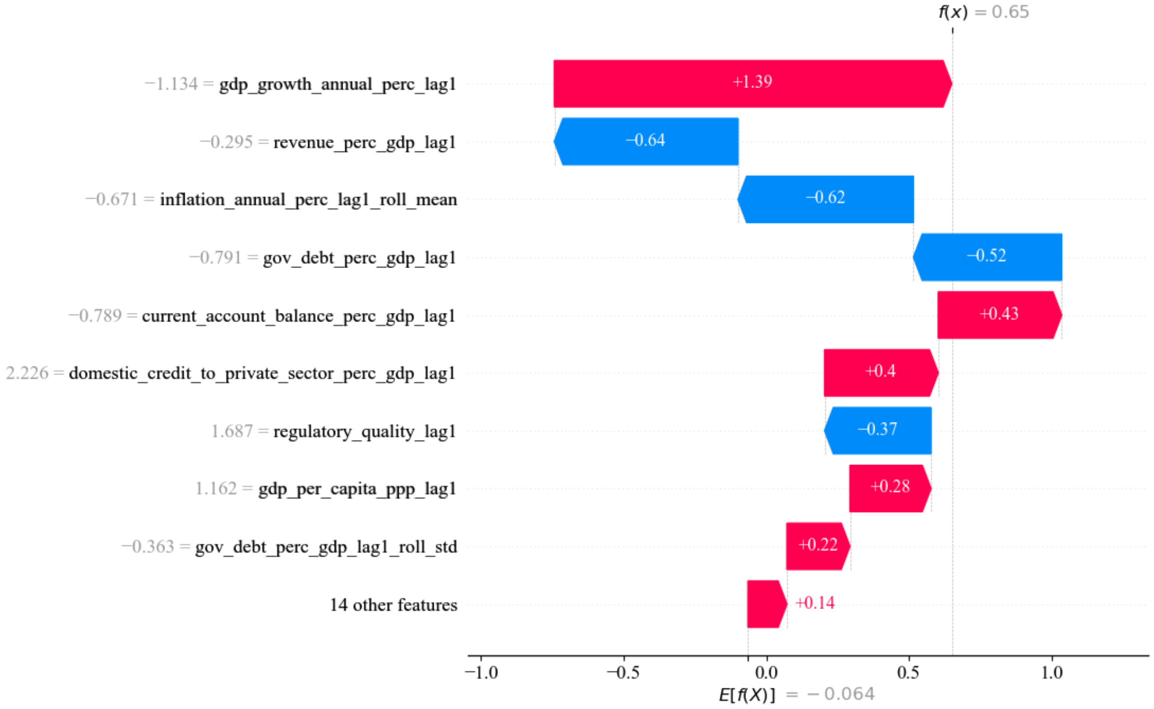
Figure 6 plots the model's predicted downgrade probability for Chile over the test period, with actual downgrade events marked by vertical lines. The plot for the 2020 downgrade is consistent with our aggregate findings: the risk score is low in the preceding years and spikes only in the year of the event. However, the plot reveals a different and compelling pattern for the subsequent 2023 downgrade. Following the 2020 event, the model's predicted risk did not recede to its previous low baseline. Instead, it remained at a highly elevated level throughout 2021 (a predicted probability of 0.72) and 2022 (0.90). This sustained period of diagnosed high risk served as a clear and persistent early warning for the next downgrade that eventually occurred in 2023.

Figure 6: Model-Predicted Downgrade Risk for CHL



This finding suggests that the model's timeliness is conditional on the nature of the crisis. To understand why the model's risk assessment remained so high, Figure 7 presents the SHAP waterfall plot for Chile's 2020 prediction, which initiated this period of sustained alert. The plot reveals that the initial risk prediction was driven primarily by a sharp deterioration in lagged GDP growth, which contributed a large positive SHAP value, pushing the prediction towards a downgrade. The model's subsequent high-risk predictions in 2021 and 2022 reflect the fact that this weak growth and the associated political uncertainty did not quickly resolve in the subsequent years' data. The model correctly interpreted this lack of recovery as a sign of entrenched vulnerability.

