MACHINE LEARNING FOR FINANCIAL CRISIS PREDICTION

# MACHINE LEARNING FOR FINANCIAL CRISIS PREDICTION

BY

JOSEPH VOSKAMP, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS AND STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2024) McMaster University

(Mathematics and Statistics) Hamilton, Ontario, Canada

TITLE: Machine Learning for Financial Crisis Prediction

AUTHOR: Joseph Voskamp

B.Sc., (Mathematics)

McMaster University, Hamilton, Canada

SUPERVISOR: Dr. Matheus R. Grasselli

NUMBER OF PAGES: xiv, 104

# Abstract

We investigate the potential applications of using machine-learning models in financial crisis prediction. We aim to identify crises one or two years ahead of their start dates by recognizing trends in a variety of economic variables. We look at two different datasets of banking crises, as well as currency and inflation crises. For consistency in analysis, we manually construct the crisis variables for the years 2017-2020. By analyzing the models in both cross-validation and forecasting experiments, we show that machine-learning models can outperform logistic regression in financial crisis prediction. We employ a Shapley value framework in an attempt to mitigate the *black box* nature of the machine-learning models. We show that the global economic climate is of vital importance in identifying banking and currency crises. Wages are shown to be the most important predictor of inflation crises. We then investigate the nonlinear relationships between the predictors and their Shapley values to further understand the driving forces behind the model predictions.

# Acknowledgements

I would like to thank my supervisor, Dr. Matheus Grasselli for his guidance and oversight of my work. His patience with me as I grappled with economic ideas was greatly appreciated.

I would also like to thank my wife and my family members for always supporting me and encouraging me as I pursue my goals.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Financial crises often shock financial systems, resulting in large economic and social costs. Spotting a financial crisis before it occurs is a worthwhile task for policymakers. The key to this is that the crisis must be spotted sufficiently early so policymakers have time to implement policy to either prevent the crisis or dampen its severity.

Pinpointing a set of reliable early warning predictors can be challenging. First, complex early warning models must be simplified into simple indicators for policy implementation. Second, achieving a robust model can be challenging since the number of observed crises over time is relatively low. Lastly, typical indicators for crises often send a signal when it is too late for policy to make a difference.

This work aims to provide solutions to these problems. Following the work in Bluwstein et al., 2021, we employ a wide variety of machine learning techniques and show that these models can achieve better results than a baseline logistic regression model. Machine learning models also offer greater flexibility, allowing them to learn the nonlinear relationships between the predictors and the predictions. Identifying these relationships can be useful in policy making.

Machine learning models are often criticized for being *black boxes*. This critique

is about how machine learning models can achieve good results, but it is difficult to determine *how* the model is making its predictions. A recent development in addressing this problem is called the **SH**apley **A**dditive ex**P**lanations, or SHAP method(Awan, 2023),(Lundberg & Lee, 2017),(Joseph, 2019). The SHAP method has desirable properties for analysis (Bluwstein et al., 2021). This framework allows the identification of the economic drivers of the model predictions. Policymakers often face a challenge in trying to use machine learning models to help inform their decisions, but the SHAP method gives insight into how the predictions of the models are being made. The benefits of using the SHAP method are important since policymakers need to be transparent and give justification when they implement policy. Having a way to interpret the models is desirable for everyone involved.

The goal of this thesis is to predict different types of financial crises 1 or 2 years in advance of the crisis year. This is the typical horizon given in the literature for early warning systems as it gives policymakers time to potentially prevent the crisis or mute its effects (Bluwstein et al., 2021). We use the Macrohistory Database from Jordà et al., 2017, henceforth denoted as JST, as our primary source of information for financial crises. This database contains a large number of macroeconomic variables for 18 countries from 1870-2020. It also contains a binary variable which indicates whether a country experienced a banking crisis that year or not. The definition of a systemic banking crisis used in this database is as follows: major bank failures, banking panics, substantial losses in the banking sector, significant recapitalization, and/or significant government intervention (Jordà et al., 2017). We test the performance of out-of-sample prediction for the following machine learning models: decision tree, random forest, extremely randomized trees, support vector machines, artificial neural networks, extreme gradient boosting, and K-nearest neighbours. The main finding is that the tree-based machine learning models often perform better than the benchmark logistic

regression model. We then apply these same methods to different types of financial crises, including a different type of banking crisis, currency crises, and inflation crises. The definitions of these 3 types of financial crises are obtained from Reinhart and Rogoff, 2009, who have a narrative-based definition for banking crises that differs slightly from that of the JST database, and threshold-based definitions for the currency and inflation crises.

When looking at the two definitions of banking crises, we find that similar results hold. Consistent with Bluwstein et al., 2021, we find that the major predictors for this type of crisis are the credit growth and the slope of the yield curve. The literature has identified credit growth as an important predictor of financial crises, but the importance of the yield curve is less present (Schularick & Taylor, 2012), (Jorda et al., 2013). Our results show that a flatter or inverted yield curve is indicative of a higher probability of a crisis. These predictors are found to be important both domestically and globally. However, it was found that global credit growth was the most important during the global financial crisis, whereas the global slope of the yield curve was important over the entire sample. The CPI, wages, and debt servicing ratio rank as the next set of important features across both types of banking crises.

For currency crises, the global variables continue to rank as the top two predictors, followed by the CPI, the public debt, and the debt servicing ratio. The weight given to the global variables is quite significant, implying that currency crises could be heavily influenced by global economic trends.

When predicting inflation crises, wages are the most important predictor. This is followed by consumption, the global slope of the yield curve, and the debt servicing ratio. The significant weight given to the wages variable implies that a significant rise in nominal wages is indicative of a pre-inflation crisis period.

We employ a method of uncovering nonlinear relationships from Bluwstein et al.,

2021 for the key predictors in the models based on the SHAP framework. An example of a nonlinear relationship is that the predicted probability of a crisis increases dramatically when the global credit growth exceeds 10%, but the global credit growth has little to no effect on the predicted probability when credit growth is between 0% and 5%. Stock and Watson, 1999 concludes in their work that there are nonlinear relationships between indicators and inflation crises. The authors also show the importance of the unemployment rate as a predictor of inflation (Stock & Watson, 1999).

The remainder of the thesis is structured as follows. Section 1.1 presents the problem and dives into the existing literature on early warning systems and financial crisis prediction, including the recent work done with machine learning, and Section 1.2 describes the main datasets used in this work. Chapter 2 outlines the procedure and features used for the modelling in this work. This chapter concludes with a brief description of all the machine learning methods used in this work, followed by the Shapley value framework. Chapter 3 reports on the results of the experiment for the 2 types of banking crises. Chapter 4 reports on the results of the experiment for currency and inflation crises. Chapter 5 concludes and summarizes.

## 1.1 Literature Review

Our research stems from two main lines of work. The first inspiration comes from Carmen Reinhart and Kenneth Rogoff, with their 2009 book *This Time is Different: Eight Centuries of Financial Folly* (Reinhart & Rogoff, 2009). They look at eight centuries of many different types of financial crises for a wide range of countries. Throughout their work, they push back against the claim that financial systems learn from their past mistakes or the "this time is different" syndrome. Over the eight centuries covered in their work, they show that this time is not different. Financial catastrophes are part

of emerging and established markets (Reinhart & Rogoff, 2009). A part of their work involved creating databases and narrative accounts for different types of financial crises. In particular, we are interested in their work on banking, currency, and inflation crises.

The second, and greater, inspiration for this work comes from a paper written by the Bank of England called *Credit growth, the yield curve, and financial crisis prediction: evidence from a machine learning approach* (Bluwstein et al., 2021). In this work, the authors look at predicting banking crises using a wide variety of machine learning approaches and show that they can outperform a logistic regression in out-of-sample prediction. They employ methods for interpreting the machine learning models, tackling the *black box* problem often associated with these types of models (Bluwstein et al., 2021). We now move on to discussing what work has been done previously on early warning systems and predicting financial crises.

Early warning systems were a popular research topic following the global financial crisis of 2008. Indeed, the literature on this topic is extensive, with many different parties developing different systems to try and effectively capture trends preceding financial crises. Some of the early work on more modern early warning systems was done by the European Central Bank in the early 2000s following the financial crises of the 1990s (Bussiere & Fratzscher, 2002). In this work, the authors review the 2 main ways of constructing early warning systems at the time: extracting signals from a range of indicators and using logit models.

The signals approach was first developed by Kaminsky and Reinhart, 1997. Under this approach, each indicator is treated separately, and a threshold is chosen for each. If the indicator crosses its specific threshold, it issues a warning signal. The threshold chosen will determine the number of false alarms. The Kaminsky-Reinhart approach is to choose the threshold levels based on a grid search that minimizes the noise-to-signal ratio.

**Definition 1** (Noise-to-Signal Ratio)**.**

$$NSR = \frac{FalsePositives/(FalsePositives + TrueNegatives)}{TruePositives/(TruePositives + FalseNegatives)}$$

The shortfalls of this method are discussed in Bussiere and Fratzscher, 2002. They point out that there is a significant loss of information when using the signals approach since, for example, the current account dropping 1, 5, and 10 percentage points would provide the same signal. They also note that this approach does not give a holistic picture in terms of the vulnerability of a given country.

More recent work on the signals approach has attempted to deal with the second criticism of this framework by implementing crisis indices or aggregated indicators (Bhamani et al., 2018). The authors point out that it often does not make sense to try and predict the likelihood of a crisis based on a single indicator. Thus, they combine all of the indicators into a crisis index for each type of crisis. This approach also provides a type of answer to the first problem listed above, as a larger value for the index indicates a larger probability of crisis.

The logit models for financial crisis prediction were first presented in 1989 (Bussiere & Fratzscher, 2002) and involve modelling the probability of a financial crisis as a nonlinear function of the predictors using a logistic distribution. As described in Saab et al., 2024, the model looks at the following equation:

$$P(crisis = 1) = F_w(x).$$

In this equation, $P(crisis = 1)$ represents the probability of a crisis occurring for observation $x$, $F_w$ is the logistic function which combines the independent features into a probability using the weights (or coefficients) $w$, and $x$ is the vector of predictors being used for that observation. The nonlinear property given by this model is desir-

able because other literature on financial crises has identified that there are nonlinear relationships at work (Reinhart & Rogoff, 2009). However, the ability to capture these nonlinear relationships is restricted to the logistic function in this model. While there is potential in more effectively capturing these relationships using logistic regression, the process of achieving that success is more trial and error, for example by manually adding polynomial functions of existing predictors. Bussiere and Fratzscher, 2002 offers a potential improvement for the logit model using pooled data, and they show that their logit model outperforms a Kaminsky-Reinhart signals approach model. Saab et al., 2024 claims that using a logit model is best because it is easy to understand, is already the most common model in crisis prediction, and is more accurate than any other model. In the literature, logit models have been used to estimate the global financial crisis (Saab et al., 2024). Other works employing a logit model have found a globalization aspect to financial crises (see references in Saab et al., 2024). Borio and Drehman, 2008, Davis and Karim, 2008, and Frankel and Saravelos, 2012 all used logit models in developing early warning systems for detecting financial crises.

## 1.2 Financial Crisis Databases

A handful of databases investigate banking crises worldwide. In this section, we will look at each of the major databases and how they define a banking crisis. After an introduction to each database, we will discuss different works that use each database for financial crisis prediction.

We begin by looking at the database constructed by Laeven and Valencia, 2018. In their work, a systemic banking crisis is defined by the following two events:

1. "Significant signs of financial distress in the banking system (as indicated by significant bank runs, losses in the banking system, and/or bank liquidations).

2. Significant banking policy intervention measures in response to significant losses in the banking system."

A positive label is placed for the $\langle country, year \rangle$ observation when the country first meets both of these criteria. They do this to ensure that the beginning of the crisis indicates the beginning of major problems. However, they mention that if the losses are significant in the first criteria, it is sufficient to classify it as a banking crisis. The losses are considered severe if: "a country's banking system exhibits significant losses resulting in a share of nonperforming loans above 20 percent of total loans or bank closures of at least 20 percent of banking system assets, or fiscal restructuring costs of the banking sector are sufficiently high, exceeding 5% of GDP." One of the major goals of their work is to reduce the amount of subjectivity that is present in other financial crisis databases for classifying a crisis (Laeven & Valencia, 2018).

The next database we will introduce was created by Baron et al., 2020. Baron et al., 2020 points out that an unfortunate thing about previous crisis databases is that they often disagree about when a crisis happens or if it happens at all. This disagreement is due to inconsistent definitions of what constitutes a banking crisis. However, there is no single correct definition for a banking crisis! Thus, their work aims to provide "one possible construction of clear-cut crisis episodes based on systematic criteria" (Baron et al., 2020). Their database is constructed as the union of two separate crisis databases. The first list of crises they construct is a list of "bank equity crises." For this list, a crisis occurs when there is a cumulative 30% decline in bank equity. From this set of potential crises, they only select instances with narrative evidence of widespread bank failures (retrieved from other databases). They append a chronology of "panic banking crises" to this list. A panic banking crisis is defined as an episode when banks experience "sudden salient funding pressures" (Baron et al., 2020). They combine these two lists to create the BVX crisis database. Their work notes that they would rather be overly

inclusive in classifying banking crises.

We now move towards databases that are employed in this work. The first dataset used in this work that we will discuss is the dataset used by Reinhart and Rogoff in their work *This Time is Different: 800 Years of Financial Folly* (Reinhart & Rogoff, 2009). In their work, they build crisis lists for several different types of financial crises. The types of crises that are of interest to us are banking crises, currency crises, and inflation crises. For Reinhart and Rogoff, 2009, a banking crisis is defined by two different events: "bank runs that lead to the closure, merging, or takeover by the public sector of one or more financial institutions, or if there are no runs, the closure, merging, takeover, or large-scale government assistance of an important financial institution that marks the start of a string of similar outcomes". The authors note that this approach to dating crises is not perfect. Specifically, they note that this method may date crises too early, as the worst of a crisis may come later. This approach can also date crises too late, as financial problems often happen before a bank is fully merged or closed. The authors also classify the first event involving bank runs as a systemic crisis, whereas, in the absence of bank runs, the crisis is considered non-systemic. The authors define an inflation crisis as "an annual inflation rate of 20 percent or higher"(Reinhart & Rogoff, 2009). In this work, we use the consumer price index, or CPI, to measure inflation. We calculate the yearly change in CPI for each observation in the sample and code a new variable called crisisInflation. If the yearly change is greater than or equal to 0.2, crisisInflation is given a 1 for that observation. Lastly, we discuss currency crises. The authors define a currency crisis as "an annual depreciation versus the U.S. dollar of 15 percent or more"(Reinhart & Rogoff, 2009). The database includes an exchange rate variable against the USD for all observations. Similarly to the inflation crises, we calculate the change of this exchange rate from year to year for all observations and import a new variable called crisisCurrency. If the yearly depreciation is more than

| | X | year | country | crisisCurrencyRR | crisisInflationRR | crisisBanking | iso | crisisCurrency | crisisInflation | crisisJST |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1871 | Australia | 0 | 0 | 0 | AUS | 0 | 0 | 0 |
| 2 | 2 | 1871 | Belgium | 0 | 0 | 1 | BEL | 0 | 0 | 0 |
| 3 | 3 | 1871 | Canada | 0 | 0 | 0 | CAN | 0 | 0 | 0 |
| 4 | 4 | 1871 | Denmark | 0 | 0 | 0 | DNK | 0 | 0 | 0 |
| 5 | 5 | 1871 | Finland | 0 | 0 | 0 | FIN | 0 | 0 | 0 |
| 6 | 6 | 1871 | France | 0 | 0 | 0 | FRA | 0 | 0 | 0 |
| 7 | 7 | 1871 | Germany | 0 | 0 | 0 | DEU | 0 | 0 | 0 |
| 8 | 8 | 1871 | Ireland | 0 | NA | NA | IRL | 0 | 0 | NA |
| 9 | 9 | 1871 | Italy | 0 | 0 | 0 | ITA | 0 | 0 | 0 |
| 10 | 10 | 1871 | Japan | 0 | 0 | 0 | JPN | 1 | 0 | 1 |
| 11 | 11 | 1871 | Netherlands | 0 | 0 | 0 | NLD | 0 | 0 | 0 |
| 12 | 12 | 1871 | Norway | 0 | 0 | 0 | NOR | 0 | 0 | 0 |
| 13 | 13 | 1871 | Portugal | 0 | 0 | 0 | PRT | 0 | 0 | 0 |
| 14 | 14 | 1871 | Spain | 0 | 0 | 0 | ESP | 0 | 0 | 0 |
| 15 | 15 | 1871 | Sweden | 0 | 0 | 0 | SWE | 0 | 0 | 0 |
| 16 | 16 | 1871 | Switzerland | 0 | NA | NA | CHE | 0 | 0 | 0 |
| 17 | 17 | 1871 | UK | 0 | 0 | 0 | GBR | 0 | 0 | 0 |
| 18 | 18 | 1871 | USA | 0 | 0 | 0 | USA | 0 | 0 | 0 |

Figure 1.1: Various crisis variables used in the experiments

0.15, crisisCurrency is given a 1. Finally, we move on to the main database for this work.

The Jorda-Schularick-Taylor Macrohistory Database is the main database used in this work. It is a database constructed over several years of hard work in data collection. The foundations of the database are based on work done by the authors in years prior. This database is one of the main financial crisis databases and contains $\langle country, year \rangle$ observation pairs for 18 countries from 1870-2020. There is a variable called crisisJST, which is labelled as 1 in the first year that a country experiences a systemic banking crisis and zero elsewhere. One thing to note is that the definition mentioned above excludes the failures of small banks that do not have systemic implications. The database also contains 48 macroeconomic variables, such as unemployment rate, wages, GDP, interest rates, debt, and house prices.

Figure 1.1 shows a snapshot of the different crisis variables employed in this work. This figure only shows 1871, but these observations exist for all years up to 2020. The dataset begins in 1871 because the currencyCrisis and inflationCrisis variables are

constructed based on the changes between 1870 and 1871, thus there can be no 1870 observations. The variables crisisCurrencyRR and crisisInflationRR come from the narrative crisis lists from Reinhart and Rogoff, 2009. We report on the results of these two crisis definitions in the appendix A.3.4.

With the advent of machine learning models, it has become a focus in this area of the literature to try and use machine learning models to create early warning systems for financial crisis prediction. However, the literature shows some controversy about the usefulness of these algorithms in financial crisis prediction. Beutel et al., 2019 trains models on monthly data from 1970 to 2016 and concludes that a logistic regression almost always outperforms machine learning models. However, Bluwstein et al., 2021 looked at yearly data and a variety of machine learning models and showed that they generally outperform logistic regression in predicting financial crises. Liu et al., 2021 uses similar methods on monthly data and shows that multiple machine learning models outperform logistic regression in predicting banking, currency, and sovereign debt crises. Liu et al., 2021 came to the same conclusion as Bluwstein et al., 2021, that the decision tree is the only machine learning model that performs worse than logistic regression. Tölö, 2019 looks at the same dataset as Bluwstein et al., 2021 and concludes that neural networks can be useful models in financial crisis prediction.

The database constructed by Laeven and Valencia, 2018 has been widely used in this realm of literature for financial crisis prediction. One advantage of this database over the other main databases is that it includes monthly data, whereas the other main databases involve yearly observations. As mentioned above, Beutel et al., 2019 used a modified version of the Laeven and Valencia database and concluded that machine learning models do not provide an advantage over standard logistic regression. A few years later, Liu et al., 2021 trained a wide variety of machine learning models on the Laeven and Valencia dataset and concluded that many machine learning models

outperform logistic regression, specifically gradient-boosted trees and random forests.

The dataset constructed by Baron et al., 2020 has been on the rise recently as potentially the strongest list of banking crises. Greenwood et al., 2020 uses this dataset and concludes that financial crises are able to be reliably predicted ahead of time. While they do not employ machine learning techniques, Greenwood et al., 2020 does show that their results are robust across 3 of the main databases Jordà et al., 2017, Reinhart and Rogoff, 2009,Baron et al., 2020. Bluwstein et al., 2021 also employs this database to check the robustness in their work.

The Reinhart and Rogoff dataset was a masterpiece and a significant contribution to the literature when it was published (Reinhart & Rogoff, 2009). More recently, it has been used for robustness checks when authors would like to check their work with a different definition of a crisis. Chen et al., 2023 uses this dataset to confirm their findings on using textual data to predict financial crises. Casabianca et al., 2022 employs a synthesized version of the Laeven and Valencia database with the Reinhart and Rogoff database to create a wider sample of countries for analysis. They found that machine learning methods outperformed logistic regression.

When looking for annual data that is useful for financial crisis prediction, the Macrohistory database (Jordà et al., 2017) is more commonly used. Bluwstein et al., 2021, the main inspiration behind this work, use this database for their work. Bluwstein et al., 2021 and Reimann, 2024 both identify that the JST database is desirable because it contains accurate yearly data for advanced economies. They find that machine learning models outperform logistic regression in both cross-validation experiments and forecasting experiments. Tölö, 2019 also uses the Macrohistory database but look specifically at the neural network. They conclude that recurrent neural networks can outperform logistic regression models in both forecasting and cross-validation, although their models are generally weaker than Bluwstein et al., 2021. Reimann, 2024 has recently updated

the work done by Bluwstein et al., 2021, using the fifth version of the Macrohistory database and showing similar findings. Namely, that tree-based methods can be useful in predicting financial crises.