Figure 7: SHAP Explanation for Chile's 2020 Downgrade Risk Prediction



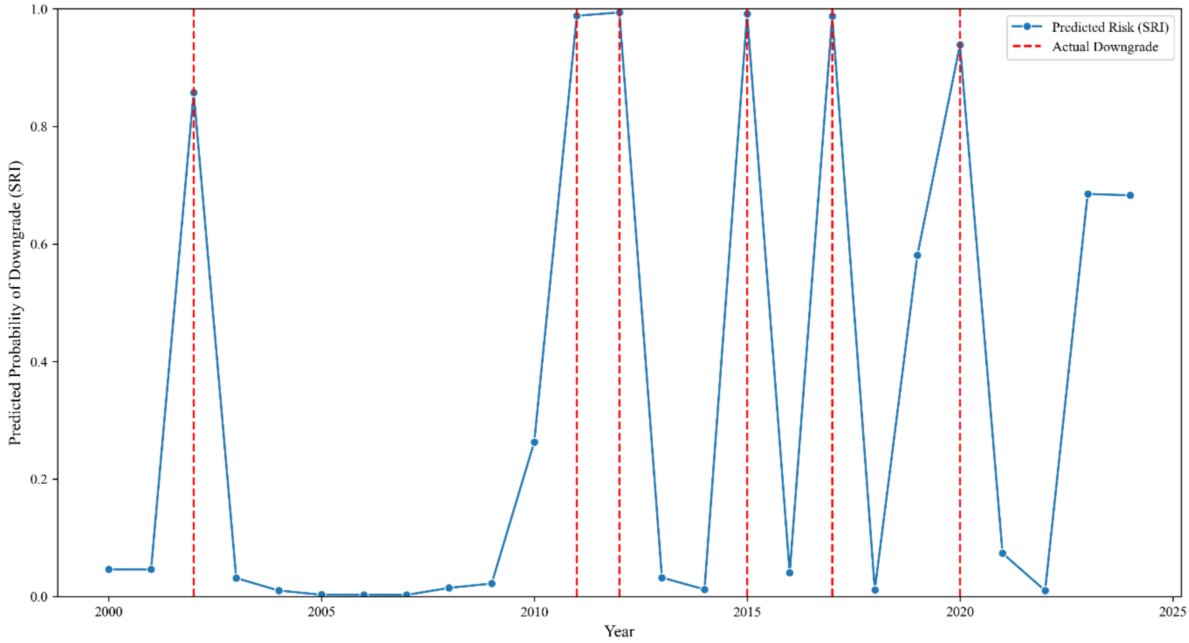
This case study, therefore, allows us to refine our conclusion. The model appears particularly effective at identifying "slow-burn" crises, where a period of sustained, unresolved economic weakness precedes the final rating action. In these instances, the model's ability to track dynamic features over time allows it to recognize the lack of recovery and maintain a high-risk signal, providing a valuable early warning that is not apparent in the aggregate, cross-country average.

iii. A Comparative Case Study: Disentangling "Slow-Burn" vs. "Sudden-Stop" Crises

The aggregate event study presented in Section 4.3.1 provided a powerful, generalizable conclusion: on average, our model functions as a highly accurate coincident indicator of sovereign downgrades. The case of Chile's 2023 downgrade, however, revealed that this aggregate result masks important heterogeneity; in specific contexts, the model is indeed capable of providing clear and sustained early warnings. This section seeks to move beyond this initial observation by conducting a deep, comparative diagnostic analysis. By contrasting the nature of the model's signals in two distinct country cases — the "slow-burn" crisis in Chile and a "sudden-stop" style crisis in South Africa — we aim to develop a richer theoretical understanding of what our model is truly detecting. We hypothesize that the *type* of signal produced by the model (coincident versus leading) is not a limitation, but rather a diagnostic feature, revealing crucial information about the underlying structure of a country's economy and its specific mode of vulnerability.

The temporal risk plot for South Africa, presented in Figure 8, tells a story that is highly consistent with our aggregate findings and serves as a quintessential example of the model functioning as a coincident indicator. For the downgrade event in 2020, as well as for prior events in 2017 and 2015, the model's predicted risk probability remains at a low baseline in the preceding years. A sharp, unambiguous spike in the risk score occurs only in the precise year that a downgrade was formally announced by the rating agencies. There is no significant, sustained run-up in the predicted risk in the t-1 or t-2 periods.

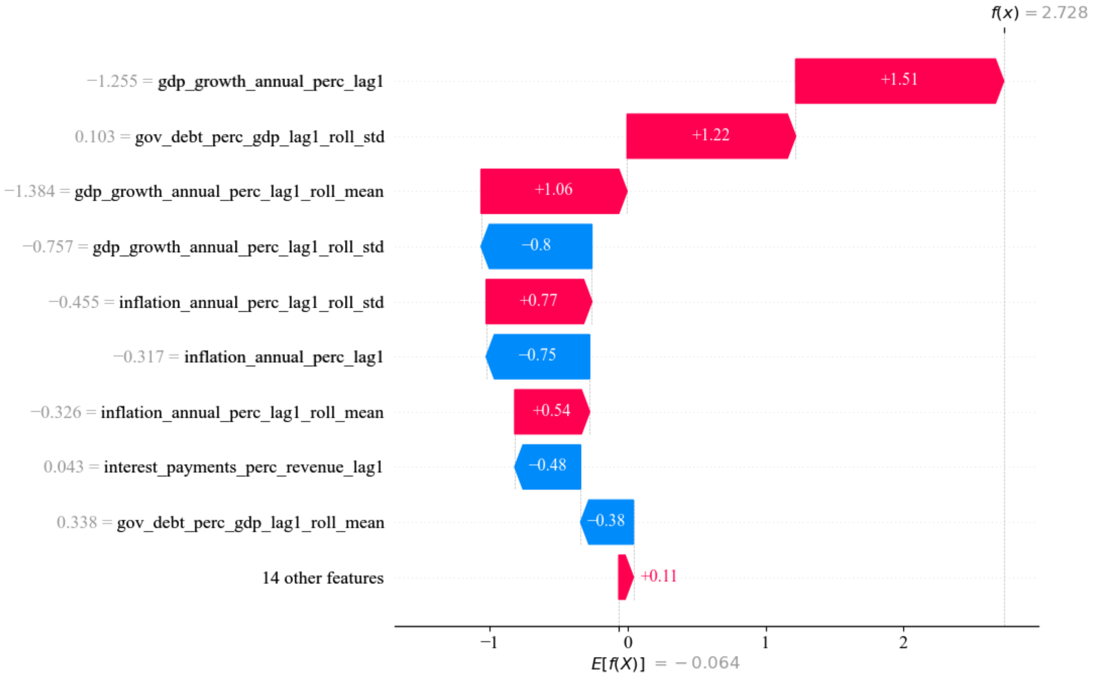
Figure 8: Model-Predicted Downgrade Risk for South Africa