A common critique of machine learning models is that they are *black boxes*. The models can achieve high accuracy in prediction, but it is often difficult to determine what is driving the predictions of these models. A recent line of work, developed by Lundberg and Lee, 2017, made use of Shapley values from cooperative game theory. This framework allows for the identification of important predictors in the models. Combining this with the Shapley regressions presented in Joseph, 2019, one can attribute statistical significance to the predictors and better understand the models. Bluwstein et al., 2021 employs both of these methods to gain insight into what are the driving indicators of banking crises. Liu et al., 2021 also employs both the SHAP method and the Shapley regressions. Reimann, 2024 employs a different method for interpreting machine learning models, namely accumulated local effects (ALEs) built upon partial dependence plots. They allow for the decomposition of machine learning models similar to the Shapley value framework (Reimann, 2024).

This work contributes to the literature in the following ways. Firstly, the dataset used in this work is the first to our knowledge to use the sixth version of the Macrohistory database (Jordà et al., 2017). Our work investigates whether or not the updated database affects the results found in the literature, either positively or negatively. Reimann, 2024 updated the work of Bluwstein et al., 2021, and our work is one step further in updating this work. Second, our work employs the commonly used SHAP method and Shapley regressions to crises that are not banking crises (Joseph, 2019). We also investigate the driving predictors and nonlinear relationships between the predictors of currency and inflation crises as described in Bluwstein et al., 2021. Our work also ensures that all the main machine learning algorithms discussed in the literature

are being compared in this study. To our knowledge, our work is also the first to use machine learning methods to look at a long sample for currency and inflation crises.

# Chapter 2

# Methods

## 2.1 Updating the Reinhart and Rogoff crises

The Macrohistory database contains all the variables of interest from 1870-2020. We discuss the missing values in the appendix. The Reinhart and Rogoff crisis definitions are only updated to 2016. Our first task is to manually determine the values for the crisis variables for 2017-2020. Reviewing the literature shows many discrepancies in when a banking crisis happens across the different narrative approaches. Baron et al., 2020 notes that a problem with these narrative approaches is that some will place a banking crisis in a year because some previous database had a crisis for that country in that year. Since Baron's crisis list aims to be overly inclusive so as to not miss a crisis, we append Baron's definition of a banking crisis to the Reinhart and Rogoff list for the years 2017-2019. Tomczak, 2023 concludes that the COVID-19 pandemic increased risk in the banking sector but did not lead to any banking crises. Applying both of these works, we conclude that no banking crises occurred in any of the 18 countries for the years 2017-2020. For a currency crisis, we calculate the value change of the currency relative to the USD by calculating 1-year growth rates in the JST database on the

`xrusd` variable (Jordà et al., 2017). We then apply the 15% threshold to the growth rates and determine when a currency crisis occurred. A similar approach is used to determine inflation crises, but the 20% threshold is applied to the consumer price index growth rate data in the JST database (Jordà et al., 2017).

## 2.2 Features

The problem presented in this work is treated as a classification problem, where each observation is treated as independent. The following predictors are used in the baseline analysis: debt service ratio, the slope of the yield curve, credit growth, consumption, investment, the current account, public debt, broad money, consumer price index, nominal wages, and the unemployment rate. The slope of the yield curve is calculated as the difference between short and long-term interest rates and is left in levels (i.e., with no further transformation applied to it). The consumer price index and consumption are transformed into percentage growth rates of the given variables. Wages and unemployment rate are transformed into 2-year percentage growth rates. All other variables are 2-year differences of GDP ratios. In addition to these variables, we also use two variables in an attempt to capture any global trends that may be present when crises happen (Bluwstein et al., 2021). In particular, we define global credit growth and the global slope of the yield curve. These variables are calculated for a $\langle c, y \rangle$ observation by taking the mean of the domestic variable in all countries except $c$ in year $y$. For example, the global credit growth feature for Canada in 2000 is calculated in the following way. All ratios are computed over a two-year horizon. The domestic credit is calculated for all countries except Canada in 2000 and in 1998. The two-year difference is calculated for each country, and then the mean is taken to be the global credit variable for Canada. Next, we will discuss why these features are appropriate for this

problem and how they have been used in the past.

In the literature, credit growth has been found to be a crucial predictor of banking crises (Schularick & Taylor, 2012). Oftentimes, high credit growth is reflective of high risk-taking. Intuitively, periods of high risk can lead to financial instability. Further, Aliber and Kindleberger, 2015 describes financial crises as "credit booms gone wrong". In fact, Bernanke and Blinder, 1992 shows that even a relatively small credit bubble can be detrimental due to financial accelerator effects. There have been several studies that have argued for the importance of looking at global credit growth. The main argument for global variables lies in the fact that "financial crises often occur on an international scale and may reflect global financial cycles" (Bluwstein et al., 2021). Furthermore, Cesa-Bianchi et al., 2019 shows that global credit growth is a stronger predictor than domestic credit.

The difference between long and short-term interest rates, referred to in this work as the slope of the yield curve, has been identified in the literature as a strong predictor of recessions (Bluwstein et al., 2021). Some work has also identified the slope of the yield curve as an important predictor of crises, but those papers have not investigated why this feature has such high predictive power. The yield curve aims to reflect the risk premium for holding an asset. Bluwstein et al., 2021 notes that the slope will be positive during normal times, meaning that long-term interest rates are higher than short-term interest rates. Bluwstein et al., 2021 notes that there are two main reasons why a flat or negative yield curve slope may have power for predicting financial crises. First, a flatter yield curve is typically associated with lower banking sector profitability and net interest margins. Because of this, the banking sector's resilience may be in jeopardy, particularly if these effects are severe. Second, flat or inverted yield curves will produce lower-term premiums. If this is the case, investors will likely have to search for riskier investments to obtain higher returns. For example, Bluwstein et al., 2021 notes that

house prices rose as the yield curve flattened before the global financial crisis. This suggests "the hunger for spread during this period of a flat yield curve could have been fuelling sub-prime and other alternative mortgage activity" (Bluwstein et al., 2021). This system-wide build-up can leave the financial system quite fragile and susceptible to crises. Similarly to credit growth, there are also arguments for looking at a global yield curve. At a global level, a flattening yield curve could be indicative of a global economic slowdown, which could very well trigger any financial instabilities that are present, resulting in crises.

The debt service ratio has been identified in the literature as a reliable early warning indicator (Bluwstein et al., 2021). It tells us the interest rate payments relative to one's income. Thus, it provides a measurement of how overextended borrowers may be. That is the higher the debt service ratio, the more likely a borrower is to experience a fall in income or a rise in interest rate. The more overextension that is occurring in borrowing can lead to increased default rates, a decrease in new investment, or a loss in consumption smoothing capabilities. In this paper, we employ a rather simple version of the debt service ratio, which is defined as the credit multiplied by the long-term interest rate over GDP. Data availability is the main reason for this definition of the debt service ratio. The downside to this definition is that we do not get insight into short-term interest rates, capital repayments, or the maturity structure of the debt, which may all be important (Bluwstein et al., 2021).

Imbalances in the current account have been found to be strong predictors of financial crises since capital inflows push interest rates down and then, similarly to the justification above, lead to excessive risk-taking. In an attempt to capture crises caused by fiscal vulnerabilities, we include public debt as a feature. Public debt is defined as the total amount a country owes to outside lenders. We look at this as a difference of GDP ratio.

To round out the set of features, we aim to get a sense of the macroeconomic trends. Thus, real consumption, investment, the consumer price index, and the money supply are all included.

| | transformation | Build up Mean | SD | Non crisis Mean | SD | Difference | p |
|---|---|---|---|---|---|---|---|
| Yield curve slope | level | -0.14 | 1.51 | 1.01 | 1.65 | -1.15 (0.178) | 0.000 |
| Credit | 2-year difference of GDP-ratio × 100 | 7.62 | 10.34 | 1.95 | 6.10 | 5.67 (1.192) | 0.000 |
| CPI | 2-year growth rate of index | 5.60 | 9.17 | 8.33 | 9.91 | -2.73 (1.084) | 0.014 |
| Debt service ratio | 2-year difference of GDP-ratio × 100 | 0.66 | 1.16 | -0.07 | 1.14 | 0.72 (0.137) | 0.000 |
| Consumption | 2-year growth rate of index | 4.57 | 5.75 | 4.83 | 4.91 | -0.26 (0.671) | 0.703 |
| Investment | 2-year difference of GDP-ratio × 100 | 1.27 | 3.71 | 0.35 | 2.50 | 0.92 (0.429) | 0.035 |
| Public debt | 2-year difference of GDP-ratio × 100 | -1.31 | 6.71 | -0.36 | 8.51 | -0.95 (0.803) | 0.240 |
| Broad money | 2-year difference of GDP-ratio × 100 | 2.60 | 6.16 | 1.16 | 5.19 | 1.44 (0.718) | 0.049 |
| Unemployment Rate | 2-year growth rate | 12.31 | 77.45 | 14.23 | 141.50 | -1.92 (9.740) | 0.844 |
| Wages | 2-year growth rate | 10.67 | 14.20 | 12.87 | 12.77 | -2.20 (1.660) | 0.189 |
| Current account | 2-year difference of GDP-ratio × 100 | -1.06 | 3.21 | 0.01 | 2.80 | -1.07 (0.375) | 0.005 |
| Global yield curve slope | level | 0.27 | 0.75 | 0.99 | 0.83 | -0.72 (0.089) | 0.000 |
| Global credit | 2-year difference of GDP-ratio × 100 | 5.37 | 3.89 | 1.94 | 2.90 | 3.43 (0.452) | 0.000 |
| Nominal short-term rate | level | 6.29 | 3.68 | 5.01 | 3.96 | 1.28 (0.435) | 0.004 |
| Nominal long-term rate | level | 6.15 | 3.36 | 6.02 | 3.73 | 0.13 (0.398) | 0.742 |
| Household loans | 2-year difference of GDP-ratio × 100 | 3.86 | 5.08 | 1.59 | 3.68 | 2.27 (0.690) | 0.002 |
| Business loans | 2-year difference of GDP-ratio × 100 | 5.16 | 7.01 | 0.30 | 4.19 | 4.85 (0.981) | 0.000 |
| House prices (index) | 2-year growth rate of index | 17.69 | 18.68 | 14.65 | 18.21 | 3.04 (2.257) | 0.181 |

Table 2.1: Overview of the variables for observations one and two years before a crisis (build-up) and non-crisis observations. crisisJST used as the outcome variable.

Table 2.1 is an updated version of the same table from Bluwstein et al., 2021. It shows some preliminary descriptive statistics for the features used in the baseline model and robustness checks. This table investigates the differences in the features in the two years leading up to a crisis versus any of the non-crisis years. The difference between the two means is calculated, and a $t-$test is used to determine statistical significance. The key takeaway from this table is that there is often a statistically significant difference between the means of the features in a build-up versus a non-crisis observation. Indeed, we see that $p < 0.01$ for the global and domestic yield curve slope, the global and domestic credit growth, and the current account in the baseline experiment. The nominal short-term rate, household loans, and business loans used in robustness checks also have statistically significant differences of means. This table alludes to the potential importance of these features in predicting banking crises.

Table 2.2 looks at the differences of the predictors in pre-crisis periods for currency crises. We see that there are statistical differences in the means for the global slope of the yield curve, the global credit growth, and the CPI. We do notice that compared to Table 2.1, Table 2.2 has fewer statistically significant differences between the means of crisis and non-crisis observations. This table suggests that interest rate changes are very prominent in the build-up to currency crises. This is known in the literature on currency crises (Kaminsky & Reinhart, 1997). Kaminsky and Reinhart also reviewed the literature on currency crises and found that the theoretical approach uses the current account, the public debt, wages, unemployment rate, and investment as indicators for crises (Kaminsky & Reinhart, 1997).

| | | Build | up | Non | crisis | | |
|---|---|---|---|---|---|---|---|
| | transformation | Mean | SD | Mean | SD | Difference | p |
| Yield curve slope | level | 0.81 | 1.64 | 1.03 | 1.83 | -0.22 (0.192) | 0.255 |
| Credit | 2-year difference of GDP-ratio × 100 | 2.59 | 6.76 | 1.36 | 6.36 | 1.23 (0.785) | 0.121 |
| CPI | 2-year growth rate of index | 11.17 | 9.28 | 8.11 | 10.35 | 3.06 (1.091) | 0.006 |
| Debt service ratio | 2-year difference of GDP-ratio × 100 | -0.14 | 1.29 | -0.03 | 1.22 | -0.11 (0.150) | 0.450 |
| Consumption | 2-year growth rate of index | 4.95 | 5.40 | 4.31 | 5.17 | 0.64 (0.627) | 0.310 |
| Investment | 2-year difference of GDP-ratio × 100 | -0.07 | 3.85 | 0.19 | 2.38 | -0.25 (0.439) | 0.565 |
| Public debt | 2-year difference of GDP-ratio × 100 | 0.11 | 15.33 | 0.51 | 8.23 | -0.40 (1.744) | 0.820 |
| Broad money | 2-year difference of GDP-ratio × 100 | 0.36 | 5.32 | 1.10 | 5.17 | -0.74 (0.619) | 0.234 |
| Unemployment Rate | 2-year growth rate | 17.92 | 76.34 | 11.20 | 58.71 | 6.72 (8.776) | 0.446 |
| Wages | 2-year growth rate | 15.24 | 13.52 | 12.58 | 13.51 | 2.66 (1.576) | 0.094 |
| Current account | 2-year difference of GDP-ratio × 100 | 0.50 | 3.55 | 0.01 | 2.76 | 0.49 (0.408) | 0.236 |
| Global yield curve slope | level | 0.82 | 0.47 | 1.02 | 0.90 | -0.20 (0.060) | 0.001 |
| Global credit | 2-year difference of GDP-ratio × 100 | 3.15 | 2.80 | 1.50 | 3.01 | 1.66 (0.329) | 0.000 |
| Nominal short-term rate | level | 6.72 | 4.04 | 4.78 | 4.00 | 1.94 (0.470) | 0.000 |
| Nominal long-term rate | level | 7.53 | 3.54 | 5.81 | 3.84 | 1.72 (0.416) | 0.000 |
| Household loans | 2-year difference of GDP-ratio × 100 | 1.84 | 3.57 | 0.99 | 3.59 | 0.84 (0.437) | 0.057 |
| Business loans | 2-year difference of GDP-ratio × 100 | 1.01 | 4.80 | 0.03 | 4.65 | 0.99 (0.593) | 0.100 |
| House prices (index) | 2-year growth rate of index | 13.67 | 14.25 | 12.66 | 18.53 | 1.02 (1.753) | 0.562 |

Table 2.2: Overview of the variables for observations one and two years before a crisis (build-up) and non-crisis observations. crisisCurrency used as the outcome variable.

Table 2.3 shows the same descriptive statistics for the predictors for pre-crisis and non-crisis observations for inflation crises. Note that the CPI predictor is omitted since the definition of crisisInflation is based on that feature. We see that the debt servicing ratio, consumption, wages, interest rates, and house prices have statistically significant differences in their mean values in pre-crisis versus non-crisis periods. Interestingly, the

yield curve slopes, both domestically and globally, do not have statistical differences despite those predictors being constructed from the interest rates. Stock and Watson, 1999 looked at a wide variety of indicators for forecasting inflation and found that spreads between interest rates of different maturities (yield curve) are a leading indicator of inflation crises. They also talk extensively about the potential importance of the unemployment rate as a predictor of inflation.

| | | Build up | | Non crisis | | | |
|---|---|---|---|---|---|---|---|
| | transformation | Mean | SD | Mean | SD | Difference | p |
| Yield curve slope | level | 0.72 | 2.14 | 0.98 | 1.70 | -0.25 (0.574) | 0.667 |
| Credit | 2-year difference of GDP-ratio × 100 | 3.27 | 5.61 | 2.21 | 7.07 | 1.06 (1.511) | 0.494 |
| Debt service ratio | 2-year difference of GDP-ratio × 100 | 0.83 | 1.21 | -0.06 | 1.18 | 0.88 (0.325) | 0.017 |
| Consumption | 2-year growth rate of index | 9.19 | 4.02 | 4.24 | 4.82 | 4.96 (1.082) | 0.000 |
| Investment | 2-year difference of GDP-ratio × 100 | 1.75 | 3.89 | 0.19 | 2.31 | 1.56 (1.042) | 0.158 |
| Public debt | 2-year difference of GDP-ratio × 100 | -2.26 | 7.31 | 0.29 | 8.59 | -2.55 (1.968) | 0.218 |
| Broad money | 2-year difference of GDP-ratio × 100 | 2.76 | 4.98 | 1.33 | 5.24 | 1.43 (1.338) | 0.305 |
| Unemployment Rate | 2-year growth rate | 4.88 | 34.06 | 14.44 | 134.01 | -9.56 (9.813) | 0.343 |
| Wages | 2-year growth rate | 38.79 | 14.61 | 10.59 | 10.75 | 28.20 (3.917) | 0.000 |
| Current account | 2-year difference of GDP-ratio × 100 | -1.21 | 2.93 | 0.08 | 2.65 | -1.29 (0.786) | 0.124 |
| Global yield curve slope | level | 0.67 | 0.95 | 0.97 | 0.92 | -0.31 (0.255) | 0.250 |
| Global credit | 2-year difference of GDP-ratio × 100 | 1.77 | 1.42 | 2.11 | 3.35 | -0.34 (0.389) | 0.398 |
| Nominal short-term rate | level | 9.09 | 3.45 | 4.75 | 3.72 | 4.34 (0.928) | 0.000 |
| Nominal long-term rate | level | 9.81 | 4.18 | 5.73 | 3.39 | 4.08 (1.122) | 0.003 |
| Household loans | 2-year difference of GDP-ratio × 100 | 0.93 | 1.99 | 1.65 | 3.98 | -0.71 (0.614) | 0.269 |
| Business loans | 2-year difference of GDP-ratio × 100 | 1.87 | 5.52 | 0.58 | 4.89 | 1.29 (1.672) | 0.457 |
| House prices (index) | 2-year growth rate of index | 42.29 | 16.27 | 12.56 | 17.67 | 29.73 (4.723) | 0.000 |

Table 2.3: Overview of the variables for observations one and two years before a crisis (build-up) and non-crisis observations. crisisInflation used as the outcome variable.

## 2.3 Experimental Procedure

This work uses many of the same methods from *Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach* (Bluwstein et al., 2021). As mentioned in the introduction, we aim to build models that can predict crises ahead of time. When choosing which definition of crisis we are using, we transform the outcome variable to be 1s in the 2 years preceding the crisis while removing the actual crisis year and 4 years after to avoid post-crisis bias (Bluwstein et al., 2021). For example, a crisis in the United States in 2008 will result in the United States ob-

servations for 2006 and 2007 having positive outcomes and the years 2008-2012 being removed from the data. Thus, we are trying to correctly predict pre-crisis years. The main approach involves using cross-validation to evaluate the performance of the models. This method uses 5-fold cross-validation, which means that each $\langle country, year \rangle$ observation is assigned to one of 5 groups, called folds. For all assignments of folds, all observations related to the same crisis event are assigned to the same fold. Then, the models are trained using 4 of the folds and then evaluated on the remaining fold. Thus, each fold for the cross-validation contains 20% of the observations, and the models are trained using 80% of the observations. To correct for potential biases in the assignment of the folds, we repeat the assignment of the folds at least 50 times. Due to the nature of this training, crises that happened in the past will be predicted using observations from the future. In practice, these early warning models would more likely use a recursive forecasting approach, where the model only uses past data to predict whether a crisis occurred or not. Some of the models employed in this analysis require tuning hyperparameters that affect the performance. The training set cannot optimize these hyperparameters since the most flexible structure would always obtain the best results (Bluwstein et al., 2021). Thus, *nested cross-validation* is employed to optimize the hyperparameters. This involves applying 5-fold cross-validation within each of the 5 folds created in the cross-validation. The best combination of hyperparameters in this nested cross-validation is what is used to train the model on the whole fold.

The models will be analyzed by comparing their area under the receiver-operator curve (AUROC) scores. A model will be considered better than another if it obtains a higher AUROC score. The ROC curve shows the true positive rate on the vertical axis, known as the hit rate in this work. The horizontal axis shows the false positive rate, which is called the false alarm rate in this paper. A perfect model would have a hit rate of 1 and a false alarm rate of 0. However, choosing a higher hit rate in practice

will lead to more false alarms (Bluwstein et al., 2021).