To understand the mechanics of this coincident signal, the SHAP waterfall plot for the 2020 prediction (Figure 9) provides a granular explanation. The model's baseline expectation, E[f(x)] = -0.64, indicates a low average probability of a downgrade. However, for this specific prediction, the final output score is a very high f(x) = 2.728. The SHAP analysis reveals this is overwhelmingly driven by the model's reaction to a severe, sudden shock registered in the prior year's data. The gdp_growth_annual_perc_lag1 variable, reflecting the sharp economic downturn that began in 2019, contributes a massive +1.51 to the log-odds score. This primary signal is then amplified by other features capturing the weak economic environment, such as the low three-year rolling average of GDP growth.

This case allows us to theorize that a coincident signal is characteristic of an economy with significant, pre-existing structural fragilities. For years, the South African economy has been characterized by chronically low potential GDP growth, extremely high structural unemployment, and significant governance challenges. An economy with these features can be conceptualized as a “fragile house in a storm”. It may remain standing during periods of relative calm, and thus our model, looking at stable (albeit weak) t-1 data, does not flag a high risk of imminent collapse. However, when a major external shock arrives—in this case, the global disruption of the COVID-19 pandemic—it quickly overwhelms the country’s limited coping capacity and institutional buffers. The annual data used in our model can only fully register the profound economic damage from this shock *after* it has occurred. The model therefore acts as a rapid and accurate diagnostic tool for an ongoing crisis, effectively declaring, *“Based on the severe negative data from last year, this country is now in a state of acute distress”*. The signal is accurate but, by necessity of the data’s frequency, coincident.

Figure 9: SHAP Explanation for South Africa’s 2020 Downgrade Risk Prediction



In stark contrast, the temporal risk plot for Chile’s 2023 downgrade (Figure 7) tells a story of a “slow-burn” crisis. Following the initial shock and coincident signal in 2020, the model’s predicted risk did not recede. Instead, it remained at a highly elevated level of over 70% in both 2021 and 2022, serving as a clear, multi-year early warning for the subsequent downgrade in 2023. This raises the critical question: what did the model see in Chile that it did not see in South Africa?

The answer, we hypothesize, lies in the model's implicit understanding of Chile's different structural profile. Chile has historically been considered an exemplar of institutional strength and orthodox economic policy in the region, possessing what could be termed a "strong house in a hurricane". The initial shock of 2019-2020 damaged this structure, but did not cause it to collapse immediately, as evidenced by the still-strong institutional scores in the SHAP analysis for that year. The crucial difference is what happened next. The model's dynamic features, particularly the 3-year rolling averages, were designed to detect not just shocks, but the *recovery from shocks*. The persistently high-risk scores in 2021 and 2022 were the model's diagnosis of a failure to recover. It observed that for a country with historically strong institutions and growth potential, the prolonged period of sluggish growth and heightened political uncertainty following the initial shock was a profound and dangerous anomaly. The model's logic was not simply "GDP growth is low," but rather, "For an economy with Chile's structural characteristics, the *persistence* of this low growth indicates a fundamental erosion of its traditional sources of resilience." This sustained dissonance between a strong historical reputation and poor current performance is what constituted the "slow burn" and triggered the multi-year early warning.

This comparative analysis allows us to posit a more sophisticated interpretation of our model's output. The nature of the signal produced by our explainable framework is not a flaw, but a powerful diagnostic feature, revealing information about the type of crisis a country is facing.

- A coincident signal, like that for South Africa, may be indicative of a country with underlying structural fragilities that is highly vulnerable to large, exogenous shocks. The model acts as a rapid damage assessment tool.
- A sustained, "slow-burn" early warning signal, like that for Chile, may be indicative of a historically more resilient country whose core economic model or institutional framework is undergoing a fundamental, multi-year erosion. The model acts as a tool for diagnosing a chronic condition.