**Definition 2.** *Hit rate*

$$Hit\,rate = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}.$$

**Definition 3.** *False Alarm Rate*

$$FAR = \frac{False\ Positives}{False\ Positives\ +\ True\ Negatives}.$$

All models will be calibrated to have a hit rate of 80%; however, this number is somewhat arbitrary. The percentage of crises detected is more of a modeller's choice, and 80% has been seen in the literature before (Bluwstein et al., 2021),(Reimann, 2024). Another justification for this high hit rate is that it is more costly to miss a financial crisis than to be overly cautious in preventing it. Since the same hit rate is chosen for all the models, a lower false alarm indicates that one model is better than another. Another common model comparison technique is the noise-to-signal ratio. We do not use this comparison technique from Kaminsky and Reinhart, 1997 because it provides no new information compared to the FAR when a constant hit rate is chosen, since $NSR = FAR/0.8$.

We then investigate the driving forces behind the predictions of these models by applying the Shapley value approach (Lundberg & Lee, 2017),(Shapley, 1953). We decompose the models' predictions into their Shapley values and see which features contribute the most to predictions. We also investigate the role of different predictors over time. Next, we employ the Shapley regressions to test the statistical significance of the features (Joseph, 2019). Finally, we investigate the nonlinear relationships between the predictors and their Shapley values (Bluwstein et al., 2021).

This entire procedure is applied to the 4 different definitions of financial crises that

were discussed in section 1.2. We begin by looking at a logistic regression model trained on the data as a benchmark model. The machine learning models will try to outperform this model later in the paper.

## 2.4 Logistic Regression - A benchmark

In order to determine whether or not machine learning models are potentially useful for financial crisis prediction, we need a benchmark model with which to compare them. As shown throughout the literature on early warning systems, logistic regression is the most commonly used model for this task (Namaki et al., 2023). To help compare the predictive power of the individual variables in the model, all variables are standardized to have a mean of 0 and a standard deviation of 1.

All 13 of the baseline features are included. For the JST definition of a banking crisis, Table 2.4 shows that global credit is the most important feature, followed by the domestic slope of the yield curve. We see that the slope of the yield curve's weight in the model is negative, which is expected. This indicates that a negative slope will correspond to a higher probability of a crisis. The same is true for CPI, consumption, current account, and broad money. These features increase the predicted probability if the features themselves decrease.

Comparing the coefficients of these logistic regression models between the two banking crises shows quite a few discrepancies. Firstly, we see less statistically significant features and more negative coefficients for the Reinhart and Rogoff banking crisis. It appears that the domestic credit growth is much less important for predicting this type of crisis. On the other hand, the current account plays a much larger role, having a statistically significant positive coefficient. We can see from the AUROC scores that the JST definition of a crisis is slightly easier to predict than the Reinhart and Rogoff

definition.

Turning our attention to currency crises, we see an immediate drop in model performance. This suggests that currency crises are harder to predict, or that the features chosen are not able to capture the trends in the pre-crisis periods. We immediately get the sense that credit growth, both domestically and globally, as well as the CPI are the most important predictors since they have statistically significant coefficients that are larger than any of the other coefficients. We also see a very small effect from the domestic slope of the yield curve compared to the global slope of the yield curve, suggesting there may be global effects at play in currency crises. This is in line with the construction of the currency crisis variable, since it is dependent on the United States exchange rate.

Logistic regression achieves its best performance when predicting inflation crises. Note that Table 2.4 omits the CPI in the model since CPI was used to construct the crisisInflation variable. Investment, wages, and broad money are all statistically significant predictors in this model, generating some large weights compared to the other crisis types. We see a rather large weight is given to the wages, indicating that it is perhaps a good indicator of inflation crises. Both global variables have negative weights indicating that countries may be more susceptible to inflation crises if the global economy is suffering.

Table 2.4 shows the number of observations and positive indicators for each crisis as well. Since 4 years are removed after a crisis, we expect there to be less observations when there are more crises. Indeed, there are almost 100 more observations for inflation crises compared to all the other types since there are so few positive indicators. Note that a positive indicator represents 1 or 2 years before a crisis occurs.

| | JST | Banking | Currency | Inflation |
|---|---|---|---|---|
| Domestic credit | 0.3560*** | -0.0908 | 0.4598** | 0.0778 |
| Global credit | 0.9543*** | 0.0701 | 0.6128** | -0.5909 |
| Domestic slope | -0.7120*** | -0.1437 | 0.0367 | 0.1691 |
| Global slope | -0.4771** | -0.0402 | -0.2176 | -0.5999 |
| CPI | -0.6270* | -0.1788 | 0.5010** | – |
| Debt service ratio | 0.4166** | 0.3138** | -0.3554 | 0.5601 |
| Consumption | -0.1820 | 0.0718 | 0.2086 | 0.3400 |
| Investment | 0.2242 | 0.0978 | 0.0268 | 0.4470* |
| Public debt | 0.0069 | -0.0588 | -0.0360 | -0.3672 |
| Broad money | -0.2518 | -0.0049 | -0.3219 | 0.6571** |
| Wages | 0.3078 | -0.1498 | -0.2137 | 1.045** |
| Unemployment | 0.1785* | 0.0550 | 0.1169 | -0.3714 |
| Current account | -0.1615 | 0.2072** | 0.3088* | 0.2295 |
| AUROC | 0.82 | 0.79 | 0.65 | 0.84 |
| Observations | 1257 | 1108 | 1141 | 1352 |
| Positive Indicators | 77 | 83 | 79 | 14 |

Table 2.4: Baseline logistic regression model. Feature weights shown for all crisis types. Significance levels: $***: p < 0.01, **: p < 0.05, *: p < 0.10$.

## 2.5 Machine Learning Models

In this section, we provide a brief introduction to machine learning methodology. After this brief introduction, we will discuss each machine learning model used in more detail.

Let $f$ represent a particular prediction of our model $\hat{y} = f(\mathbf{X})$. $\mathbf{X}_{n\times k}$ is our predictor matrix which contains $n$ observations and $k$ features. In our classification problem framework $\hat{y} \in [0, 1]$ represents the predicted probability of a crisis. A crisis year for an observation is denoted by $y \in \{0, 1\}$ where 1 indicates a pre-crisis year and 0 indicates no crisis. Thus, pre-crisis years are the positive class for our work, and the absence of a crisis is the negative class.

In this work, we use a diverse set of machine learning models that are suitable for classification. The models range in complexity and interpretability. The following descriptions of these algorithms are rather high-level to provide a general understanding of how the algorithms work. For an in-depth explorations of any of these algorithms,

please consult the extensive literature available. The appendix includes details on the implementation of all these algorithms in Python.

### 2.5.1 Decision trees

A decision tree is a simple machine learning model that is highly interpretable. It works by splitting the data into subsets and testing a single predictor at each step. For example, it will split the data into two groups based on whether the credit growth is greater than or less than 1%. Every observation in one group has a credit growth greater than 1%, whereas every observation in the other group has a credit growth less than 1%. This process is then recursively repeated in each subset with the goal of obtaining an optimal split.

Because of their relative simplicity, decision trees are very flexible models. However, they are also prone to *overfitting*. Overfitting is the phenomenon in machine learning where a model will not only learn the patterns in the data but also fit the noise in the data. This leads the model to perform well on the data it was trained on but poorly on a new set of observations. To help reduce the effects of overfitting in decision trees, there exist many different *pruning* techniques which control the size of the decision tree. The idea behind pruning is not to allow the tree to learn too much about the data. Decision trees perform worse than other machine learning models, but they are included in this work due to their simplicity and high interpretability.

### 2.5.2 Random forests

A random forest is a collection of many decision trees. In the random forest algorithm, the prediction is made by averaging the predictions of this collection of trees. This technique drastically reduces the amount of overfitting that occurs since the overall variance in the predictions is reduced. Each individual decision tree may still overfit,

but averaging the predictions somewhat cancels out the noise and improves performance on new data. However, this only works if the trees are different from each other since similar trees will fit the noise in a similar way. The random forest algorithm employs two techniques to try and achieve this. First, each tree is trained on a different subset of the data made with replacement. This method is known as *bagging*, or bootstrap aggregating, and is a general technique for improving the stability of models Breiman, 1996. Second, a random sample of $m$ candidates from the $k$ predictors is chosen, and the algorithm optimizes the forest for this subset. In a forest, given an observation, each tree predicts either the positive or negative classes. The mean of these predictions is the probability that the algorithm gives.

Due to all of these adjustments, a random forest often performs much better than an individual decision tree. However, this comes at the cost of interpretability, a general machine learning trend. Since the prediction is determined by many trees, it cannot be decomposed into a simple flow chart like a decision tree. Fernandez-Delgado et al., 2014 compared 179 classification algorithms on 121 real-world datasets, and random forests were the best-performing algorithm on average.

### 2.5.3 Extremely randomized trees

Extremely randomized trees is yet another tree-based algorithm. It is similar to random forests but tends to produce predictions that are continuous as a function of the predictors (Bluwstein et al., 2021). This is done by creating trees that are more diverse and random. There are two main differences between extremely randomized trees and random forests. First, each tree in the extremely randomized trees algorithm is trained on the complete training dataset. Second, the splitting for each tree is randomized rather than optimized. That is, splits are made randomly across the range of the predictor for each of the $m$ predictors that are randomly sampled. Then, the best split

among the random splits is used to make the prediction.

When extremely randomized trees was first discovered, it was believed to achieve comparable performance to random forests, most of the time slightly edging it out (Geurts et al., 2006). Since then, this belief has been challenged, as it does not always outperform random forests (Fernandez-Delgado et al., 2014).

### 2.5.4 XGBoost

Since Random Forest and Extremely Randomized Trees are bagging machine learning algorithms, we decided to employ XGBoost, a boosting algorithm. XGBoost stands for extreme gradient boosting. Boosting is a machine learning term that refers to training many models sequentially to make a group of weak learners into one strong learner (Santhanam et al., 2017). XGBoost improves upon basic boosting by having each subsequent model trained to reduce the error of the previous model. XGBoost also uses parallelization to decrease computation time (Santhanam et al., 2017). XGBoost presents similar postives and negatives when compared to the other decision tree ensemble methods. It is likely to outperform a single decision tree, but at the cost of interpretability.

### 2.5.5 Support vector machines

A support vector machine shares a few similarities to logistic regression. First, a support vector machine algorithm learns a linear function of the inputs. However, the inputs in a support vector machine are first transformed using a non-linear kernel function, which allows them to model non-linear classification problems (Bluwstein et al., 2021). A kernel takes the data and transforms it into a space of higher dimension, in which the algorithm can learn to linearly separate the positive from the negative class. Fernandez-Delgado et al., 2014 concluded from their comprehensive study that support vector

machines were the second-best algorithm, only performing worse than random forests. The major downside to using support vector machines is that the use of the non-linear kernel makes the algorithm almost a complete black box. We have almost no insight into how the model is generating its predictions.

### 2.5.6 Artificial neural networks

Artificial neural networks are the most researched machine learning technique recently (Bluwstein et al., 2021) (Tölö, 2019). These models have achieved astounding results in different types of classification problems, including face recognition and speech recognition (Bluwstein et al., 2021). The basic structure of a neural network includes an input layer, at least one hidden layer, and an output layer. Inputs are passed through the hidden layers until they are aggregated into a prediction in the output layer. An interesting thing to note is that in the absence of a hidden layer, a neural network is a linear function. All the different nodes in a hidden layer are connected to the previous hidden layer by different weights. Given a dataset with $k$ features and $m$ nodes in a hidden layer, $k \times m$ weights are needed to connect the hidden layer to the input layer (Bluwstein et al., 2021). Since this is a binary classification problem, the output layer is just a single node, so only $m$ weights are needed to connect the last hidden layer to the output layer. Due to their construction, neural networks are computationally expensive. Indeed, the training time required for the support vector machines and neural networks in this work was at least 100 times larger than for the tree based ensembles.

### 2.5.7 K-Nearest Neighbours

When there is little to no prior knowledge of the dataset, the K-nearest neighbours algorithm is one of the computationally cheapest (Imandoust & Bolandraftar, 2013). The algorithm works by comparing all observations to those around them. Then, groups are

made based on the value of $K$. The observation will be given the label that matches the majority label in the $K$ observations nearest to it. The simplest form of this algorithm is when $K = 1$. In this case, an unseen observation will be given the same label as the observation in the training set that is nearest to it. Observation 1 is considered *nearer* to observation 2 than observation 3 if the distance between observation 1 and observation 2 is less than the distance between observation 1 and observation 3 (Imandoust & Bolandraftar, 2013). Thus, different metrics for calculating distance can be used in this algorithm.

## 2.6 Shapley Values

### 2.6.1 Shapley Additive Explanations

We move on to how we will address the *black box* nature of these models. There are several popular techniques in the literature on this subject. Partial dependence plots, or PDPs, show the marginal effects of 1 or 2 features on the predicted probability in a model. These plots show the relationship between a feature and the output. However, a downside to PDPs is that they cannot capture the interactions between all of the features and suffer if features are highly correlated (Apley & Zhu, 2020). Accumulated local effects, or ALEs, build upon the PDP framework. The advantage that ALEs give over PDPs is that they are faster to compute and they are unbiased (Apley & Zhu, 2020). The goal of both of these methods is the same: to visualize the relationship that 1 or 2 features have with the output of a model (Apley & Zhu, 2020).

In this work, we chose to use **SH**apley **a**dditive ex**pl**anations, or SHAP, as the method for determining feature importance. The reason we chose to use SHAP over PDPs or ALEs is that SHAP has the advantage of seeing all of the interactions between features. This is because SHAP calculates the marginal payoffs for all the different

combinations of the features. SHAP also has a well-defined package in Python which makes it easy to implement. Finally, SHAP was the method used in determining feature importance in Bluwstein et al., 2021, which serves as the foundation for this work. Thus, for comparability, we employ SHAP in this work. We now move on to a more detailed description of how these Shapley values work.

The machine learning models used in this work are all able to approximate any well-behaved function if given enough training data. However, a common critique of these machine learning models is that they can be difficult to interpret. Specifically, it can be a challenge to identify which features of the dataset are driving a given prediction. Thus, this work employs the *Shapley additive explanations* (Lundberg & Lee, 2017),(Lundberg, 2018) method for determining feature importance. The main idea behind this method is using Shapley values from cooperative game theory. Given a team of players and a payoff, one can calculate the distribution of that payoff across the players in a team (Shapley, 1953). When applied to a machine learning framework, the team of players is the set of features used to train the model, and the payoff is the value of the predicted probability of whether a crisis occurs or not. Thus, one can decompose the predicted crisis probability into a sum of crisis probabilities from each feature, which are called the Shapley values (Shapley, 1953). Using this framework allows one to understand which features play a large role in driving the predictions of the model. Lundberg and Lee, 2017 point out in their work that this framework can be applied to any model and has some nice analytical properties. It has the following properties: local accuracy, additivity, missingness, and consistency (Awan, 2023). The local accuracy property refers to the fact that the Shapley values will sum to the difference between the model prediction and the actual prediction. Additivity refers to the fact that the Shapley values can be computed independently and then summed. Because of this, the computation of Shapley values is said to be efficient, even with large

datasets (Joseph, 2019). The missingness property means that if a feature is missing or irrelevant to a prediction, then its Shapley value will be zero (Awan, 2023). This gives Shapley values robustness to missing data and ensures that irrelevant features do not distort the prediction. Lastly, the consistency property says that the Shapley values for a model do not change unless a contribution of a feature changes. Thus, this approach can provide consistent interpretations, even when the parameters of the model change.

The following description of the calculations behind the Shapley values comes from *Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach*(Bluwstein et al., 2021). Suppose we have $n$ observations and $k$ features; we begin by defining a Shapley value matrix $\Phi_{n\times k}$ and $\phi_{ij}$ as the Shapley value of observation $i$ and predictor $j$. Every observation $i$ will have a predicted probability associated with it. This predicted probability can be decomposed into the sum of the Shapley values $\hat{y}_i = \sum_{j=1}^{k} \phi_{ij} + c$, where c is the base value that is set to the mean predicted value in the training set.

Let $N$ be the set of all the players in the game, and let $f(S)$ be the payoff of a coalition of players $S$. $S$ is a subset of $N$. The following formula calculates the Shapley value for player $j$:

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

In this work, we treat the predicted probability that the models compute to be the payoff in the Shapley value framework. The $N$ players correspond to the features used in the models. As mentioned above, the Shapley values must be computed for every individual observation where we want to explain the predicted probability. To calculate the exact Shapley value for a particular variable $j$ and observation $i$, we need to compute how much predictive value feature $j$ adds to $(f_i(S \cup j) - f_i(S))$ for all the

possible subsets of the other features in ($S \subseteq N \setminus j$). To help with this explanation, the following example is explained. Suppose you have three regressors in a linear model, and $\hat{y}$ is the payoff. First, you compute all the regressions with one, two, and three regressors and then look at the marginal contributions of each regressor in each case. From there, take the weighted average of these marginal contributions and account for the number of possible permutations across one, two, or three variables.

### 2.6.2 Shapley regressions

Shapley values can give information on how a feature contributes to the prediction of a model for a given observation without regard for the model's accuracy. That is, Shapley values by themselves cannot tell us how reliable a certain feature is in predicting the *correct* outcome. We employ *Shapley regressions* (Joseph, 2019). The general framework for how this is carried out is as follows: we regress the crisis indicator $y$ on the Shapley values $\Phi_{n\times k}$ using a logistic regression (Bluwstein et al., 2021). In other words, the unobservable predictions of the black box machine learning model get transformed using Shapley values into a parametric space that is additive. Once we are here, estimating the p-values is just a standard regression task. As we will see later, a desirable property of this framework is that if it is applied to a linear function, it will reproduce the results. The Shapley values used for these regressions are the mean of the Shapley for a specific observation across all the different splits in cross-validation.

# Chapter 3

# Banking Crises

## 3.1 JST Banking Crisis

### 3.1.1 Baseline Analysis

The first section of the results is based on the methods from Bluwstein et al., 2021 applied to an updated version of the JST database. The models in this work are compared by two performance metrics: the area under the receiver operator curve (AUROC) and the false alarm rate at a fixed hit rate of 80%. We can see from the ROC curve below that extremely randomized trees and random forests outperform logistic regression, immediately indicating that these machine learning methods have value in financial crisis prediction. From the AUROC, we can quantify how much better these models perform by comparing the false alarm rates.

Figure 3.1 confirms the immediate potential for machine learning models in creating early warning models for predicting banking crises. Under the AUROC metric, both extremely randomized trees and random forests outperform logistic regression. XGBoost performs slightly worse, with a single decision tree performing poorly. To quantify just how much better these models are performing, observe that, at a fixed hit rate

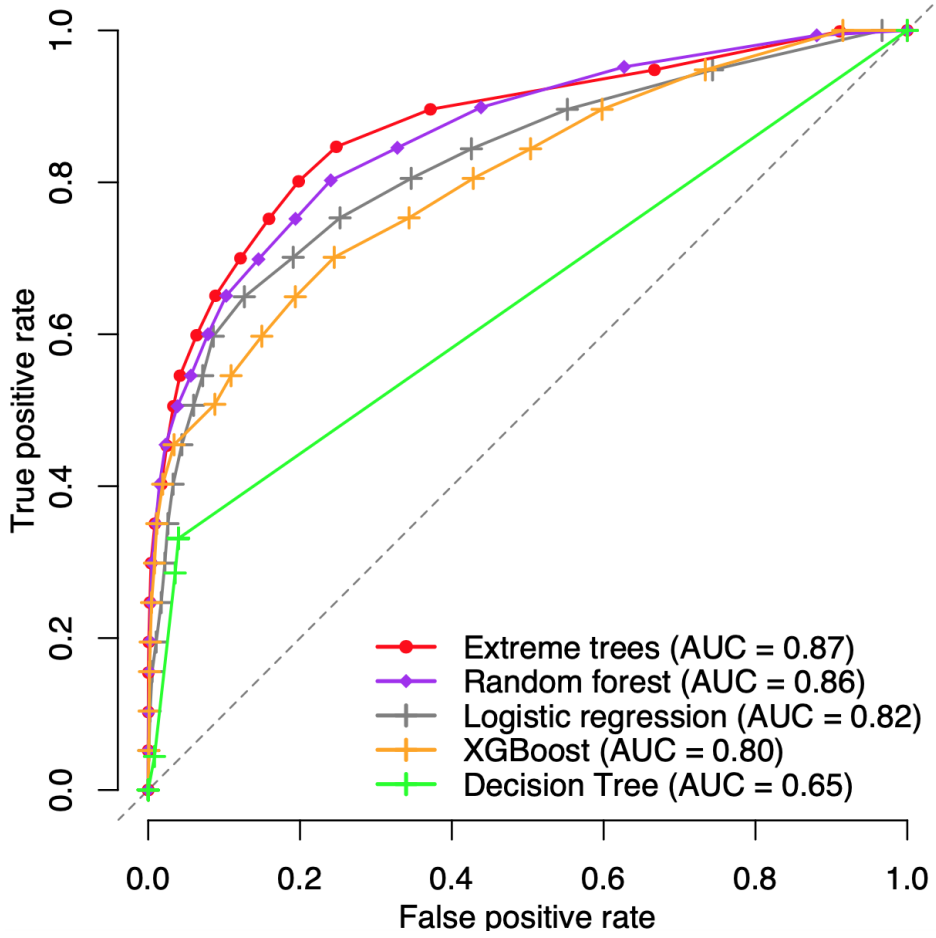


Figure 3.1: ROC curves for baseline models using the crisisJST as the outcome variable

of 80%, extremely randomized trees generate an 18.3% false alarm rate, random forest 26.3%, and logistic regression 36.2%. Compared to Bluwstein et al., 2021, extremely randomized trees obtain a marginally lower false alarm rate, whereas logistic regression produces a false alarm rate that is about 6% higher. However, the model rankings are consistent.

### 3.1.2 Robustness

Ensuring a model is robust is an important part of machine learning. A model is *robust* if it performs relatively the same on training data as new data (Li, 2018). To show that the ranking of the models is robust, we train many different models on different subsets of predictors or transformations of predictors while applying cross-validation to obtain stable results (Bluwstein et al., 2021). However, support vector machines and artificial neural networks were not tested for robustness due to the considerably long training time. Their performance in the baseline is also weaker than that of the tree-based algorithms. K-nearest neighbours generally had a weaker predictive performance as well, so the results of these models are presented in the appendix.

From the results reported in Bluwstein et al., 2021, we trained models including the house prices feature since they were able to obtain marginally better results when they included this variable. We also trained models using different filters on the variables. Both the Hamilton filter and the HP filter were used in an attempt to show that scaling the variables by the GDP was superior (Bluwstein et al., 2021). Lastly, we also investigated replacing the slope of the yield curve with the short and long-term nominal interest rates. The ranking of the models is consistent, and we observe that the best results are obtained using the slope of the yield curve.

| | AUROC | | | | |
|---|---|---|---|---|---|
| Experiment | extree | forest | xgboost | tree | logreg |
| Baseline | 0.87 | 0.86 | 0.80 | 0.65 | 0.82 |
| House prices | 0.87 | 0.86 | 0.81 | 0.63 | 0.82 |
| HP filter on GDP ratios | 0.87 | 0.86 | 0.83 | 0.61 | 0.80 |
| HP filter on nominal variables | 0.86 | 0.83 | 0.79 | 0.61 | 0.76 |
| HP filter on real variables | 0.84 | 0.84 | 0.79 | 0.64 | 0.77 |
| Hamilton filter on GDP ratios | 0.87 | 0.87 | 0.84 | 0.62 | 0.81 |
| Replace yield curve with nominal rates | 0.88 | 0.87 | 0.84 | 0.65 | 0.82 |
| Replace yield curve with real rates | 0.86 | 0.85 | 0.81 | 0.63 | 0.81 |

Table 3.1: AUROC scores for robustness check

### 3.1.3 Extremely Randomized Trees: the best predictive model

In an attempt to understand the results in more depth, we again follow the procedure in Bluwstein et al., 2021 and analyze the extremely randomized trees algorithm in more detail. Figure 3.2 shows correctly identified crises in dark green, correctly identified non-crises in light green, incorrectly predicted crises in grey, and incorrectly predicted non-crises in red. The black line in the figure shows the predicted probability for the observations in the sample. The model fully misses 6 crises across the 18 countries from 1882-2020.

The model misses the following crises: United States in 1893 and 1984, United Kingdom in 1890 and 1974, Japan in 1997, and Spain in 1977. For another 3 crises, the model correctly identifies 1 of the 2 years preceding a crisis. 5 of the 6 fully missed crises match the crises that are missed in Bluwstein et al., 2021.

The crisis in 1891 in the United Kingdom was known as the Barings crisis or the Panic of 1890 (Mitchener & Weidenmier, 2008). Some poor investments in Argentina led to the bankruptcy of the Barings Bank. The crisis in 1974 is known as the secondary banking crisis. During the 1960s, many secondary banks were trying to increase their market share in property and consumer loans. A policy was issued by the government in 1970, which was accompanied by a credit boom that led secondary banks to finance

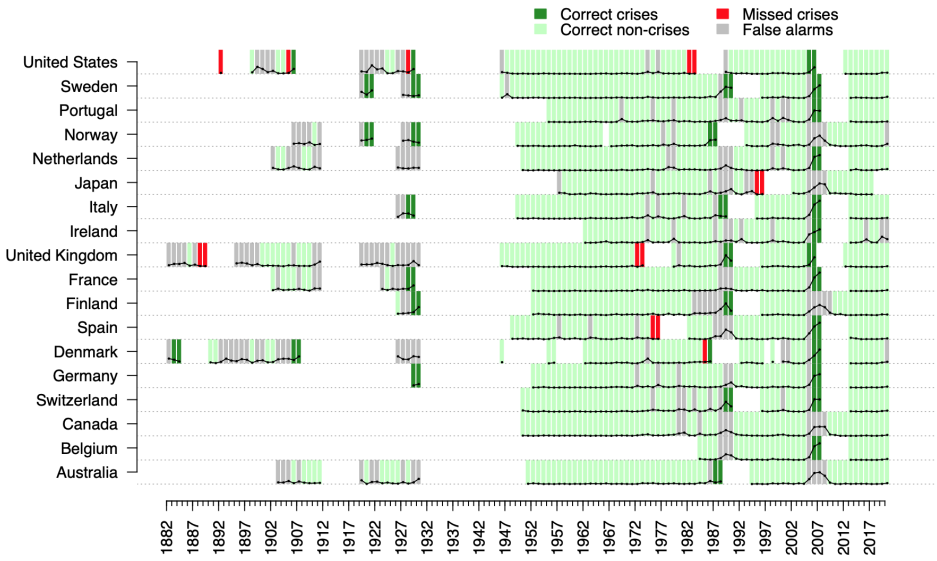


Figure 3.2: Crisis probability estimated by extremely randomized trees at a hit rate of 80%, using crisisJST as the outcome variable

long-term loans with short-term liquidity. As inflation rose in 1973, these smaller banks were unable to refinance their loans, leading to the banking crisis of 1974 (Reid, 1982).

The crisis in Spain in 1977 was supposedly caused by the government attempting to dampen the effects of the global oil price shock. This led to a delay in firm consolidation and increased public spending (Betran & Pons, 2013). Another explanation could be the institutional change in the financial sector because of the end of the Franco regime (Martin-Acena, 2014). All of these factors are not captured in the features used in this model, so missing this crisis is not surprising.

The 1984 crisis in the United States is commonly known as the savings and loan crisis, which was preceded by the global recession (Reinhart & Rogoff, 2009). There was a significant increase in the discount rate, which led to the interest rates of existing long-term loans being lower than the borrowing rates. This caused the saving and loan associations to make risky investments to remain profitable. The risks did not pan out, and a crisis occurred (Bluwstein et al., 2021). The 1893 crisis in the United States was one of the most severe financial crises since the switch to national banking (Jordà

et al., 2023). Over 500 national banks failed as a result of many railway companies failing and the price of silver dropping. These two events, taken together, caused a stock market crash in New York, leading to many withdrawals from banks, resulting in many failures. Unsurprisingly, this would be one of the missed crises since none of the features used in the model relate to the railway industry, the value of silver, or the stock market. Interestingly, in the work done by Bluwstein et al., 2021, which included a stock market variable, their model partially detected this crisis.

Lastly, the 1997 crisis in Japan was missed. This crisis is linked in part to the Asian financial crisis of 1997 (Bluwstein et al., 2021). However, as we can see from Figure 3.2, many 'false alarms' lead up to the missed crisis. Reinhart and Rogoff, 2009 identify this crisis as beginning in 1992 for Japan. Combining these false alarms with our knowledge about how difficult it is to date the beginning of a banking crisis can provide some understanding and comfort in missing this crisis.

Another thing to note about the model is that the number of false alarms is somewhat misleading. Recall that the model is calibrated to correctly predict 80% of the crises, so the model is risk-averse by construction. Being risk-averse will come with a higher false alarm rate. Part of the justification for this is that missing a crisis is often much more expensive than preparing for a crisis that does not occur (Bluwstein et al., 2021). Figure 3.2 also shows that many false alarms occur just before the 1-2 year window preceding a crisis. So, these false alarms still prove useful as early warning signals. We can also see that many false alarms occur in countries when other countries are experiencing a crisis in that same year. For example, there are false alarms in every country that did not experience a crisis during the global financial crisis. These false alarms indicate the elevated global risks and help justify the usefulness of the global variables. We also do not have any features in the dataset that account for any policy that may have been implemented to prevent a financial crisis. Because of this, the

model may predict false alarms in countries where they would have suffered a crisis in the absence of policy. Keeping all these things in mind means that the false alarms are not necessarily bad and may still be useful for policymakers (Bluwstein et al., 2021).

Looking at pre-World War 2 in Figure 3.2, we can see the proportion of false alarms is much higher than after the war. This can be in part due to how much the global economy has changed over time. Thus, a model trying to capture macroeconomic trends over 150 years will most likely not perform well in all sample periods. Since the number of observations was fewer pre-World War 2, this period also had a lower weight in the training of the model. Thus, the model will learn more about trends in recent observations and perform better on such observations. Also, in the computer age, the recorded data now includes much more detail. Thus, the quality of recent data is likely higher than that of older data (Bluwstein et al., 2021).

### 3.1.4 Forecasting

We employ a recursive forecasting experiment to investigate how early warning models are used in practice. All the results thus far have been based on cross-validation. Thus, all observations in the training set need to be from years earlier than those in the test set. For $1946 \leq t \leq 2020$, we use all observations up to year $t-2$ to train and then test them against observations from year $t$.

Under this construction, the training sets will have significantly different numbers of crises for different periods of time. An obvious example of this is after the global financial crisis. When the forecasting model is being trained for $t > 2009$, the proportion of crises in the training set will be much higher than when $t < 2008$. Since the AUROC is highly sensitive to this proportion problem, we resample all the training sets so that they contain an equal number of crisis and non-crisis observations (Bluwstein et al., 2021). We employed two types of sampling for this experiment and will only report on

the type of sampling that creates the best result. Upsampling involves resampling crisis observations to match the number of non-crisis observations, whereas downsampling involves resampling non-crisis observations to match the number of crisis observations.

Figure 3.3 shows the prediction chart for extremely randomized trees in the forecasting experiment. The model correctly identifies the global financial crisis and many of the crises that occurred in the 90s. Compared to Figure 3.2, the location of the missed crises is almost identical. The model fully misses 2 crises: United States in 1985 and Japan in 1997. The model partially misses 5 crises, and in most cases, it is the first indicator that is missed. These partially missed crises are: United States in 2007, Italy in 1989, United Kingdom in 1974, Spain in 1977, and Australia in 1989. This shows that across the different experiments, the models recognize the same indicators of crises.

In general, we see from these forecasting experiments that the identification of crises is often quite similar to the cross-validation experiment, but there are substantially more false alarms. In the cross-validation experiment, extremely randomized trees had a false alarm rate of 18.3%, compared to logistic regression's false alarm rate of 36.2%. Under the forecasting framework, these false alarm rates change to 38.3% and 36.1%, respectively. Extremely randomized trees suffer under the forecasting approach, but logistic regression is able to achieve a marginally better false alarm rate. Indeed, in terms of the false alarm rate, it appears that logistic regression is better than the machine learning models. However, extremely randomized trees obtain a higher AUROC score, meaning it would be more versatile in choosing hit rates other than 80%. Logistic regression obtains an AUROC of 0.78, which is lower compared to the 0.81 achieved by extremely randomized trees. Random forest sits in between with an AUROC of 0.80.

One thing to note about Figure 3.3 is that in every country in the sample, the model gives a false alarm in the year 2020. Of course, the COVID-19 pandemic began in 2020,

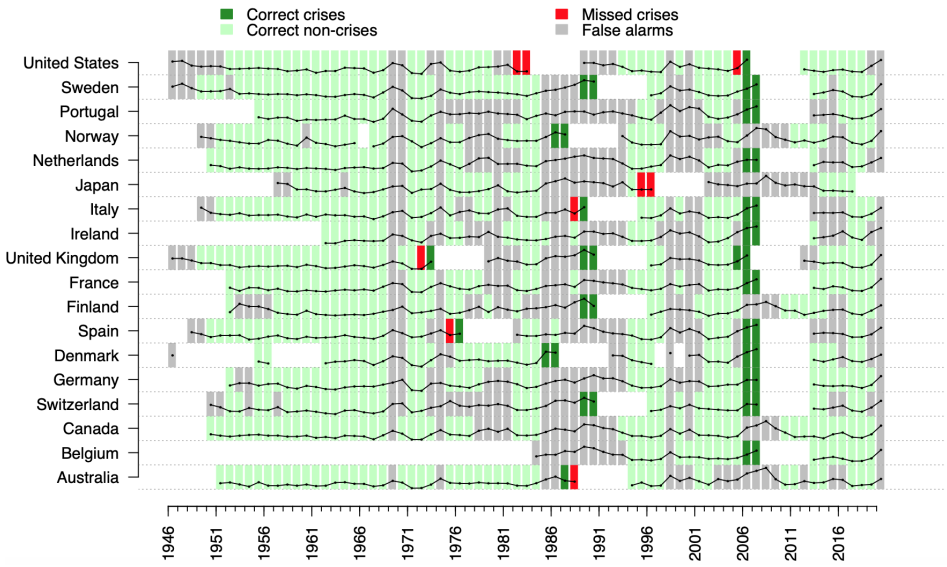


Figure 3.3: Forecasting crisis probability estimated by extremely randomized trees at a hit rate of 80%, using crisisJST as the outcome variable

and many of the model's features had introduced shocks. Where this gets interesting is combining this with the knowledge of the banking crises that occurred in 2023. This forecasting model provides a useful early warning indicator for a crisis that does not even exist in the database.

### 3.1.5 Shapley value decomposition: variable importance

We employ the same SHAP method as described in Bluwstein et al., 2021. We define the *predictive share* of a variable as the mean of that variable's absolute Shapley value across all observations. Figure 3.4 shows the normalized mean absolute Shapley values for all the predictors in the baseline analysis. By comparing Figure 3.4 to Table 2.4, we can immediately see the potential in using Shapley values as a way to determine the feature importance. The ordering of the mean absolute Shapley value is nearly identical to the order of logistic regression weights from Table 2.4. Since logistic regression is interpretable, we can gain confidence in the SHAP approach according to the weights

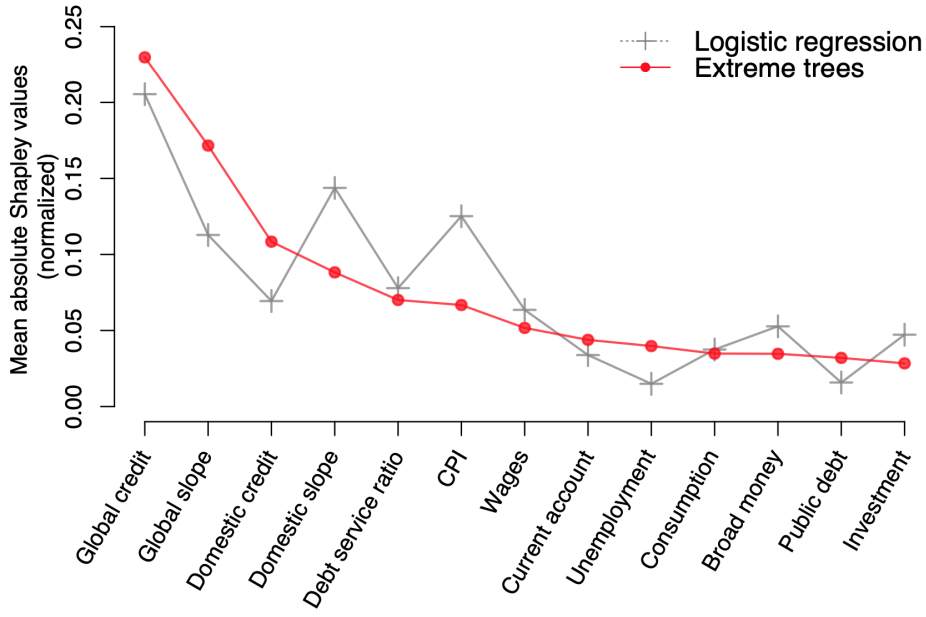


Figure 3.4: Mean absolute Shapley values of individual features for logistic regression and extremely randomized trees, using crisisJST as the outcome variable

given to each variable. If the SHAP method works, it should somewhat agree with the rankings given by logistic regression when applied to logistic regression.

In looking at the extremely randomized trees results, we get a similar ordering to the results from Bluwstein et al., 2021, with both of the global features having the highest predictive share. However, the global credit feature outranks the global slope of the yield curve using the updated JST database. The new features to the database prove themselves to be of moderate importance, with the wages and unemployment rate features ranking 7th and 9th, respectively. These rankings are relatively robust across all the models, yet again showing the importance of the global variables.

To show the value of being able to decompose predictions into their Shapley values for interpretation, Figure 3.5 shows the decomposition of each prediction for each observation over time for several countries. The red and blue parts of the bars represent the domestic and global slope of the yield curve. The yellow and green parts of the bar

represent the domestic and global credit growth. The black dot is the predicted probability of a crisis in that year, and the red vertical bars represent the years leading up to a crisis. The black horizontal line shows the threshold for an 80% hit rate. In keeping with the analysis done thus far, this figure is based on the predictions of extremely randomized trees. Unsurprisingly, the performance of the model varies across different countries. We see a general trend of noisier predictions before World War 2. The expected results from Bluwstein et al., 2021 hold, as we see, the slope of the yield curve dominating many of the pre-World War 2 predictions, whereas the predictions during the global financial crisis are primarily driven by global credit. We also see domestic credit growth plays a much more substantial role during the global financial crisis in Denmark than in the United States or Sweden. Using this framework to investigate which features drive which predictions is a further justification for using machine learning models for financial crisis prediction. This approach makes great strides toward uncovering the black-box nature of these models.

### 3.1.6 Shapley regressions: variable significance

The previous section investigated variable importance. We now move on to investigating the statistical significance of the features in the models. To do this, we regress the predicted crisis probability on the calculated Shapley values. This works as an additive feature transformation (Bluwstein et al., 2021). Table 3.2 shows the results of this transformation for the extremely randomized trees model. The *share* column contains the normalized mean absolute Shapley values, obtained from Figure 3.4. The coefficients are the result of a one standard deviation change of the Shapley values on the predicted log-odds of crisis ($\log \frac{\hat{y}}{1-\hat{y}}$) (Bluwstein et al., 2021). The sign of the coefficients does not show the association between the predicted probability and the predictor. That is captured in the *direction* column. The interpretation of the coefficient column is that

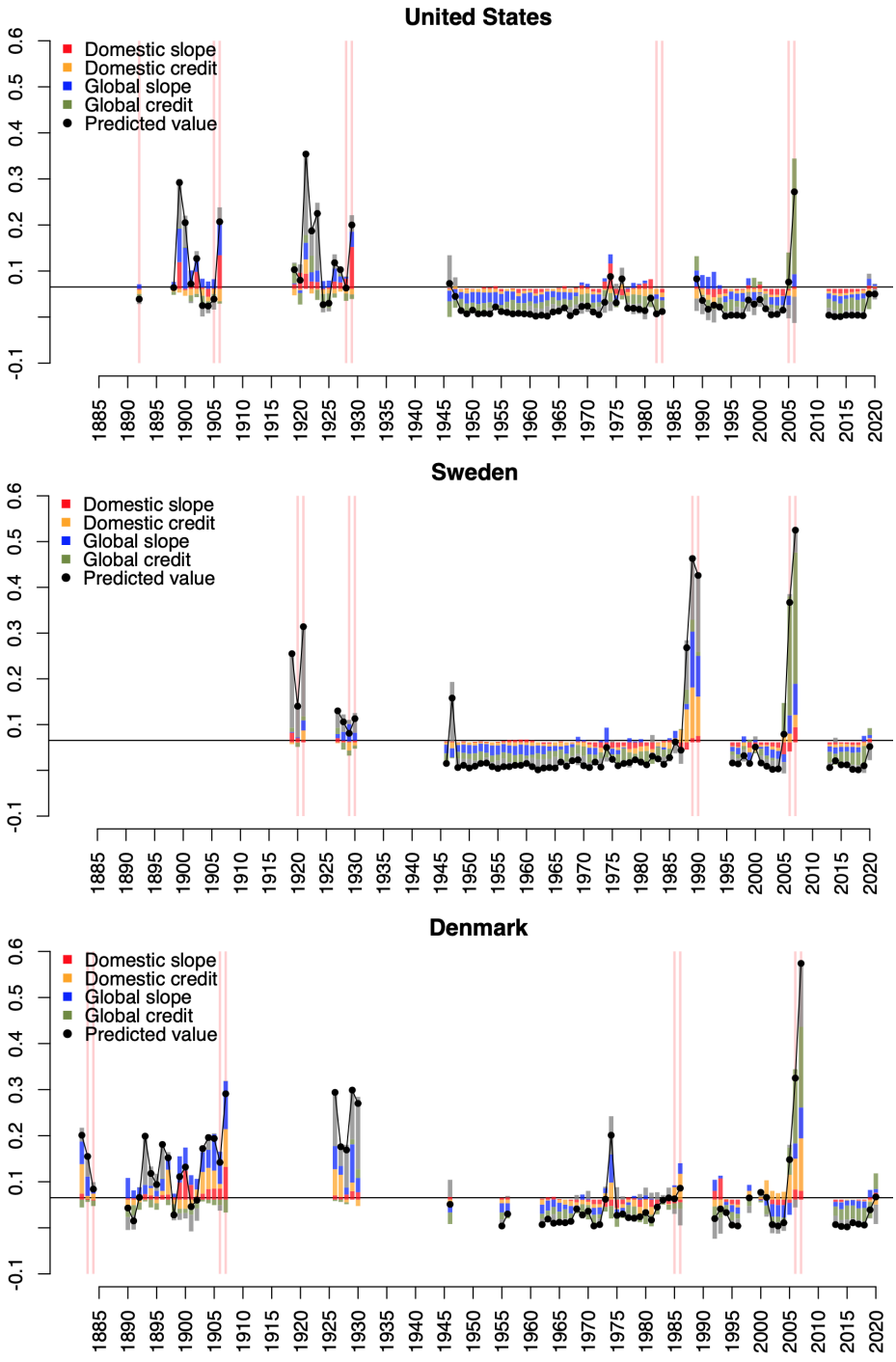


Figure 3.5: Shapley values as a function of time using crisisJST as the outcome variable

the higher the number, the greater the significance of the variable. If the model made little use of a variable, then the coefficient is statistically insignificant and is represented by a negative number in the table.

| Variable | Direction | Share | Coefficient | p |
|---|---|---|---|---|
| Global credit | + | 0.230 | 0.4750 | 0.000 |
| Global slope | - | 0.172 | 0.4325 | 0.015 |
| Domestic credit | + | 0.108 | 0.3546 | 0.000 |
| Domestic slope | - | 0.088 | 0.3467 | 0.000 |
| Debt service ratio | + | 0.070 | 0.1280 | 0.157 |
| CPI | - | 0.067 | 0.2438 | 0.012 |
| Wages | + | 0.052 | 0.0406 | 0.353 |
| Current account | - | 0.044 | 0.0642 | 0.327 |
| Unemployment | + | 0.040 | 0.1152 | 0.170 |
| Consumption | - | 0.035 | 0.1426 | 0.089 |
| Broad money | - | 0.035 | -0.2236 | 1.000 |
| Public debt | + | 0.032 | 0.0045 | 0.486 |
| Investment | + | 0.028 | -0.0545 | 1.000 |

Table 3.2: Results of the Shapley regression. Shows the direction of the relationship between predictor and crisis, coefficients, p-value against the null hypothesis, and the predictive share. crisisJST used as the outcome variable.