This insight—that the timeliness of a signal is contingent on the nature of the crisis—would be impossible to uncover without the transparent, feature-by-feature explanation provided by the SHAP framework. It demonstrates that the true value of this approach lies not in a single, universal predictive capability, but in its ability to provide a rich, multi-layered narrative about the specific nature of a country's sovereign risk. This also provides the ultimate justification for the primary avenue of future research: the integration of high-frequency market data, which would be better suited to providing earlier warnings for the “sudden-stop” style of crises that annual data struggles to anticipate.

d. Discussion

The empirical results presented in this chapter, taken in totality, offer a multi-faceted and sophisticated answer to our central research question. We have demonstrated that a non-linear machine learning model, specifically an XGBoost classifier, can predict sovereign downgrade events with a significantly higher degree of balanced accuracy than a traditional econometric baseline. Furthermore, through the application of a state-of-the-art explainability framework, we have shown that this model’s predictive power is not derived from spurious or uninterpretable correlations, but from a set of economically sensible drivers that align with established theories of sovereign risk. Finally, our temporal analysis has revealed a crucial nuance in the model’s performance: while it functions as a powerful and accurate coincident indicator on average, its ability to provide true early warnings is context-dependent, emerging most clearly in cases of slow, persistent economic deterioration.

This set of findings requires us to address the question — “Do explainable machine learning models provide earlier signals of sovereign credit downgrades?” — with a conclusion that is more refined than a simple affirmative or negative. The evidence suggests that the quest for a universal, one-year-ahead early warning system based on annual macroeconomic data may be misdirected. For “sudden-stop” crises, often triggered by sharp external shocks or unforeseen political events, the lagged nature of annual data inherently limits the predictive horizon. The economic damage from such shocks often only becomes fully visible in the official statistics with a considerable delay, making a coincident signal the most that can be reliably achieved. Our aggregate event study, which showed the model’s risk signal peaking in the year of the downgrade itself, provides strong evidence for this conclusion.

However, this does not diminish the value of the framework we have developed. In fact, it sharpens our understanding of its true, practical contribution, which lies less in its universal predictive timing and more in its capacity to provide a transparent, granular, and real-time risk audit. This represents a paradigm shift away from the traditional model of credit risk assessment.

The primary limitation of the ratings issued by major agencies is their inherent opacity, particularly the “qualitative overlay” applied by their rating committees. Our approach directly resolves this “black box” problem. By pairing the XGBoost model with the SHAP framework, we have created a “glass box”. For any country, at any point in time, we can produce a detailed, quantitative attribution of its predicted risk, pinpointing the specific factors that are driving the assessment. A policymaker or investor is no longer faced with just a rating, but with a clear, evidence-based explanation. They can see, for instance, that a country’s elevated risk score is being driven primarily by a weakening current account balance and high inflation, while being partially mitigated by its strong institutional quality. This transparency is a profound advantage, fostering trust, enabling accountability, and facilitating a more constructive policy dialogue.

Furthermore, our model produces a continuous probability score—our Sovereign Risk Index—which offers a much more granular view of evolving risk than the static, discrete letter grades issued by agencies. A country’s underlying risk profile can deteriorate significantly for many months or even years while its official rating remains unchanged, giving a false sense of security. Our SRI, in contrast, would reflect these subtle shifts in real-time, providing a more dynamic and responsive measure of vulnerability. This granularity allows for a more proactive approach to risk management.

This reframes the very concept of an “early warning”. Perhaps the most valuable warning is not that a downgrade will happen in one year, but an immediate and transparent warning about the specific vulnerabilities that exist right now. A policymaker armed with this knowledge can take targeted action to address those weaknesses, potentially averting the crisis altogether. Ultimately, while the search for a perfect predictive crystal ball continues, this research demonstrates that the combination of machine learning and explainability provides a powerful, immediately applicable tool for understanding and managing sovereign risk in the present. The final chapter will conclude by summarizing these findings and discussing their broader policy implications.

V. Conclusion

a. Summary of Findings

This thesis embarked on a systematic inquiry to assess whether a modern, explainable machine learning framework could overcome the well-documented limitations of traditional sovereign credit ratings—namely, their opacity, discreteness, and often lagging nature. The central objective was to develop and rigorously test a data-driven Sovereign Risk Index (SRI) to determine if it could provide a more transparent, granular, and, most critically, timely assessment of sovereign risk. After a comprehensive process of data construction, feature engineering, model development, and multi-faceted evaluation, the empirical results of this research yield a series of clear, albeit nuanced, conclusions.

First, this research established the superior predictive power of non-linear machine learning models for the task of sovereign downgrade prediction. A state-of-the-art XGBoost classifier was developed and benchmarked against a traditional Logistic Regression model, a common tool in the econometric literature. The out-of-sample evaluation on the post-2019 test set demonstrated that the XGBoost model achieved a significantly higher F1-Score (0.46 vs. 0.35) and overall accuracy (77% vs. 53%). This performance confirms that the complex, non-linear interactions between economic and institutional variables, missed by linear models cannot capture, are critical for classifying sovereign risk. The XGBoost model's higher precision further indicates that its downgrade signals are substantially more reliable, reducing the rate of false alarms, which is a crucial feature for any practical risk management tool.