Table 3.2 shows results that are consistent with the previous section. Both global variables return the highest coefficients with low p-values. Consumption also seems to be significant, with a p-value of 0.089. This means that even though consumption has a relatively small predictive share, the value of consumption is significantly aligned with the indicators. On the other hand, the debt service ratio, wages, and the current account have moderate predictive shares, but the signals they provide are difficult for the model to distinguish from the null. That is, there is not a clear alignment between these variables and the actual crises. Broad money and investment both obtain p-values of 1 and negative coefficients. This means that the model made very little use of the information given by these variables and that their signals were not able to be recognized against the null hypothesis.

### 3.1.7 Nonlinearities

The benefits of using the SHAP method do not end here. We can investigate the nonlinear relationship between the predictor values and their Shapley values. These nonlinearities are unique to the machine learning methods. If we treat the Shapley values as the weights in a machine learning model, then these nonlinearities act as a weight function based on the predictor value. Figure 3.6 shows the Shapley values of important predictors as a function of the predictor value. Crisis observations are represented by the red dots, and non-crisis observations are the black dots. Shapley values that are greater than 0 represent an increase in the predicted probability for that observation (Bluwstein et al., 2021). The top 4 predictors in terms of Shapley share are represented in Figure 3.6.

For each predictor, we fit a linear and cubic regression and compare the goodness-of-fit in terms of $R^2$. We can see that for the key predictors, the $R^2$ is substantially higher for the cubic polynomial line. From an economic view, these relationships are also not surprising. Figure 3.6 shows that a flattening or inversion of the yield curve, either domestically or globally, results in higher Shapley values, which means an increase in predicted probability. Similarly, severe spikes in credit growth, either domestically or globally, result in higher Shapley values. This relationship also makes sense the other way around. When the yield curve is upwardly sloping, or if credit growth is relatively muted, then these features have little effect on the predicted probability of a crisis. What we can take away from Figure 3.6 is that these financial systems are much more susceptible to crises if variables are in the far tails of their distributions (Bluwstein et al., 2021). This is most easily seen in the global credit plot of Figure 3.6. When global credit growth is above 10%, we see many clustered crisis observations.

We now ask the question of how the awareness of these nonlinearities affects the predictions of the features. We will employ a simple test to judge whether the models

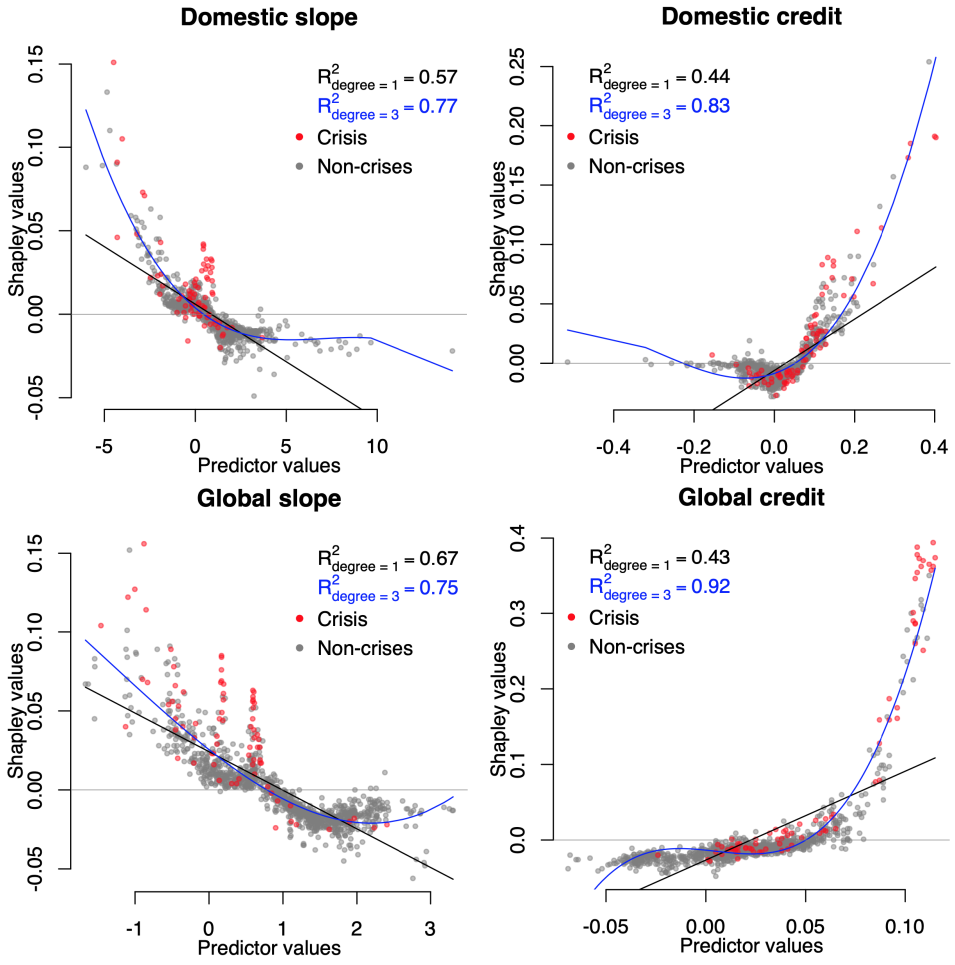


Figure 3.6: Shapley values of key predictors as functions of the predictor values. JST crisis observations shown in red.

can use these nonlinear relationships. We train linear models for each of the predictors on their own. The first model is trained on the actual predictor values, and the second is trained on the Shapley values extracted from the extremely randomized trees model. Table 3.3 shows the comparison of the AUROC for the univariate models under the baseline approach.

| Feature | Linear | Shapley |
|---|---|---|
| Domestic slope | 0.71 | 0.77 |
| CPI | 0.56 | 0.69 |
| Broad money | 0.59 | 0.47 |
| Unemployment | 0.54 | 0.51 |
| Wages | 0.58 | 0.68 |
| Consumption | 0.51 | 0.56 |
| Public debt | 0.55 | 0.55 |
| Investment | 0.60 | 0.55 |
| Current account | 0.61 | 0.61 |
| Domestic credit | 0.69 | 0.63 |
| Debt service ratio | 0.67 | 0.68 |
| Global credit | 0.74 | 0.78 |
| Global slope | 0.74 | 0.82 |

Table 3.3: AUROC of univariate linear regressions compared to the crisisJST variable being regressed on the individual predictor Shapley values.

We immediately see from Table 3.3 that regressing the crisis outcome on the Shapley values improves the AUROC score for most of the key predictors. The AUROC scores are significantly better for the domestic and global slope of the yield curve, the global credit growth, and the CPI. The only key predictor for this type of financial crisis that does not have its univariate AUROC improved is the domestic credit, which is in line with the results from Bluwstein et al., 2021. The two new variables added in the 6th version of the JST database, the unemployment rate and the wages, have differing results in this experiment. There is a 0.11 improvement for the wages variable when it is regressed on its Shapley values and a 0.04 loss when the unemployment rate is regressed on its Shapley values. These univariate regressions are not particularly useful

on their own, but when we combine the results from Table 3.3 and Figure 3.6, we can see that the nonlinear relationships found in these predictors are most likely important and useful for predicting financial crises.

## 3.2 Reinhart and Rogoff Banking Crisis

### 3.2.1 Baseline Analysis

We now attempt to answer the following question: can machine learning be beneficial in predicting other types of financial crises? We use the banking crisis narrative list from Reinhart and Rogoff, 2009. Figure 3.7 shows a generally worse AUROC performance across all models compared to the JST crisis case, but extremely randomized trees and random forests still outperform logistic regression. The false alarm rates at an 80% hit rate for extremely randomized trees, random forest, and logistic regression are 22.7%, 28.2%, and 35.3%, respectively. Extremely randomized trees produces the lowest false alarm rate of all models.

### 3.2.2 Robustness

Using the same approach as for the JST crisis, the same transformations were used to test the robustness of the model. We see that the model ranking is indeed robust across these different experiments, with extremely randomized trees consistently performing the best, followed by random forests. Under the Reinhart and Rogoff definition of a crisis, XGBoost performs similarly to logistic regression. Table 3.4 shows these rankings over the different experiments. We can note that replacing the yield curve with the nominal short and long-term interest rates improves the performance of extremely randomized trees by 1 percentage point. However, the analysis will be carried out on the baseline model for comparability throughout this work.

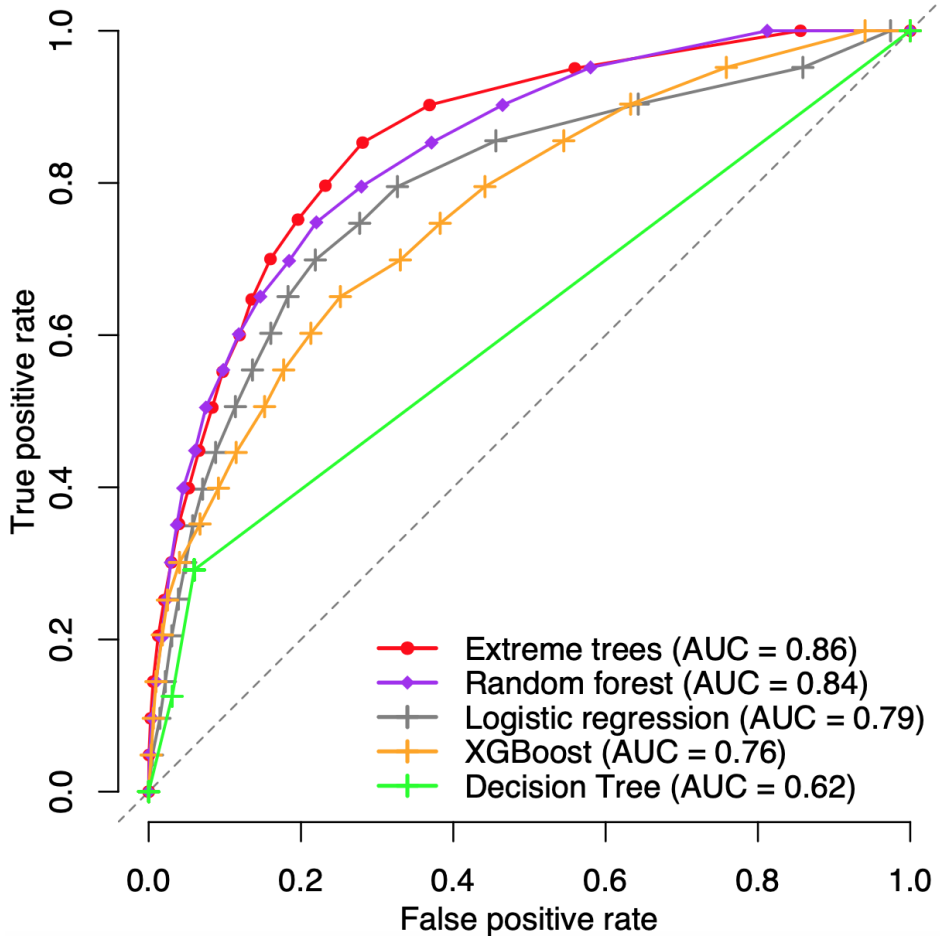


Figure 3.7: ROC curves for baseline models, crisisBanking used as the outcome variable

| | AUROC | | | | |
|---|---|---|---|---|---|
| Experiment | extree | forest | xgboost | tree | logreg |
| Baseline | 0.86 | 0.84 | 0.72 | 0.62 | 0.79 |
| House prices | 0.84 | 0.82 | 0.75 | 0.63 | 0.78 |
| HP filter on GDP ratios | 0.86 | 0.83 | 0.80 | 0.58 | 0.76 |
| HP filter on nominal variables | 0.85 | 0.82 | 0.79 | 0.60 | 0.76 |
| HP filter on real variables | 0.84 | 0.83 | 0.77 | 0.61 | 0.76 |
| Hamilton filter on GDP ratio | 0.86 | 0.85 | 0.77 | 0.61 | 0.77 |
| Replace yield curve with nominal rates | 0.87 | 0.84 | 0.76 | 0.64 | 0.80 |
| Replace yield curve with real rates | 0.85 | 0.83 | 0.79 | 0.56 | 0.80 |

Table 3.4: AUROC scores for robustness check, crisisBanking used as the outcome variable

### 3.2.3 The best predictive model

We have a similar figure for analysis as in the previous section. The model fully misses 4 crises while partly missing 9 crises. The 4 fully missed crises are 4 of the same fully missed crises when using the JST definition of a banking crisis, namely: United States in 1893 and 1984, United Kingdom in 1974, and Spain in 1977. The 9 partially missed crises are: United States in 2007, Norway in 1988, Japan in 1992, United Kingdom in 1890 and 1985, France in 1995, Denmark in 1987, Germany in 1977, and Canada in 1984. It is interesting to note that the same crises are being missed despite changing the definition of a crisis. The partially missed crises are split almost evenly, depending on whether the first indicator (2 years prior to a crisis) or the second indicator (1 year prior to a crisis) was missed. In 4 of the partially missed crises, the second indicator was detected, while for the other 5, the first indicator was detected. Similarly to the results of using the JST definition of a crisis, many of the false alarms occur before World War 2 or during a time when many countries are experiencing crises.

The model missing the first indicator year of a crisis means that the conditions two years before the crisis were not as severe as the year directly preceding the crisis. This could be problematic because 1 year may not be enough time to implement effective

policy to help mitigate the crisis. This model misses the first indicator year of the global financial crisis in the United States, but this can be attributed to the dating of the crisis by Reinhart and Rogoff. According to their dating, the United States experienced the global financial crisis one year before all the other countries in the sample. Since the models are trained with some global variables, it is not surprising that the second indicator year is the one that is identified. The model also misses the first indicator of the crisis in Norway in 1988. The 1980s involved an international trend of liberalizing the entire banking system so that there was more competition between financial institutions. Norway was no stranger to this trend. Real estate and stock market prices increased at least twofold, leading to the three largest Norwegian banks being supported by the government (Jordà et al., 2023). Since our model does not contain house price data or stock market data, it makes sense that this crisis would be partially missed. The Denmark banking crisis in 1987 also had the first indicator missed. Due to similar international trends as the Norway crisis, Denmark's banking was deregulated in the 1980s as well. Asset prices also rose, and several major banks required government assistance (Jordà et al., 2023). The last crisis, in which the first indicator was missing, was the crisis in Germany in 1977. This crisis is not included in the JST database and is the first of its kind in this analysis. Caprio and Klingebiel, 2003 classified this crisis as non-systemic and described it as a period when "so-called giroinstitutions faced problems."

The model partially missed the second indicator for 5 crises. Missing the second indicator is somewhat less problematic than missing the first indicator in an early warning approach. Missing the second indicator could be due to effective policy having been implemented to prevent the crisis. In the absence of that, correctly predicting the first indicator should give policymakers ample time to design and implement effective policy. The first indicator was detected for the Japanese banking crisis in the 1990s.

During this period, Japan's real estate and stock prices were sharply falling, resulting in Sanyo Securities collapsing. Japan implemented two rounds of capital injections, in which many major banks received support from the government (Jordà et al., 2023). We can see from Figure 3.8 that Reinhart and Rogoff actually classify all the years from 1992 until 2001 as crisis years in Japan. They did not consider the crisis to be systemic until 1997, which then lines up with the dating system of the JST database. The United Kingdom in 1890 and 1984 experienced crises that were partially detected by the model. The 1890 crisis is commonly referred to as the Baring crisis, which was already mentioned in the previous section. It is difficult to accurately try and justify why this crisis would be missed due to the very high proportion of false alarms during that time period. Reinhart and Rogoff classify the 1984 crisis as a non-systemic crisis and label it as a crisis because the "Johnson Matthey Bankers failed" (Reinhart & Rogoff, 2009). This crisis is an example of the model trying to capture a crisis that had somewhat successful policy implemented to mitigate. The Johnson Matthey bank was going to fail and was bailed out by the Bank of England ("Bank failures case study: - Johnson Matthey", n.d.). The crisis in France in 1994 is described by Reinhart and Rogoff as a period when Credit Lyonnaise, a major French bank, experienced severe solvency problems. Again, this crisis is not classified as systemic by Reinhart and Rogoff, 2009. Lastly, and perhaps surprisingly, Reinhart and Rogoff date a crisis in Canada in 1983, which the model partially detects. During this period, 15 Canadian Deposit Insurance Corporation members, including 2 major banks, failed. This crisis was also classified as not systemic.

All 5 of the crises that had the first indicator detected were non-systemic, according to Reinhart and Rogoff. Reinhart and Rogoff, 2009 makes the distinction between systemic and non-systemic banking crises. Essentially, a systemic banking crisis is a more severe event. In light of this, all the partially missed crises are classified as less

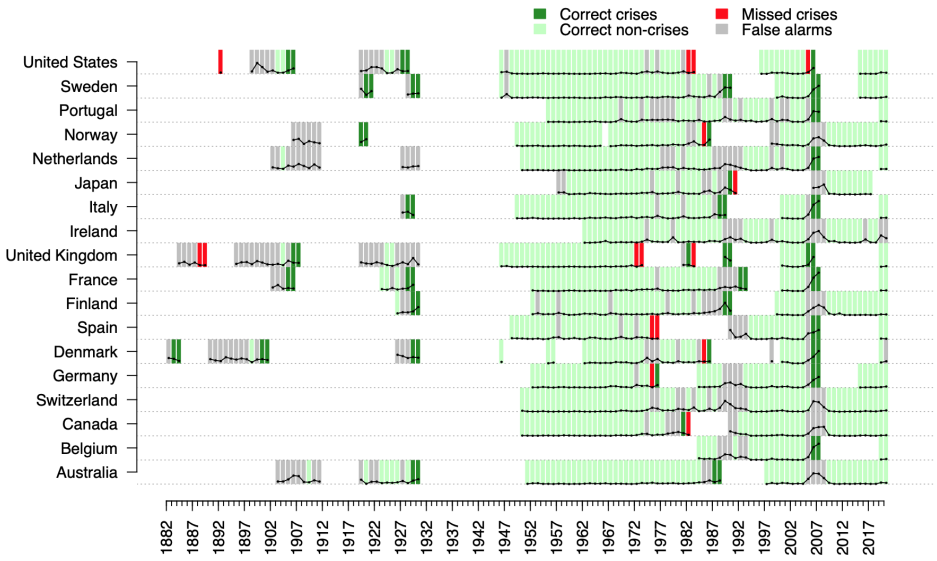


Figure 3.8: Crisis probability estimated by extremely randomized trees at a hit rate of 80%, crisisBanking used as the outcome variable

severe. This would make these crises harder for the model to detect since if the change in features is less severe, the predicted probability is less severe, which means the model is less likely to classify that observation as a crisis. We also see that the majority of false alarms happen before World War 2, or during the banking crises of the 1990s. This is in line with the analysis of the other banking crisis.

### 3.2.4 Forecasting

We again shift our focus now to investigating the traditional way that these early warning models are used in practice. We can see by comparing Figures 3.8 and 3.9 that the same result holds as when the JST crisis definition was used. The crises that are missed or partially missed in the forecasting experiment almost match perfectly with those missed in the cross-validation experiment. The forecasting model fully misses 3 crises and partially misses 3 more crises. The 3 missed crises are: the United Kingdom in 1974 and 1985 and Germany in 1977. The partially missed crisis are: the United

States in 1985 and 2007 and Canada in 1983. There is a period during the mid 1980s until the early 1990s where almost every prediction in the model is a positive indicator prediction. This is in contrast to the model mostly predicting correct non-crises during the period of the 1950s to the 1970s. What we see here is perhaps an elevated level of global stress that the model is picking up on. Indeed, we see many of the false alarms occurring in years when one or more countries are on the verge of experiencing a crisis. An example of this is the forecasting model correctly identifying the global financial crisis. Every prediction during the years of 2006 and 2007 is either a true positive or a false positive.

In terms of the false alarm rates for comparing the different types of models, we see that extremely randomized trees do perform best using this definition of a banking crisis. Extremely randomized trees, random forest, and logistic regression produce false alarm rates of 38.6%, 44.8%, and 43.6%, respectively. In comparison to the forecasting results when using the JST definition of a crisis, extremely randomized trees produce an almost identical false alarm rate, whereas random forest and logistic regression suffer. This shows that the extremely randomized trees model is more robust to the definition of a banking crisis that is used. In terms of the AUROC scores, extremely randomized trees achieve the highest score of 0.79, beating the scores of 0.77 and 0.75 achieved by random forest and logistic regression, respectively.

### 3.2.5 Shapley value decomposition: variable importance

We next investigate if there is any change in the importance of variables when we redefine the crisis variable. Figure 3.10 shows the ranking of features in terms of the mean absolute Shapley values while using the Reinhart and Rogoff definition of a banking crisis. This ordering more closely matches that from Bluwstein et al., 2021, with the global slope of the yield curve proving itself to be the most important feature, followed

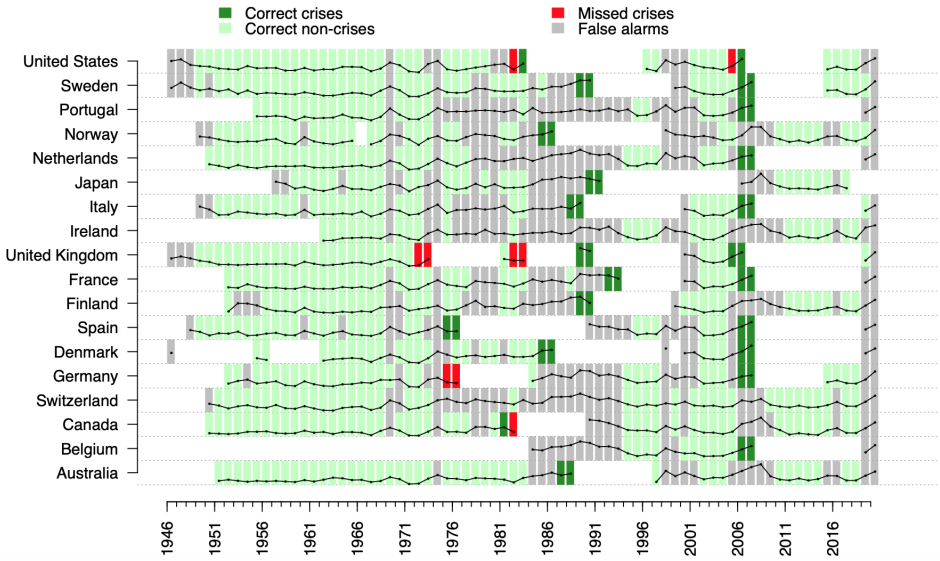


Figure 3.9: Forecasting crisis probability estimated by extremely randomized trees at a hit rate of 80%, crisisBanking used as the outcome variable

by the global credit growth. The unemployment rate plays a less significant role under this definition of a crisis, dropping from 9th to 10th. However, the wage feature shows itself to be more important, ranking 6th rather than 7th. We see more agreement in the ranking of the features compared to the JST crisis case. Both logistic regression and extremely randomized trees agree that the pre-crisis period of the Reinhart and Rogoff definition of a banking crisis is dominated by the global credit growth and slope of the yield curve.

Again, we can decompose the predictions into their Shapley values over time. Figure 3.11 shows the same trends as in the JST crisis case. The slope of the yield curve plays a very important role in the early part of the sample, whereas the identification of the global financial crisis is driven by global credit growth. Looking closer at the plot for the United States, we can note that the domestic slope of the yield curve also seems to play a dominant role during the 1920s, which is not seen in Sweden and Denmark.

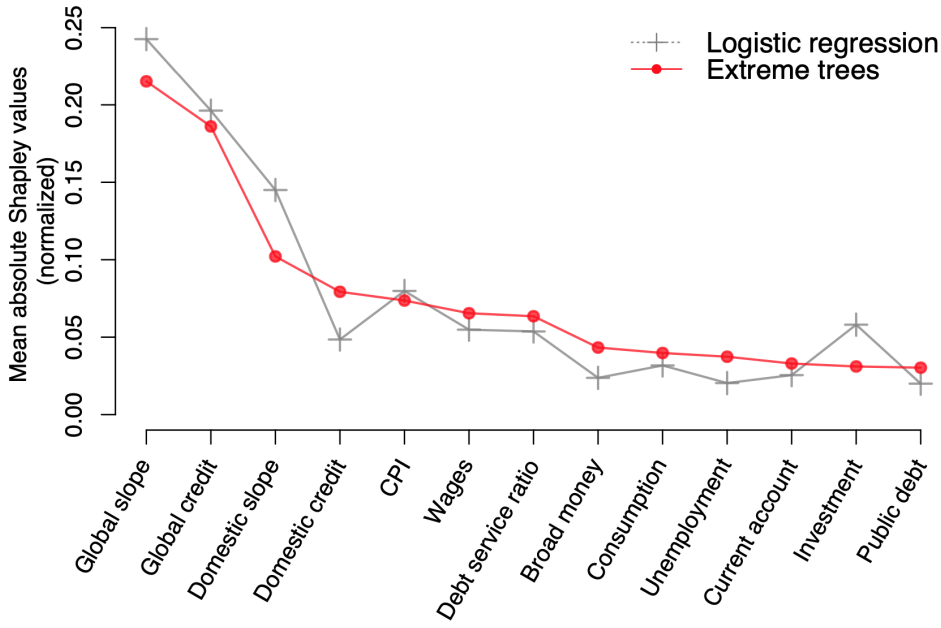


Figure 3.10: Mean absolute Shapley values of individual features for logistic regression and extremely randomized trees, crisisBanking used as the outcome variable

### 3.2.6 Shapley regressions: variable significance

Table 3.5 shows broad money has a low p-value. This means that the value of the broad money predictor is well aligned with the crisis indicators. Contrarily, although domestic credit growth has the 4th highest predictive share, it has a p-value of 0.403, meaning that the signals it is issuing cannot be distinguished from the noise. The same can be said for consumption, the current account, and public debt, although their predictive shares are all much lower than domestic credit. We also see that the unemployment rate and investment have p-values of 1 and negative coefficients. This suggests that the model made little to no use of the information from these variables and that the signals issued were not aligned with the crisis indicator. The three features with the highest predictive share also happen to have the largest coefficients and smallest p-values in this experiment. Table 3.5 shows just how aligned global credit growth and the slope

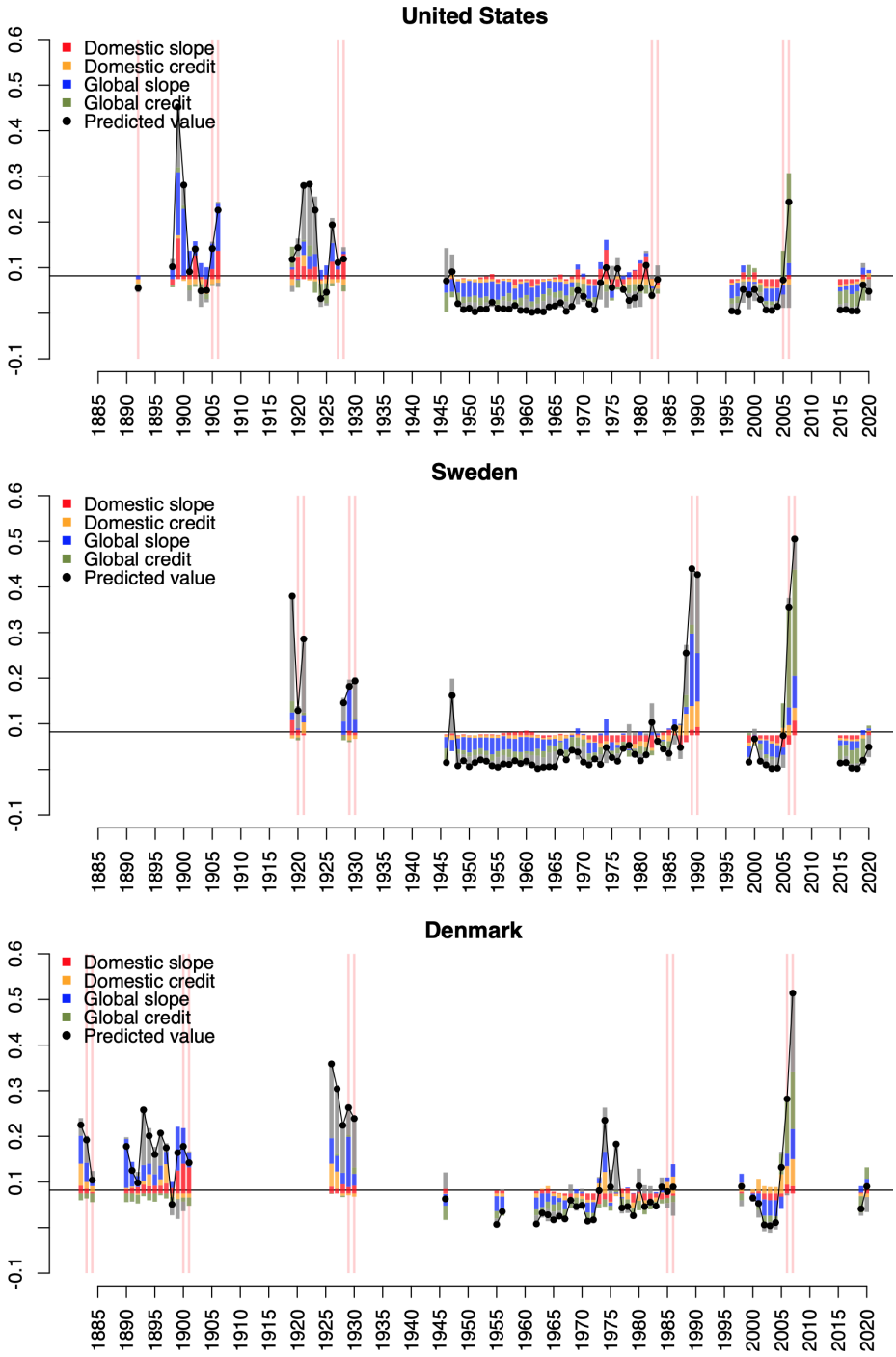


Figure 3.11: Shapley values as a function of time using crisisBanking as the outcome variable

of the yield curve, both domestically and globally, are to the crisis variable.

| Variable | Direction | Share | Coefficient | p |
|---|---|---|---|---|
| Global slope | - | 0.215 | 0.3982 | 0.003 |
| Global credit | + | 0.186 | 0.4080 | 0.000 |
| Domestic slope | - | 0.102 | 0.2991 | 0.008 |
| Domestic credit | + | 0.079 | 0.0259 | 0.403 |
| CPI | - | 0.074 | 0.2770 | 0.018 |
| Wages | + | 0.065 | 0.1364 | 0.114 |
| Debt service ratio | + | 0.063 | 0.1036 | 0.152 |
| Broad money | - | 0.043 | 0.2127 | 0.027 |
| Consumption | - | 0.040 | 0.0237 | 0.426 |
| Unemployment | + | 0.037 | -0.0270 | 1.000 |
| Current account | - | 0.033 | 0.0109 | 0.469 |
| Investment | + | 0.031 | -0.0887 | 1.000 |
| Public debt | - | 0.030 | 0.0594 | 0.332 |

Table 3.5: Shapley regression. Shows the direction of the relationship between predictor and crisis, coefficients, p-value against the null hypothesis, and the predictive share. crisisBanking used as the outcome variable.

### 3.2.7 Nonlinearities

Figure 3.12 shows similar results to Figure 3.6. The same 4 predictors are shown since they are the top 4 for both banking crisis definitions. The same results hold as in the JST crisis case. We see that a cubic polynomial fits the relationship for each predictor much better than the linear fit. The biggest improvement in goodness-of-fit is with the global credit growth, where we see the clustered crisis observations again when the credit growth rises above 10%. We see that inverted or flattening yield curve slopes result in higher Shapley values, as does a rise in credit growth. The spike in Shapley values when credit growth is in the tail of its distribution is in line with the description from Aliber and Kindleberger, 2015, who describe a financial crisis as "credit booms gone wrong". Indeed, across the 4 key predictors, the predicted probability of a crisis increases when these variables are in the tails of their distributions.

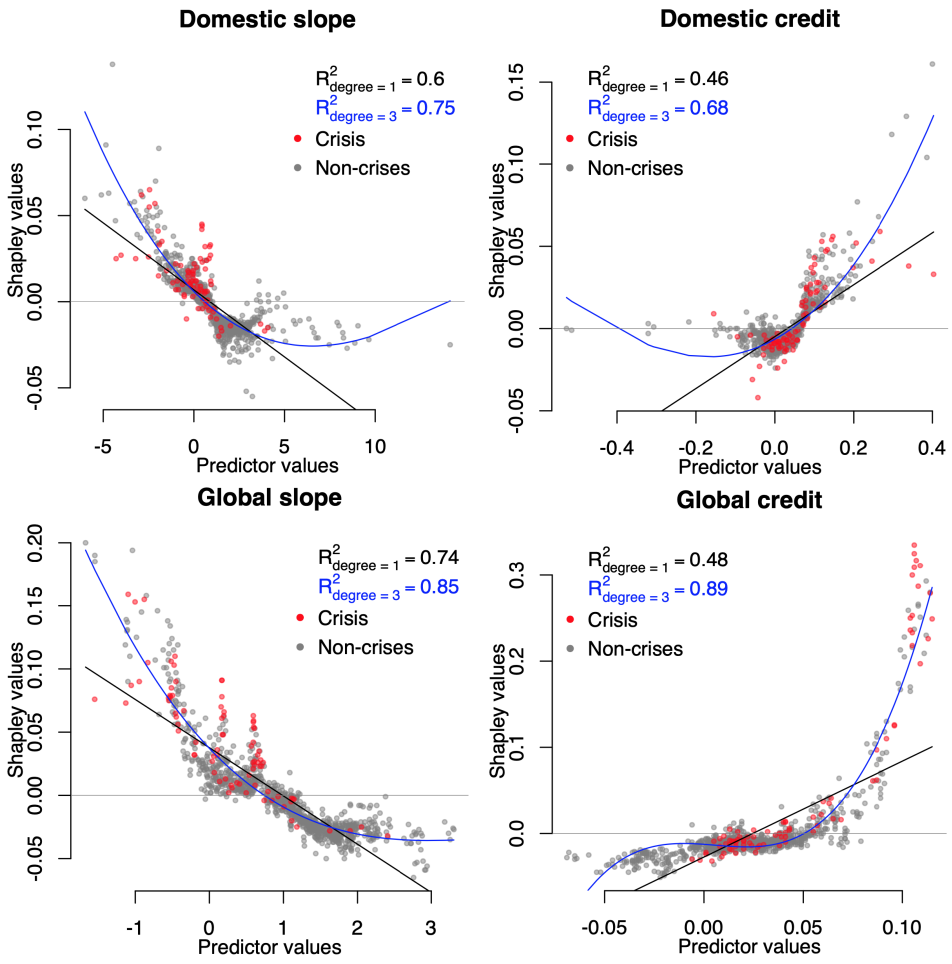


Figure 3.12: Shapley values of key predictors as functions of the predictor values. Banking crisis observations shown in red.

| Feature | Linear | Shapley |
|---|---|---|
| Domestic slope | 0.71 | 0.75 |
| CPI | 0.55 | 0.70 |
| Broad money | 0.59 | 0.62 |
| Unemployment | 0.52 | 0.52 |
| Wages | 0.59 | 0.69 |
| Consumption | 0.53 | 0.52 |
| Public debt | 0.48 | 0.52 |
| Investment | 0.56 | 0.50 |
| Current account | 0.57 | 0.56 |
| Domestic credit | 0.63 | 0.57 |
| Debt service ratio | 0.62 | 0.64 |
| Global credit | 0.67 | 0.70 |
| Global slope | 0.76 | 0.81 |

Table 3.6: AUROC of univariate linear regressions compared to the crisisBanking variable being regressed on the individual predictor Shapley values.

Table 3.6 shows the results of the univariate regressions on the predictor values and then repeats the univariate regressions on the Shapley values. The same general results hold as in the JST crisis case. The AUROC is usually much better for these regressions when the Shapley values are used, particularly for the important features. The wages feature receives a similar improvement in AUROC compared to the JST case, and the unemployment rate performs equally well rather than worse when regressed on the Shapley values. 8 of the 13 predictors obtain better AUROC scores when regressed on the Shapley values. The biggest loss in AUROC is for investment, which dropped 6 percentage points. However, we know from Figure 3.10 that investment is nowhere near being an important predictor. Most losses are 1 or 2 percentage points, whereas the improvements can be as high as almost 15 percentage points. Again, combining these results with those from Figure 3.12 confirms the importance of recognizing the nonlinear relationships for predicting financial crises.

# Chapter 4

# Currency and Inflation Crises

We now move on to our investigation of machine learning models' potential usefulness in predicting different types of financial crises. This first section focuses on the currency crisis definition given by Reinhart and Rogoff: An observation is classified as a currency crisis if the country's currency has "an annual depreciation versus the U.S. dollar of 15 percent or more" (Reinhart & Rogoff, 2009).

## 4.1 Currency Crisis

### 4.1.1 Baseline Analysis

Model performance according to Figure 4.1 is comparable to the performance of the models when investigating banking crises. Something that is of immediate interest is that while extremely randomized trees and random forests maintain their performance, the performance of logistic regression suffers. The only change given to these models is where the 0s and 1s are placed in the output variable. Instead of the machine learning models outperforming logistic regression by just a few percentage points, they now are performing over 20 percentage points better! Considering that our models are

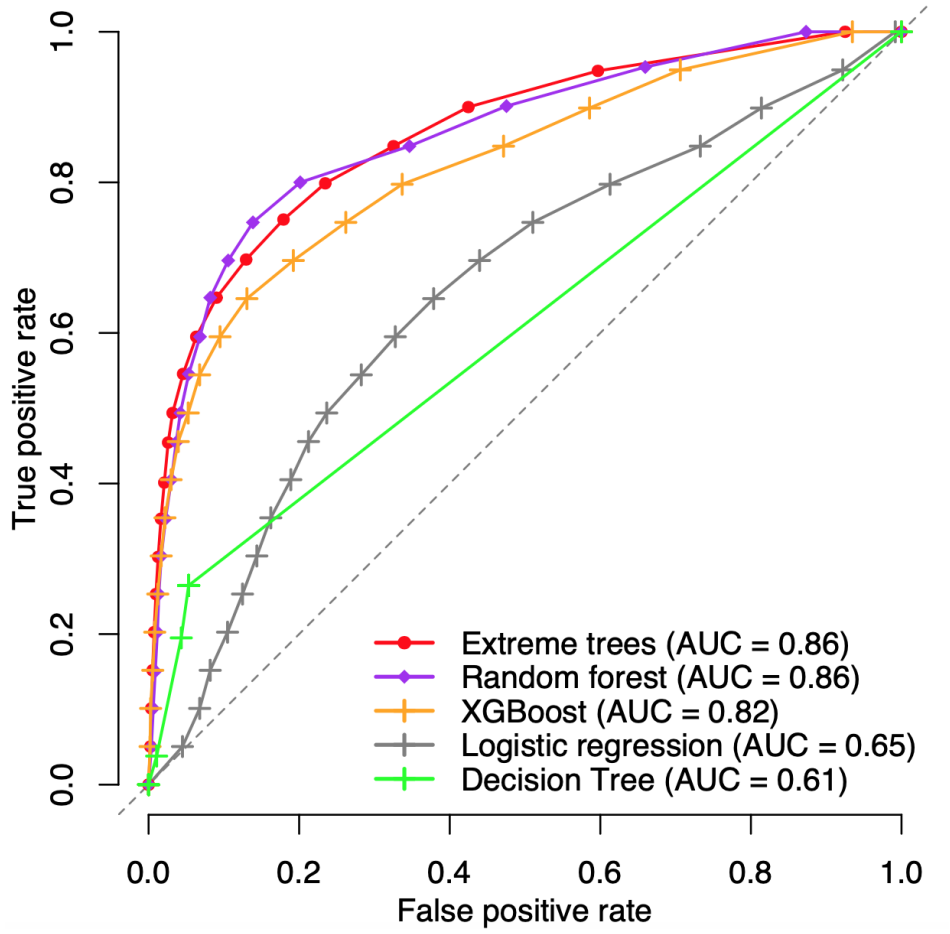


Figure 4.1: ROC curves for baseline models, crisisCurrency used as the outcome variable

constructed to have an 80% true positive rate, Figure 4.1 shows us that random forest will produce a lower false alarm rate than extremely randomized trees, despite having the same AUROC score. Extremely randomized trees and logistic regression produce false alarm rates of 20.6% and 53.7%, respectively. Random forest beats both of these with a false alarm rate of 17.4%. We also see XGBoost performing 17 percentage points better than logistic regression. This confirms our hypothesis that XGBoost is a suitable model for financial crisis prediction.