Second, having established the model's predictive efficacy, we addressed its "black box" nature through the application of a state-of-the-art Explainable AI (XAI) framework, SHAP. The analysis of global feature importance provided compelling evidence that the model's predictions are not based on spurious correlations but are driven by economically sound and theoretically grounded principles. The model correctly identified prior-year GDP growth, measures of institutional quality such as regulatory quality and rule of law, and the dynamic volatility of government debt as the most significant drivers of its risk assessments. This step was critical in validating the model as not just a statistically powerful tool, but as an analytically sensible one, thereby building the trust necessary for its use in policy and investment contexts.

Finally, the central research question regarding the timeliness of the model's signals was addressed through a detailed temporal analysis. The aggregate event study, which averaged the model's predicted risk probability across all downgrade events in the test set, revealed that the model functions, on average, as a highly accurate coincident indicator. The risk signal was found to increase most sharply and peak in the year of the downgrade itself (t=0), rather than showing a consistent, rising trend in the years prior. However, this aggregate finding was enriched by a deeper case study analysis. The analysis of Chile's recent downgrade history showed that the model provides specific early warnings in particular situations. The model detected a prolonged high-risk period for two consecutive years before the 2023 downgrade which proved correct in identifying a gradual crisis through worsening fundamentals.

This research demonstrates that the developed machine learning framework functions as a strong transparent and evidence-based tool for real-time sovereign risk diagnosis, but it does not serve as a solution for predicting all crisis types in advance. The model surpasses conventional models by delivering sophisticated data-based insights about the intricate process of sovereign risk evaluation.

b. Contribution to the Literature

This thesis contributes to the academic literature on sovereign risk assessment, predictive modelling, and applied financial machine learning across three primary dimensions: methodological, empirical, and conceptual. By bridging the gap between high-performance prediction and model interpretability, this research offers a novel framework and a set of nuanced findings that extend the current body of knowledge.

The primary methodological contribution of this work is its demonstration of a systematic and integrated framework that pairs a high-performance, non-linear machine learning algorithm (XGBoost) with a state-of-the-art Explainable AI (XAI) technique (SHAP). Much of the existing literature on financial forecasting has implicitly or explicitly accepted a trade-off between predictive accuracy and model interpretability. Researchers have often had to choose between simple, transparent models (like logistic regressions) that may fail to capture the complexity of the data, and more powerful "black box" models whose decision-making processes remain opaque. This thesis directly challenges that dichotomy. By providing a complete, end-to-end workflow — from rigorous feature engineering to model training, to post-hoc explanation via SHAP — this research provides a practical and robust blueprint for future studies. It demonstrates how to build what can be termed a "glass box" model: one that leverages the full predictive power of machine learning without sacrificing the crucial requirements of transparency, trust, and auditability demanded by financial regulators, policymakers, and market participants.

The main empirical contribution is the nuanced and data-driven answer provided to the long-standing "early warning" question in the context of machine learning models. While many studies have focused on optimizing out-of-sample accuracy metrics, this thesis provides a more rigorous assessment of the timeliness of the model's predictive signals. The application of a formal event study methodology to the model's out-of-sample predictions allowed for a clear, aggregate distinction to be drawn between coincident and leading indicators. Our central finding—that the model functions, on average, as a powerful coincident indicator but that its ability to provide true early warnings is context-dependent on the nature of the crisis—adds a new layer of sophistication to the EWS debate. The identification of “slow-burn” deteriorations as a scenario where early warnings are more likely provides a testable hypothesis for future research and suggests that the effectiveness of any EWS may be contingent on the type of shock it is designed to predict.

Finally, this research makes a conceptual contribution by proposing and developing a prototype for a Sovereign Risk Index (SRI) as a new paradigm for risk assessment. This thesis argues that the output of our framework should be viewed not merely as a prediction, but as a new class of risk metric that is superior to traditional credit ratings across several dimensions. Unlike the discrete, opaque, and often lagging letter grades issued by agencies, the SRI is, by design, transparent, with its drivers quantifiable via SHAP; granular, offering a continuous probability score that captures subtle shifts in risk; and dynamic, capable of being updated in near real-time as new data becomes available. This thesis, therefore, does not just present a model; it presents a proof-of-concept for a new class of analytical tools designed for a more dynamic and data-rich era of sovereign risk management.

c. Policy and Practical Implications

The academic framework of this thesis provides practical applications for multiple actors in the global financial system. The main worth of this research extends beyond its predictive capabilities because it enables better sovereign risk management through transparent and proactive methods. The SRI functions as a powerful decision-making instrument as it provides transparent near real-time risk assessments instead of using opaque lagging ratings for governments and private investors and international financial institutions.