### 4.1.2 Robustness

We can see from Table 4.1 that the performance gap between logistic regression and the tree-based methods is robust. Across different experiments, extremely randomized trees and random forests consistently perform much better than logistic regression. We see most of the robustness checks providing better AUROC scores. However, for consistency and comparability throughout the paper, we keep our analysis focused on the baseline experiment.

| | AUROC | | | | |
|---|---|---|---|---|---|
| Experiment | extree | forest | xgboost | tree | logreg |
| Baseline | 0.86 | 0.86 | 0.82 | 0.61 | 0.65 |
| House prices | 0.86 | 0.86 | 0.83 | 0.62 | 0.68 |
| HP filter on GDP ratios | 0.87 | 0.87 | 0.83 | 0.63 | 0.68 |
| HP filter on nominal variables | 0.90 | 0.89 | 0.84 | 0.65 | 0.52 |
| HP filter on real variables | 0.87 | 0.85 | 0.83 | 0.65 | 0.53 |
| Hamilton filter on GDP ratios | 0.87 | 0.87 | 0.81 | 0.65 | 0.56 |
| Replace yield curve with nominal rates | 0.88 | 0.87 | 0.84 | 0.59 | 0.66 |
| Replace yield curve with real rates | 0.88 | 0.88 | 0.84 | 0.60 | 0.66 |

Table 4.1: AUROC scores for robustness check, crisisCurrency used as the outcome variable

### 4.1.3 Extremely Randomized Trees in detail

In keeping with comparability across the sections of this work, we take a deeper look at the extremely randomized trees model, despite random forests producing a better false alarm rate in the baseline analysis.

The model can correctly predict the majority of the early 2000s currency crises. It only partially misses 1 crisis during this period. Similarly, the model is able to predict most of the mid-80s currency crises, only fully missing 1 crisis and partially missing 2. The model fully misses 5 currency crises and partially misses 5 more crises throughout the sample. The 5 fully missed crises are Sweden in 1921, Japan in 1971 and 2007,

Finland in 1985, and Germany in 1973.

The model misses a currency crisis in Sweden that happened right after the First World War. During the war, Sweden's economy was booming and stock and asset prices rose (Jordà et al., 2023). However, this was accompanied by rising inflation rates, leading to Sweden returning to the gold standard. Indeed, this fall in the value of the Swedish currency predated the banking crisis in 1922.

Japan experienced great economic growth after the Second World War. This crisis is most likely due to the *Nixon shock* that happened in 1971. This is when former United States president Richard Nixon ended the gold standard. In the post World War 2 era, there were fixed exchange rates around the world, known as the Bretton Woods System (Butkiewicz & Ohlmacher, 2021). For example, 360 yen was always worth 1 USD (Butkiewicz & Ohlmacher, 2021). Since the gold standard was ended, the Japanese government had to pay a substantial sum of money so that the yen would not devalue against the US dollar, but that led to the foreign exchange reserves rapidly increasing and the yen depreciating in value. The model also misses a crisis in Germany in 1973. Butkiewicz and Ohlmacher, 2021 notes that the *Nixon shock* primarily told Japan and Germany to devalue their currency for the sake of US trade. Germany was not free from the shock of Nixon's "America first" international policy (Butkiewicz & Ohlmacher, 2021). This would be a difficult crisis for the model to predict since it was heavily influenced by factors outside of the features used to train the model. Japan also experienced a currency crisis just before the global financial crisis. This was due to the *lost decade*, where the Japanese yen exchange rate was higher than the historical average (Kawai & Takagi, 2011). The higher value yen encouraged the production of goods, however, this reduced the price competitiveness of the manufacturers. Kawai and Takagi, 2011 note that the production of these goods was beneficial in a nominal sense, but it did not translate to a real sense. As Japan implemented a policy to climb out of

this recession, its economy became very sensitive to the US housing price bubble, which burst in the summer of 2006 (Kawai & Takagi, 2011). This led to the yen's devaluation against the USD, which is classified as a currency crisis. We can see that for both of these missed Japanese currency crises their main causes were potentially driven by the United States.

The missed crisis in Finland in 1985 is the only fully missed crisis caused by an event known as *The Plaza Accord* (Frankel, 2015). This event was a result of the *Nixon shock* mentioned above. The decade after Nixon implemented his policy, the US dollar had risen 44% against the other major currencies internationally (Frankel, 2015). Since the US dollar was worth so much, other countries could not afford to trade with the United States. The G-7 countries agreed to devalue the US dollar to get US trade back online (Frankel, 2015). Nixon's policy made most currencies dependent upon the US dollar, so intentionally dropping the value of the US dollar caused currency crises in 14 countries in our sample. Another question is why Finland was the only country in which this crisis was missed. Jonung et al., 2009 reports that during the mid-1980s, Finish banks were doing a lot of lending in foreign currency, which could be a potential cause. The model also partially missed the Finland currency crisis in 1994. Indeed, Jonung et al., 2009 notes that between the currency crises of 1985 and 1994, the Finnish economy was booming. However, stock prices began to fall, and the Finnish currency became overvalued. Between 1990 and 1993, the GDP fell by 13 percent and was accompanied by a much higher unemployment rate (Jonung et al., 2009).

The 5 partially missed crises by the model are as follows: Norway in 1985, the United Kingdom in 1985, France in 1927, Finland in 1994, and Canada in 2003. We have already discussed the global economic climate for the 1985 crises.

During the 1920s, France was accumulating more gold than any other country (Irwin, 2010). It is common knowledge that much of the Great Depression was facilitated by

the gold standard (Irwin, 2010). "Countries not on the gold standard managed to avoid the Great Depression almost entirely, while countries on the gold standard did not begin to recover until after they left it" (Irwin, 2010). The post World War 1 world had the gold standard reevaluated. France experienced some high inflation in the mid-1920s and stabilized the franc in 1926. However, the franc was stabilized at an undervalued rate (Irwin, 2010). Essentially, France was selling off their foreign reserves to accumulate more gold. This led to the franc dropping in value compared to the US dollar.

Lastly, we briefly discuss the Canadian currency crisis that occurred in 2003. All but three countries in our sample are classified as having experienced a currency crisis in either 2002 or 2003. This is due to, yet again, a sharp depreciation of the US dollar.

The dating system used for identifying currency crises is based on Reinhart and Rogoff's threshold of 15% depreciation against the USD. However, because of this construction, the United States is classified as having no currency crises. This is because the positive cases of a currency crisis are based on the exchange rates given in the sixth version of the JST database, which presents all exchange rates against the USD. This means that applying this threshold to the JST database produces no positive cases for the United States since the USD/USD exchange rate is always 1. We also train a model on the narrative currency crisis list given by Reinhart and Rogoff. This model is reported in the appendix.

### 4.1.4 Forecasting

We saw from the 2 cases of a banking crisis that machine learning has potential use in forecasting financial crises. However, the results may not be as pretty for currency crises. Due to the clustered nature of the currency crises, the construction of the forecasting experiment causes some problems. The vast majority of the currency crises occurred in 1985 or 2002, mostly due to the United States devaluing the USD as mentioned

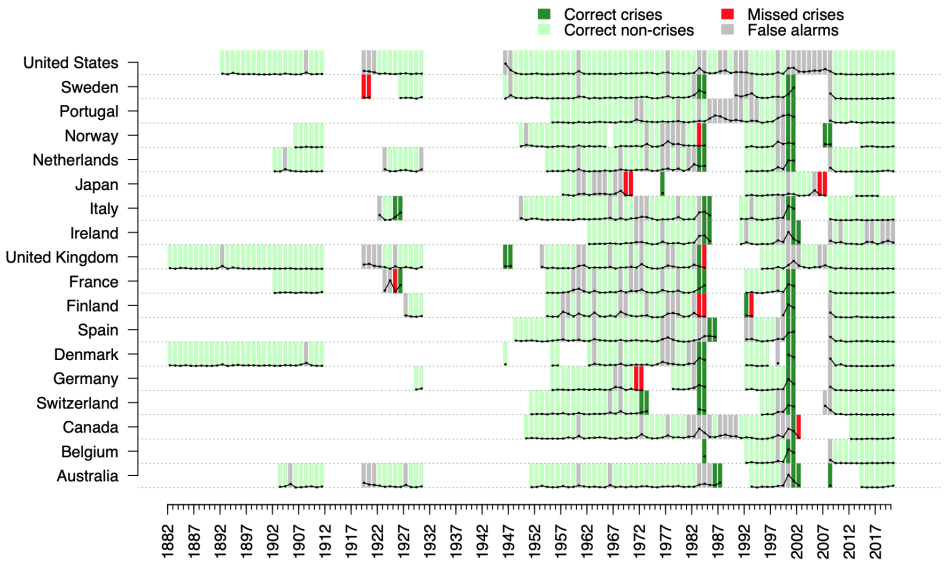


Figure 4.2: Crisis probability estimated by extremely randomized trees at a hit rate of 80%, currency crisis

above. Since the forecasting required a minimum number of crises in the training set, the forecasting model in Figure 4.3 is only able to predict the early 2000s currency crises, along with a few of the later currency crises in the 1980s. When the model is calibrated to an 80% hit rate, we see that the majority of the predictions are false alarms or correct identifications of crises. This forecasting model is very risk-averse. It misses only 1 crisis completely and 6 other crises partially. All of the missed crises occur during the 2002 currency crisis episode. The results do not get much better, even lowering the hit rate to 60% or 30%. Those models obtain marginally more true negatives at the cost of many more false negatives. This is not worth it since false negatives are a much bigger problem than false positives. Extremely randomized trees does produce the lowest false alarm rate of 86.2%. Random forest and logistic regression both have false alarm rates of about 95%. The AUROC scores confirm this ranking but also show that the models may be performing worse than random chance. Extremely randomzied trees, random forest, and logistic regression produce AUROC scores of 0.44, 0.40, and

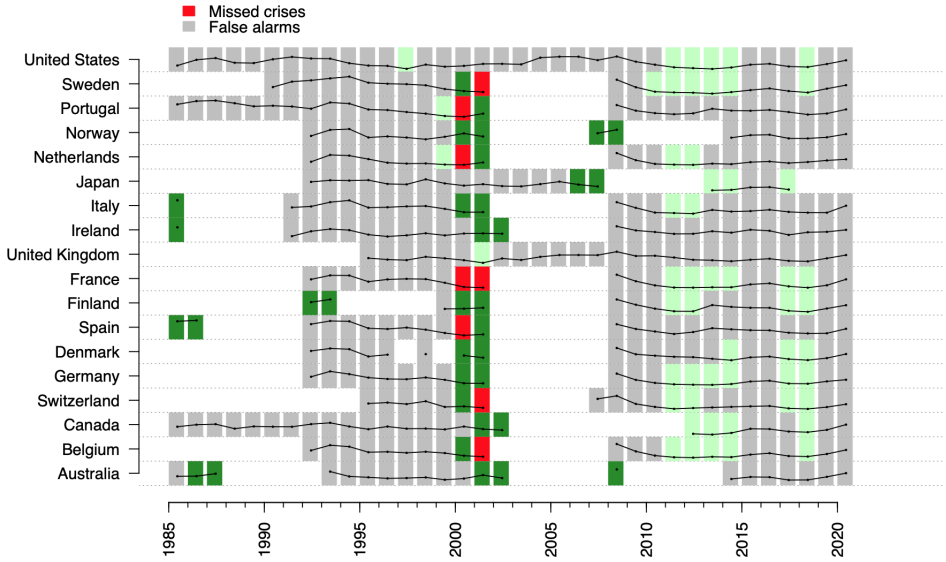


Figure 4.3: Forecasting crisis probability estimated by extremely randomized trees at a hit rate of 80%, crisisCurrency used as the outcome variable

0.41, respectively.

### 4.1.5 Shapley value decomposition: variable importance

Looking at currency crises we see a major change in the ranking of the features. However, the global variables still rank as the top predictors in the extremely randomized trees model. The CPI plays a slightly more prominent role in predicting currency crises, whereas public debt ranks 4th instead of being one of the bottom predictors for banking crises. The two new variables introduced rank as the 3rd and 4th worst in predicting currency crises. We can also see that there is not a lot of agreement between the logistic regression model and the extremely randomized trees model. The only thing these models agree on is that global credit growth is the most important predictor. It is rather interesting that a global variable would be the top predictor in currency crises. However, this could be due to the construction of the currency crisis variable. Since the currency crisis variable is constructed using the exchange rate against the United States

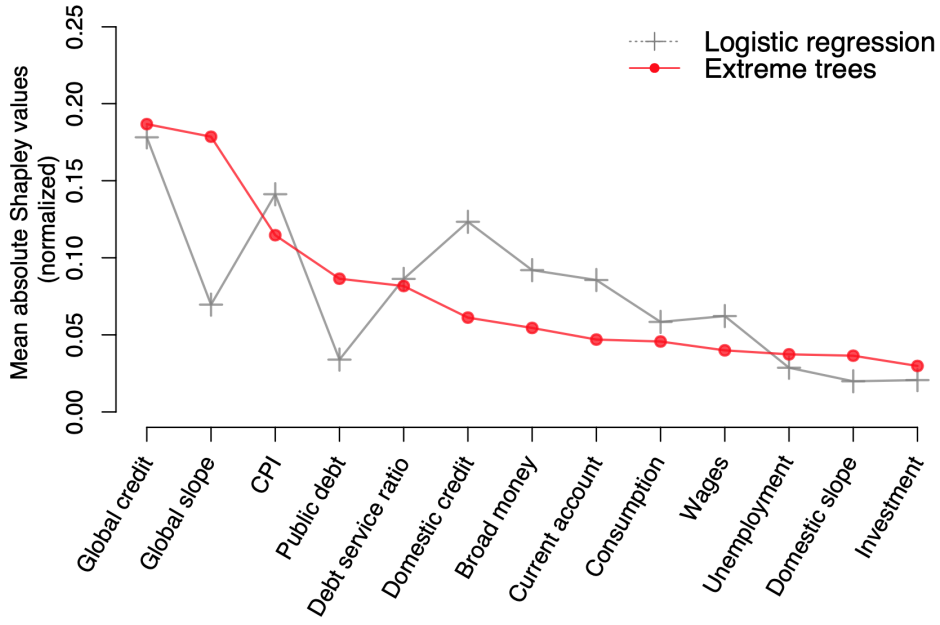


Figure 4.4: Mean absolute Shapley values of individual features for logistic regression and extremely randomized trees, crisisCurrency used as the outcome variable

dollar, whatever is happening in the United States is going to be very important to the other countries. Not only that, we saw from Figure 4.2 that many of the currency crises are clustered in terms of the years that they occur. Thus, there will be elevated levels of global risk during these periods, reflected in the top two predictors in the extremely randomized trees model. Figure 4.4 is perhaps showing that logistic regression cannot produce a good AUROC score for currency crises because it cannot identify trends in predictors in pre-crisis periods. Extremely randomized trees achieve an AUROC score of more than 20 percentage points higher, and the predictions of this model are driven by very different features. After decomposing the predictions into their Shapley values, we can see from Figure 4.5 that the signals are generally quite strong when the model identifies a currency crisis. The advantage of this plot over the prediction charts from earlier sections is that it exaggerates the actual predicted probability of a crisis rather than just its classification as right or wrong. For example, the model produces its high-

est predicted probability of a crisis 3 years in advance of the France crisis in the 1920s. Due to the construction of our models, this comes across as a false alarm. However, this false alarm would still have been a useful early warning indicator due to the high predicted probability. Figure 4.5 also shows just how unexpected the Nixon shock was in Japan, as the predictions during those years are nowhere near indicating a currency crisis would occur.

### 4.1.6 Shapley regressions: variable significance

Table 4.2 shows the results of the Shapley regression. We see that there are only 3 statistically significant features for the currency crisis, and they are the 3 with the highest predictive share. They are the global credit growth, the global slope of the yield curve, and the CPI. It appears as though the CPI and the global slope obtain the highest coefficients and lowest p-values, but the global credit growth still has a p-value $< 0.01$. What this tells us about the extremely randomized trees model for currency crises is that the majority of the information being used for the predictions comes from these 3 features and that the signals they issue are statistically different from those of the null hypothesis. In contrast, there are 5 features that have p-values of 1 and negative coefficients, meaning that the model made little use of the information given by those predictors. For the remaining predictors, we see p-values between 0.2 and 0.5 and positive coefficients of less than 0.1. This means that the model has a hard time distinguishing between these predictors and the null hypothesis in terms of the signals issued.

### 4.1.7 Nonlinearities

We now investigate the importance of nonlinear relationships for other types of financial crises. From Figure 4.4, we can gather that the 2 global features, the CPI and the public

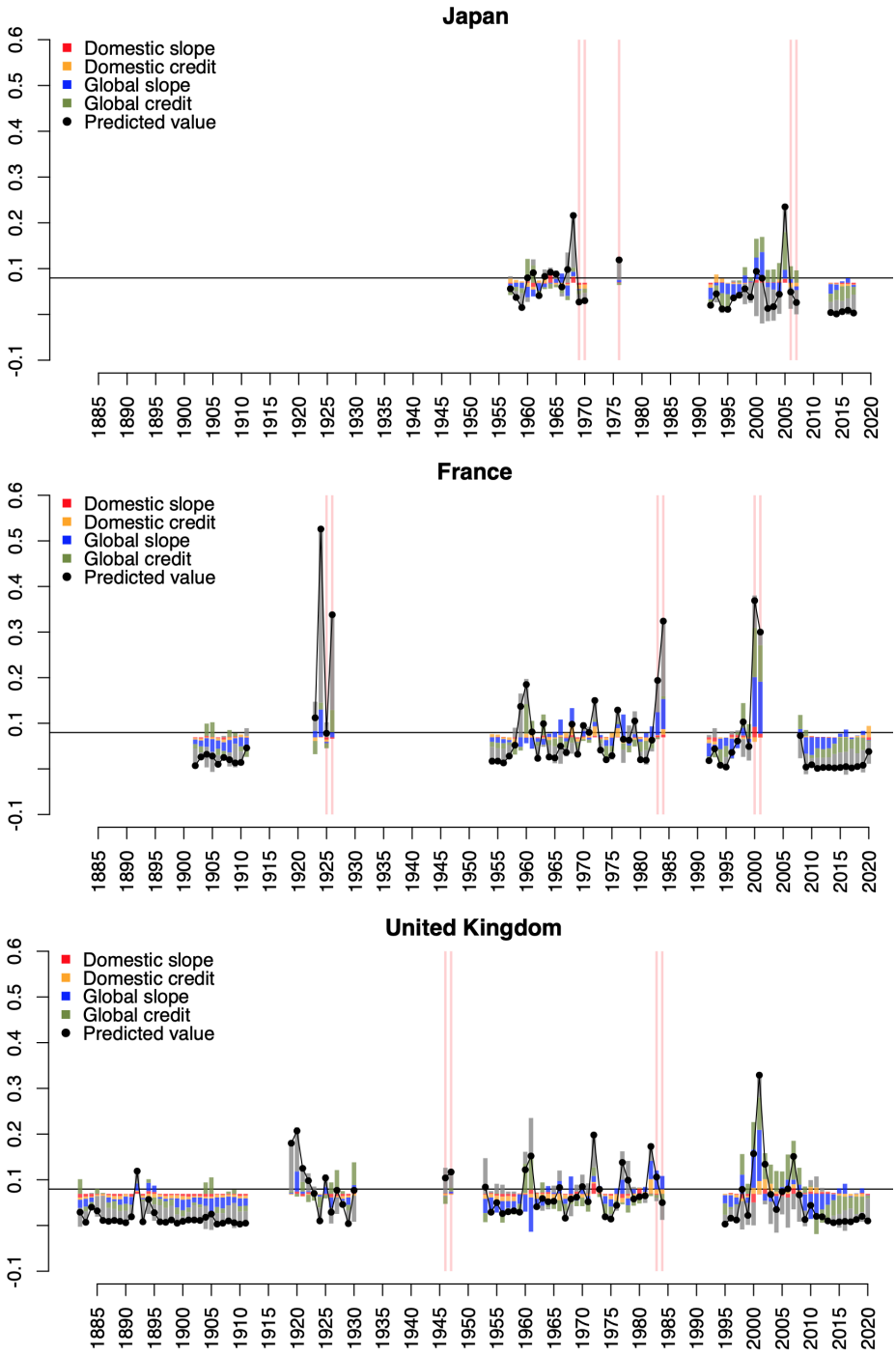


Figure 4.5: Shapley values as a function of time using crisisCurrency as the outcome variable

| Variable | Direction | Shapley Share | Coefficient | p |
|---|---|---|---|---|
| Global credit | + | 0.1870 | 0.5376 | 0.009 |
| Global slope | - | 0.1781 | 0.7972 | 0.000 |
| CPI | + | 0.1151 | 0.8917 | 0.000 |
| Public debt | - | 0.0858 | 0.0411 | 0.411 |
| Debt service ratio | - | 0.0813 | 0.0677 | 0.274 |
| Domestic credit | + | 0.0610 | -0.1703 | 1.000 |
| Broad money | - | 0.0541 | -0.1610 | 1.000 |
| Current account | + | 0.0470 | -0.1092 | 1.000 |
| Consumption | + | 0.0460 | 0.0168 | 0.452 |
| Wages | - | 0.0411 | -0.2475 | 1.000 |
| Unemployment | + | 0.0376 | 0.0241 | 0.441 |
| Domestic slope | - | 0.0363 | 0.0342 | 0.382 |
| Investment | + | 0.0296 | -0.3574 | 1.000 |