For governments and policymakers, particularly within emerging market economies, the SRI framework offers a paradigm shift from a reactive to a proactive policy stance. Traditionally, a sovereign government’s interaction with credit risk is often punctuated by discrete, high-stakes reviews from rating agencies. The SRI, in contrast, provides a continuous, dynamic “risk audit” capability. A ministry of finance or central bank could use this framework to monitor its country’s risk profile on a regular basis, perhaps quarterly. More importantly, the integrated explainability of the model allows for an immediate decomposition of any change in the risk score. For instance, if the SRI indicates a rising risk, policymakers could use the SHAP analysis to determine if the driver is a deterioration in the fiscal balance, declining investor sentiment reflected in market volatility, or weakening external accounts. This allows for a highly targeted and evidence-based policy response. It enables a more nuanced policy dialogue, both internally and with external stakeholders like investors and the rating agencies themselves, as the government can demonstrate a clear, data-driven understanding of its own specific vulnerabilities and the actions being taken to mitigate them.

The SRI serves as an advanced due diligence tool and risk management system for private sector investors alongside asset managers. The discrete letter-grade ratings of CRAs place different countries with varied risk patterns into identical broad risk categories (e.g., ‘BB’). Investors who use the SRI with continuous probability scoring can identify risk cases with higher precision than the broad groupings of traditional ratings. A ‘BB’ rated country evaluation reveals different fundamental risk levels when comparing Country A which maintains a stable SRI of 20% while Country B shows a SRI increase from 15% to 40% during the last two quarters. The SRI generates precise risk profile indicators which traditional ratings systems cannot detect. The SRI functions as a numerical input for sovereign debt allocation models and serves as a warning indicator to initiate fundamental analysis and enables better management of portfolio-level risks through agile adjustments.

International financial institutions such as the International Monetary Fund (IMF) and the World Bank can enhance their essential country surveillance and crisis prevention work through the SRI framework. The monitoring of economic health across all member countries represents a resource-intensive responsibility for these institutions. A transparent SRI system that runs across many countries functions as an efficient initial evaluation system. These institutions can use the model to detect continuous deterioration in fundamentals across countries and direct their attention to these areas for additional scrutiny and assistance or policy engagement. The explainability feature provides immediate understanding of risk sources so analysts can direct their engagement efforts with better precision. A data-driven method enables institutions to direct their scarce resources toward the most critical areas thus promoting worldwide financial stability.

d. Limitations and Avenues for Future Research

While this thesis presents a robust and transparent framework for a machine-learning-driven Sovereign Risk Index, it is important to acknowledge the limitations inherent in its scope and methodology. These limitations, however, do not diminish the study’s contributions; rather, they illuminate a clear and promising roadmap for future research at the dynamic intersection of sovereign risk, machine learning, and explainable AI. The challenges encountered in this work provide the very questions that can motivate the next generation of inquiry in this field.

The most significant limitation of the current study, which directly impacts its timeliness, is its reliance on annual data for its core macroeconomic predictors. This low frequency inherently constrains the model's ability to provide timely, intra-year warnings. Economic conditions and market sentiment can shift dramatically within a few months, yet a model based on annual data can only observe these changes with a considerable lag. The primarily coincident, rather than leading, nature of our model's signals in the aggregate event study is likely a direct consequence of this data structure. Therefore, the most critical and immediate avenue for future research is the reconstruction of this entire framework using quarterly or, ideally, monthly data. The integration of high-frequency, market-based indicators—such as sovereign CDS spreads, government bond yield spreads, and measures of FX volatility—would provide the model with the real-time information necessary to detect the build-up of risk much earlier. A high-frequency version of the SRI would represent a major step towards a true Early Warning System.

Second, our dependent variable, while robustly defined, is a simple binary flag (downgrade vs. no downgrade). This simplification, while necessary for a clear classification approach, discards valuable information about the magnitude and type of rating change. A single-notch downgrade from 'A-' to 'BBB+' is fundamentally different from a multi-notch collapse from 'B' to 'CCC'. Future research could explore more sophisticated target variables. For instance, a multi-class classification model could be trained to predict the specific rating bucket a country will transition to (e.g., 'no change', 'single-notch downgrade', 'multi-notch downgrade'). Alternatively, a regression framework could be used to predict the exact number of notches a rating will change. Furthermore, explicitly redefining the target to a two-year forward prediction window, as suggested by external feedback, remains a key experiment to more directly test the model's early warning potential.

Third, this thesis developed a single, general model trained on a diverse panel of 23 emerging markets. While this approach ensures the generalizability of our findings, it may average over important regional or structural differences between countries. The economic drivers of risk for a major commodity exporter in Latin America may differ significantly from those for a manufacturing hub in Southeast Asia. An interesting path for future research would be to explore the development of more specialized models. One could apply clustering algorithms to group countries into archetypes based on their economic structure (e.g., commodity exporters, low-income countries, highly dollarized economies) and then train a tailored model for each cluster. This could potentially yield more precise, customized, and ultimately more useful risk assessments.

Finally, while XGBoost represents a state-of-the-art algorithm for structured, tabular data, it is not inherently designed to capture very long-range temporal dependencies. As the dataset is migrated to a higher frequency in future work, more advanced model architectures could be explored. Deep learning models specifically designed for time-series data, such as Long Short-Term Memory (LSTM) networks or even Transformers, may be able to capture more complex patterns and temporal dependencies that are not accessible to tree-based models. Integrating these architectures with the SHAP framework would represent the next frontier in building a truly dynamic and explainable Sovereign Risk Index. This thesis provides a solid foundation upon which this future work can be built.

Bibliography

Akter, F., Min Su and Amankwah, O. (2021) ‘Determinants &amp; Impact of Sovereign Credit Ratings’. Available at: https://doi.org/10.5281/ZENODO.4515670.

Alessi, L. and Detken, C. (2011) ‘Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity’, *European Journal of Political Economy*, 27(3), pp. 520–533. Available at: https://doi.org/10.1016/j.ejpoleco.2011.01.003.

Bar-Isaac, H. and Shapiro, J. (2011) ‘Credit Ratings Accuracy and Analyst Incentives’, *The American Economic Review*, 101(3), pp. 120–124.

Berg, A. and Pattillo, C. (1999) ‘Predicting currency crises:: The indicators approach and an alternative’, *Journal of International Money and Finance*, 18(4), pp. 561–586. Available at: https://doi.org/10.1016/S0261-5606(99)00024-8.

Bluwstein, K. *et al.* (2023) ‘Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach’, *Journal of International Economics*, 145, p. 103773. Available at: https://doi.org/10.1016/j.jinteco.2023.103773.

Butler, A.W. and Fauver, L. (2006) ‘Institutional Environment and Sovereign Credit Ratings’, *Financial Management*, 35(3), pp. 53–79.

Chen, T. and Guestrin, C. (2016) *XGBoost: A Scalable Tree Boosting System*, p. 794. Available at: https://doi.org/10.1145/2939672.2939785.

Eichengreen, B., Hausmann, R. and Panizza, U. (2005) ‘The Pain of Original Sin, The Mystery of Original Sin, and Original Sin: The Road to Redemption’, *Eichengreen & Hausmann 2005* [Preprint]. Available at: https://doi.org/10.7208/chicago/9780226194578.003.0010.

Ferri, G., Liu, L. -G. and Stiglitz, J.E. (1999) ‘The Procyclical Role of Rating Agencies: Evidence from the East Asian Crisis’, *Economic Notes*, 28(3), pp. 335–355. Available at: https://doi.org/10.1111/1468-0300.00016.

Frankel, J. and Rose, A. (1996) ‘Currency Crashes in Emerging Markets: An Empirical Treatment’, *Journal of International Economics*, 41, pp. 351–366. Available at: https://doi.org/10.1016/S0022-1996(96)01441-9.

Kaminsky, G., Lizondo, S. and Reinhart, C. (1998) ‘Leading Indicators of Currency Crises’, *IMF Staff Papers*, 45(1), pp. 1–48.

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2010) ‘The Worldwide Governance Indicators: Methodology and Analytical Issues’, *World Bank, Working Paper Series 5430*, 3. Available at: https://doi.org/10.1017/S1876404511200046.

Krugman, P. (1979) ‘A Model of Balance-of-Payments Crises’, *Journal of Money, Credit and Banking*, 11(3), pp. 311–325. Available at: https://doi.org/10.2307/1991793.

Krugman, P. (1999) ‘Balance Sheets, the Transfer Problem, and Financial Crises’, *International Tax and Public Finance*, 6(4), pp. 459–472. Available at: https://doi.org/10.1023/A:1008741113074.

Lundberg, S.M. and Lee, S.-I. (2017) ‘A Unified Approach to Interpreting Model Predictions’, in I. Guyon et al. (eds) *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

MacKinlay, A.C. (1997) ‘Event Studies in Economics and Finance’, *Journal of Economic Literature*, 35(1), pp. 13–39.

North, D.C. (1990) *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press (Political Economy of Institutions and Decisions). Available at: https://doi.org/10.1017/CBO9780511808678.

Obstfeld, M. (1996) ‘Models of currency crises with self-fulfilling features’, *Papers and Proceedings of the Tenth Annual Congress of the European Economic Association*, 40(3), pp. 1037–1047. Available at: https://doi.org/10.1016/0014-2921(95)00111-5.

Reinhart, C.M. and Rogoff, K.S. (2009) *This Time Is Different: Eight Centuries of Financial Folly*. Princeton University Press. Available at: https://doi.org/10.2307/j.ctvcm4gqx.

Annex

Annex I: Country Sample and Recorded Downgrade Events

Country Name	ISO Code	Sample	Downgrade Event Year(s)
Brazil	BRA	2001-2024	2002, 2015, 2016, 2018
Chile	CHL	2002-2024	2003, 2008, 2013, 2017, 2020, 2023
Colombia	COL	2002-2024	2003, 2017, 2021
Ecuador	ECU	2001-2024	2003, 2005, 2008, 2014, 2016, 2018, 2023
Egypt	EGY	2001-2024	2002, 2011, 2012, 2017, 2023
Ghana	GHA	2005-2024	2013, 2014, 2016, 2021, 2022
Hungary	HUN	2001-2024	2002, 2005, 2008, 2009, 2011, 2012, 2018, 2020
India	IND	2001-2024	2001, 2009, 2018
Indonesia	IDN	2001-2024	2018
Kenya	KEN	2007-2024	2008, 2021, 2023
Sri Lanka	LKA	2006-2024	2008, 2016, 2018, 2020, 2022
Mexico	MEX	2001-2024	2009, 2019, 2020, 2023
Malaysia	MYS	2001-2024	2020
Peru	PER	2001-2024	2006, 2021
Philippines	PHL	2001-2024	2004, 2005, 2006, 2015
Poland	POL	2001-2024	2003, 2016
Romania	ROU	2001-2024	2008, 2013
Russia	RUS	2001-2022	2008, 2014, 2015, 2017, 2022
Saudi Arabia	SAU	2003-2024	2005, 2015, 2017, 2019
Thailand	THA	2002-2024	2009, 2011
Tunisia	TUN	2001-2024	2011, 2012, 2013, 2017, 2020, 2021, 2022, 2023
Ukraine	UKR	2001-2024	2008, 2009, 2012, 2014, 2018, 2022, 2024
South Africa	ZAF	2001-2024	2002, 2011, 2012, 2015, 2017, 2020

Annex II: Logistic Regression Coefficient Estimates

Predictor Variable	Coefficient	Std. Error	P-value
Intercept	-1.6250	0.1541	0.000 **
current_account_balance_perc_gdp_lag1	-0.7623	0.4530	0.092 *
inflation_annual_perc_lag1	-0.6001	0.3770	0.111
domestic_credit_to_private_sector_perc_gdp_lag1	0.2498	0.1808	0.167
lending/borrowing_perc_gdp_lag1	0.0463	0.2053	0.822
revenue_perc_gdp_lag1	0.4967	0.2714	0.067 *
interest_payments_perc_revenue_lag1	0.5088	0.2897	0.079 *
gdp_growth_annual_perc_lag1	0.1623	0.2114	0.443
gdp_per_capita_ppp_lag1	0.3689	0.3046	0.226
gov_debt_perc_gdp_lag1	0.9015	0.9328	0.334
control_of_corruption_lag1	-0.0308	0.3377	0.927
gov_effectiveness_lag1	-0.0886	0.3710	0.811
political_stability_lag1	0.0704	0.2623	0.788
rule_of_law_lag1	0.1635	0.3933	0.678
regulatory_quality_lag1	-0.3552	0.3144	0.258
voice_and_accountability_lag1	-0.0648	0.2502	0.796
gov_debt_perc_gdp_lag1_roll_mean	-1.2136	0.9039	0.179
gov_debt_perc_gdp_lag1_roll_std	0.2599	0.1898	0.171
current_account_balance_perc_gdp_lag1_roll_mean	0.2291	0.4800	0.633
current_account_balance_perc_gdp_lag1_roll_std	-0.2275	0.2069	0.271
inflation_annual_perc_lag1_roll_mean	0.5180	0.4383	0.237
inflation_annual_perc_lag1_roll_std	-0.1785	0.3327	0.592
gdp_growth_annual_perc_lag1_roll_mean	-0.1543	0.2332	0.508
gdp_growth_annual_perc_lag1_roll_std	-0.1577	0.1710	0.356

* and ** denote statistical significance at the 10% and 5% levels, respectively.

Annex III: Top 20 Feature Importances from XGBoost Model

Rank	Feature Name	Mean |SHAP Value|
1	gdp_growth_annual_perc_lag1	0.8400
2	gov_debt_perc_gdp_lag1_roll_std	0.7265
3	current_account_balance_perc_gdp_lag1	0.4513
4	domestic_credit_to_private_sector_perc_gdp_lag1	0.4065
5	regulatory_quality_lag1	0.3895
6	gov_debt_perc_gdp_lag1	0.3411
7	inflation_annual_perc_lag1_roll_mean	0.3385
8	gdp_growth_annual_perc_lag1_roll_std	0.3318
9	interest_payments_perc_revenue_lag1	0.3252
10	current_account_balance_perc_gdp_lag1_roll_std	0.3121
11	inflation_annual_perc_lag1	0.3055
12	gdp_growth_annual_perc_lag1_roll_mean	0.2944
13	rule_of_law_lag1	0.2668
14	voice_and_accountability_lag1	0.2566
15	gdp_per_capita_ppp_lag1	0.2557
16	inflation_annual_perc_lag1_roll_std	0.2524
17	political_stability_lag1	0.2323
18	control_of_corruption_lag1	0.2298
19	lending/borrowing_perc_gdp_lag1	0.2201
20	gov_debt_perc_gdp_lag1_roll_mean	0.2173