Table 4.2: Results of the Shapley regression. Shows the direction of the relationship between predictor and crisis, coefficients, p-value against the null hypothesis, and the predictive share. crisisCurrency used as the outcome variable.

debt, are the 4 most important predictors of currency crises for our model. We plot the relationship between the Shapley values of the predictor, and the predictor values themselves in Figure 4.6 for these 4 features. Immediately, we can see that there are indeed nonlinear relationships present in these features, as a cubic polynomial fits the data significantly better than a straight line. What we notice about public debt is that the Shapley values spike when the change in public debt drops by a significant amount. For the CPI, we notice that the vast majority of crisis observations fall when the CPI growth is between 0 and 20%. The highest Shapley values occur around the 20% growth mark. Interestingly, as you head towards the tails of this distribution, there are very few crisis observations. The global slope of the yield curve, which is listed as the second most important predictor in Figure 4.4, yields a very interesting relationship between the predictor value and its Shapley values. What we see is a large spike in the Shapley values for observations where the global slope of the yield curve is right around or slightly below 1. This means that the long-term interest rates are slightly higher than the short-term interest rates. The model learned that when the

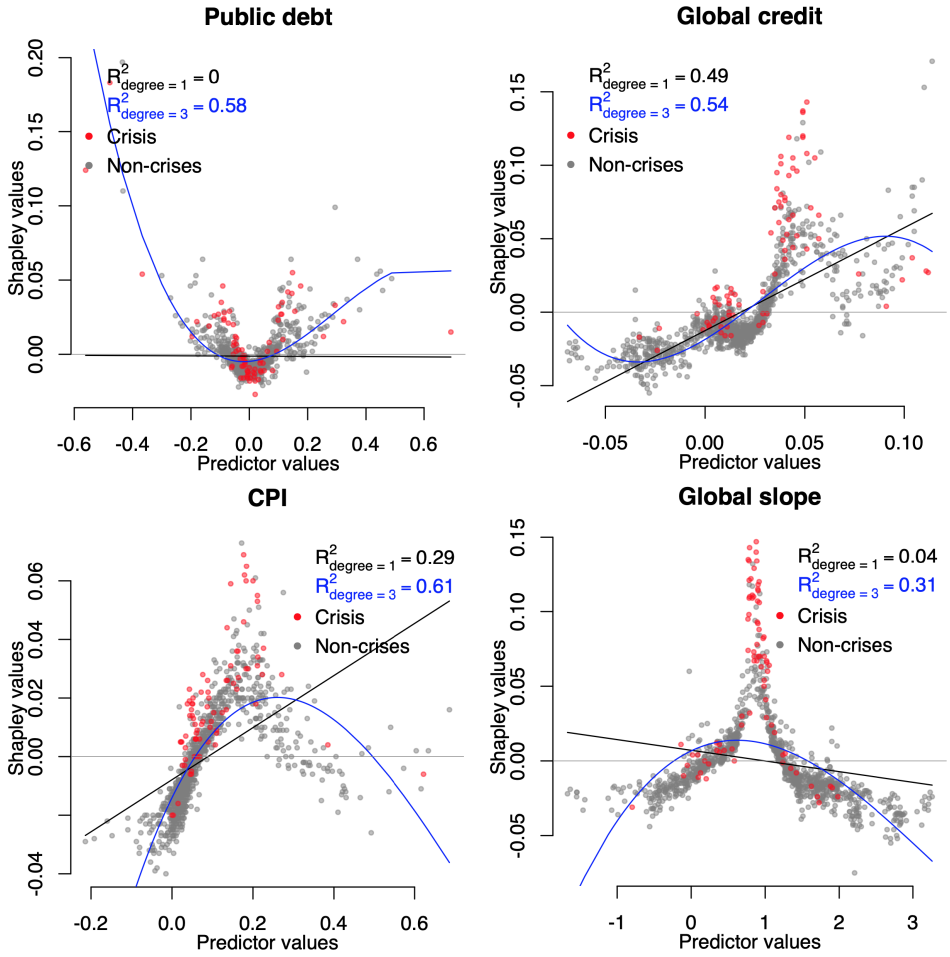


Figure 4.6: Shapley values of key predictors as functions of the predictor values. Currency crisis observations shown in red.

difference between the long- and short-term rates is just below one, it is more likely to be a pre-crisis period. The global credit growth relationship shows a general trend of rising global credit growth, resulting in greater Shapley value for that observation. Again, this trend does not continue into the tails of the distribution but rather peaks just shy of 5% global credit growth. Just before 5% global credit growth, we see a cluster of crisis observations with high Shapley values. However, as the credit growth rises past 7%, the number of crisis observations decreases greatly.

We apply the same experiment to check whether these nonlinearities can help im-

prove predictive performance. Table 4.3 shows the results of this experiment. For 9

| Feature | Linear | Shapley |
|---|---|---|
| Domestic slope | 0.49 | 0.63 |
| CPI | 0.65 | 0.79 |
| Broad money | 0.57 | 0.55 |
| Unemployment | 0.51 | 0.67 |
| Wages | 0.59 | 0.72 |
| Consumption | 0.53 | 0.62 |
| Public debt | 0.52 | 0.61 |
| Investment | 0.50 | 0.49 |
| Current account | 0.56 | 0.53 |
| Domestic credit | 0.55 | 0.53 |
| Debt service ratio | 0.49 | 0.70 |
| Global credit | 0.66 | 0.81 |
| Global slope | 0.59 | 0.82 |

Table 4.3: AUROC of univariate linear regressions compared to the crisisCurrency variable being regressed on the individual predictor Shapley values.

of the 13 predictors, the AUROC improves. The improvements are even more pronounced, with the biggest improvement being for the global slope of the yield curve, which improves by 22 percentage points. For the 4 key predictors shown in Figure 4.6, we see the AUROC improve by at least 10 percentage points. The biggest loss of AUROC occurs for broad money, which drops 3 percentage points when regressed on the Shapley values. Compared to the banking crisis cases, the nonlinear relationships seem even more important for predicting currency crises.

## 4.2 Inflation Crisis

We now shift our attention to the definition of an inflation crisis given by Reinhart and Rogoff: An observation is classified as an inflation crisis if the country experiences "an annual inflation rate of 20 percent or higher" (Reinhart & Rogoff, 2009).

### 4.2.1 Baseline Analysis

The first thing to note about the inflation crisis case is that the number of positive cases in the dataset is much less than in the other 3 cases we have looked at. Indeed, there are only 14 positive indicators for inflation crises, representing 8 distinct crises. We do see from Figure 4.7 that the ranking of the models still holds. These are the highest AUROC scores of all 4 cases. Some of the literature suggests that the AUROC does not fare well with severely imbalanced datasets. However, in a more recent work, Richardson et al., 2023 shows that the AUROC is indeed robust to class imbalance. One difference between the baseline models for this definition of a financial crisis is that the CPI variable is removed. This is because the CPI is basically the definition of inflation in our sample, as the 0s and 1s are generated by that variable in the JST database. Extremely randomized trees produce a false alarm rate of 13.1%, followed closely by random forests with 15.7%. Both of these false alarm rates are significantly lower than the 25.2% false alarm rate achieved by logistic regression. XGBoost does not compare well to the other machine learning models in this case, still outperforming a decision tree but not logistic regression.

### 4.2.2 Robustness

Table 4.4 that the model ranking is relatively robust across the different experiments. Logistic regression performs best in one experiment, namely when the HP filter is applied to the GDP ratios of the features. However, the AUROC that logistic regression obtains is still lower than the baseline experiment using the extremely randomized trees model, so we do not feel this is an issue. In general, across the table, we see that random forest performs quite comparably to extremely randomized trees, even outperforming extremely randomized trees in 1 experiment.

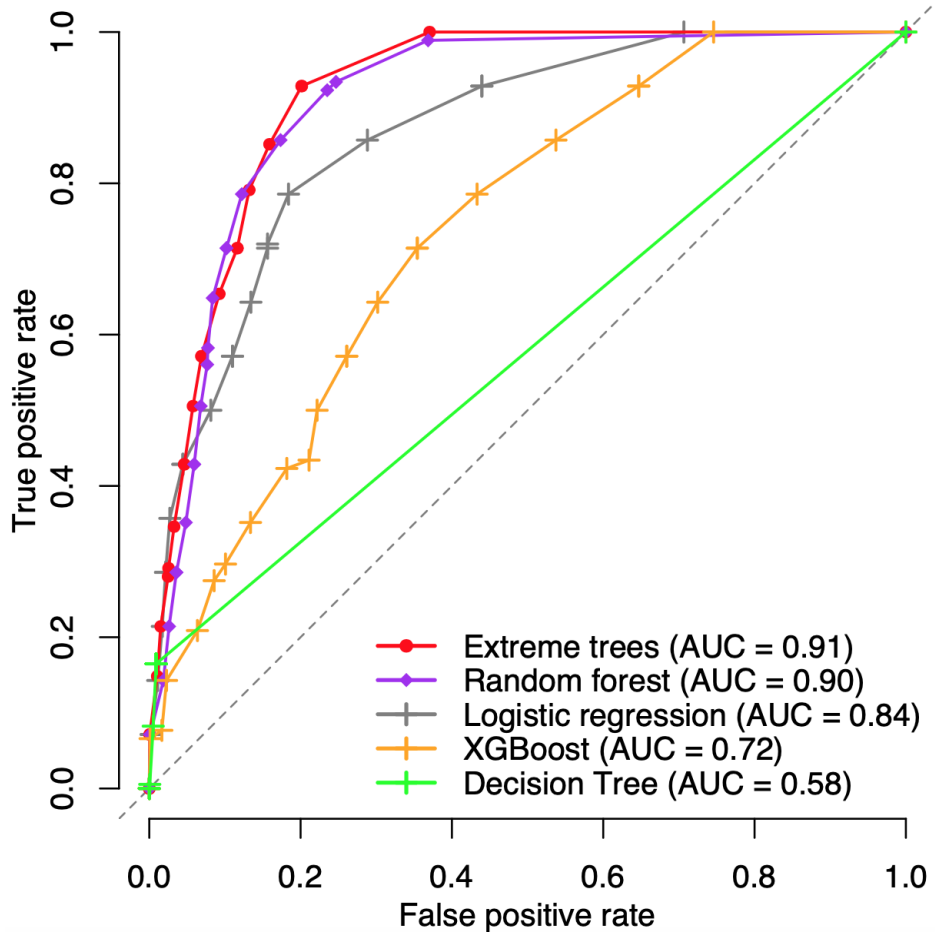


Figure 4.7: ROC curves for baseline models, crisisInflation used as the outcome variable

| | AUROC | | | | |
|---|---|---|---|---|---|
| Experiment | extree | forest | xgboost | tree | logreg |
| Baseline | 0.91 | 0.90 | 0.72 | 0.58 | 0.84 |
| House prices | 0.82 | 0.75 | 0.61 | 0.54 | 0.75 |
| HP filter on GDP ratios | 0.88 | 0.85 | 0.66 | 0.56 | 0.90 |
| HP filter on nominal variables | 0.91 | 0.87 | 0.67 | 0.55 | 0.83 |
| HP filter on real variables | 0.91 | 0.89 | 0.71 | 0.52 | 0.83 |
| Hamilton filter on GDP ratios | 0.92 | 0.92 | 0.71 | 0.58 | 0.85 |
| Replace yield curve with nominal rates | 0.89 | 0.91 | 0.66 | 0.58 | 0.89 |
| Replace yield curve with real rates | 0.92 | 0.92 | 0.67 | 0.55 | 0.84 |

Table 4.4: AUROC scores for robustness check, crisisInflation used as the outcome variable

### 4.2.3 Extremely Randomized Trees in detail

With just a quick glance at Figure 4.8, we can see that since there are so few positive indicators, the majority of the predictions will be true negatives. We can also see that the predicted probability given by the model for many of these observations is very close to zero. This figure represents this by the black lines being almost flat on the bottom. There are 8 inflation crises in our sample, and the model misses 1 of them fully and 1 of them partially. The fully missed crisis happened in France in 1926. The partially missed crisis happened in Portugal in 1974.

The inflation crisis in France in 1926 is closely related to the currency crisis of 1927. We mentioned above that the franc was stabilized at an undervalued rate as an attempt to get France out of its stretch of inflation. This inflationary episode is confirmed by Tryon, 1979, who looked in depth at the franc between world wars. "In April 1925 it was announced that the legal limit of the currency in circulation had been exceeded earlier in the year by the central banks" (Tryon, 1979). This inflation crisis was somewhat brought about by poor government fiscal policy, making it unsurprising that the model would miss it.

Portugal had the first indicator year of its inflation crisis of 1974 missed. This was

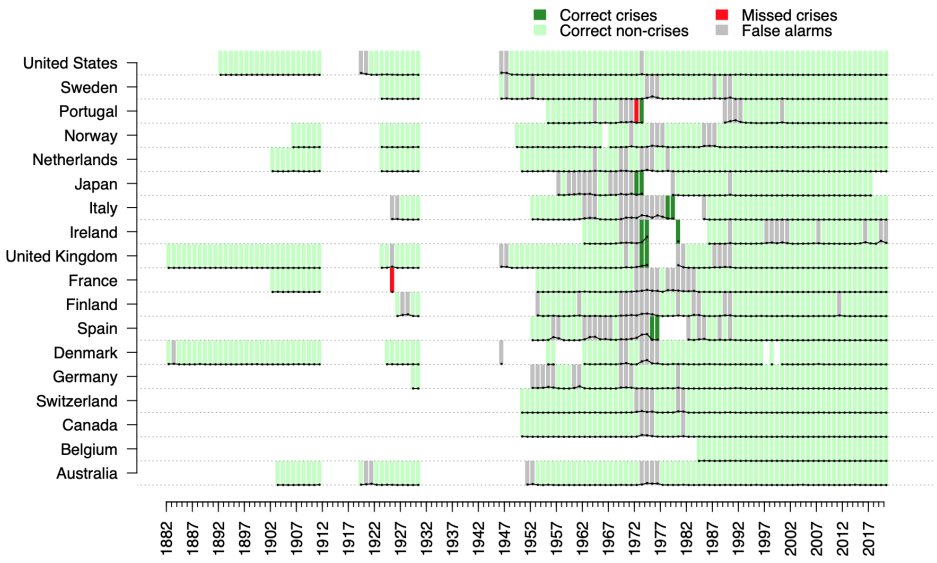


Figure 4.8: Crisis probability estimated by extremely randomized trees at a hit rate of 80%, crisisInflation used as the outcome variable

during the oil crisis of the 1970s. Fundação Francisco Manuel Dos Santos, 2023 notes a couple of things about Portugal during this time. First, Portugal had a stock market boom along with an expansion in credit and money supply (Fundação Francisco Manuel Dos Santos, 2023). Secondly, there was a political revolution in 1974, which resulted in 80000 workers being hired. Of course, all of these new salaries that were created out of nothing caused the inflation rate to spike later that year.

### 4.2.4 Forecasting

As was the case with the currency crisis case, the forecasting period for inflation crises is also much smaller than for banking crises. This is simply due to the low number of positive cases in the dataset for inflation crises. We set the model to need 7 positive indicators in the training set before training. This results in 3 inflation crises that the model can try to predict while forecasting. Again, similarly to the currency crisis case, we cannot choose a lower number than 7 since the forecasting training would begin

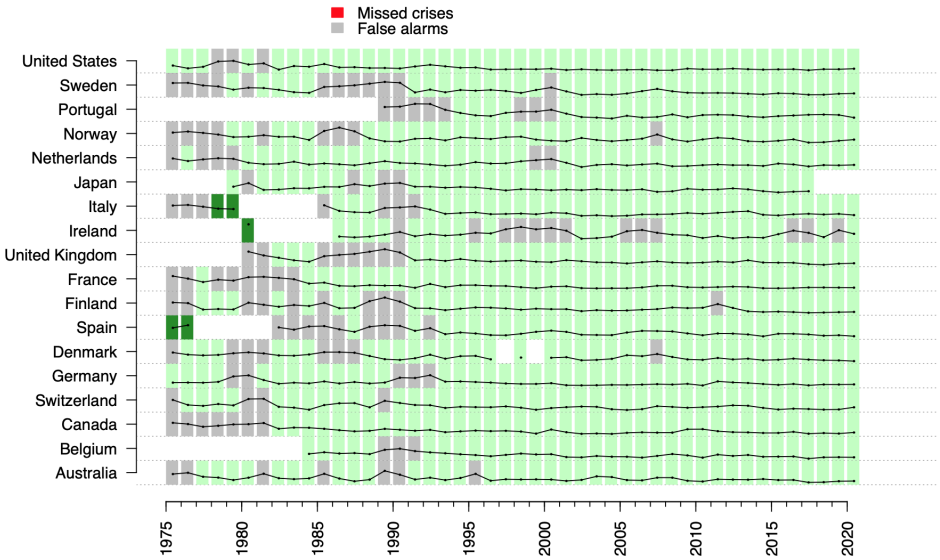


Figure 4.9: Forecasting crisis probability estimated by extremely randomized trees at a hit rate of 80%, crisisInflation used as the outcome variable

with only 1 positive case. Figure 4.9 shows us that after 1995, every country except Ireland was never really at risk of experiencing an inflation crisis. We also note that since there are only 5 indicators in the forecasting experiment, a minimum hit rate of 80% means that all crises will be identified correctly. Changing the minimum hit rate so that 1 indicator was missed did improve the false alarm rate of the model that much. Extremely randomized trees produce a false alarm rate of 17.1%. This is a drastically better false alarm rate than logistic regression, which obtains a false alarm rate of 78.1%. Random forests produce the lowest false alarm rate in this case at 16.4%. So, in terms of how these early warning models are used in practice, the machine learning models outperform logistic regression by a wide margin, confirming their potential usefulness in inflation crisis prediction. The AUROC scores confirm this ranking, with random forests achieving an AUROC of 0.910, extremely randomized trees an AUROC of 0.909, and logistic regression an AUROC of 0.727.

### 4.2.5 Shapley value decomposition: variable importance

We turn our attention to the most important features for predicting inflation crises. Figure 4.10 shows that the wage variable is far and away the most important predictor for the extremely randomized trees model. It has a mean absolute Shapley value of more than double any other predictor. We knew from Table 2.4 that the wage variable was potentially important for inflation crisis prediction, but extremely randomized trees clearly recognized something different than logistic regression. Consumption ranks second, followed by the global slope of the yield curve. Most of the other features following have a relatively equal mean absolute Shapley value. The unemployment rate has the definitively lowest mean absolute Shapley value for predicting inflation crises in the extremely randomized trees model. There is not a lot of agreement between logistic regression and extremely randomized trees. Logistic regression gives much more weight to global credit growth and broad money. However, both of these models agree that changes in wages are indicative of inflation crises. The decomposition of the predictions into their Shapley values over time is slightly different for inflation crises compared to the other types of financial crises. Figure 4.11 shows that the predicted probabilities of inflation crises are, across the board, much lower than the other financial crises. Spain is perhaps the most sporadic country in this sample but it seems to get a hold of inflation by the 1990s. One thing to note here is that all countries have a spike in predicted probability in the mid-1970s. We mentioned above the effect of the Nixon shock, and Figure 4.11 shows just how global its effects were.

### 4.2.6 Shapley regressions: variable significance

Table 4.5 shows the results of the Shapley regression on the predictors. Compared to the other 3 cases, the extremely randomized trees model for inflation crises is not

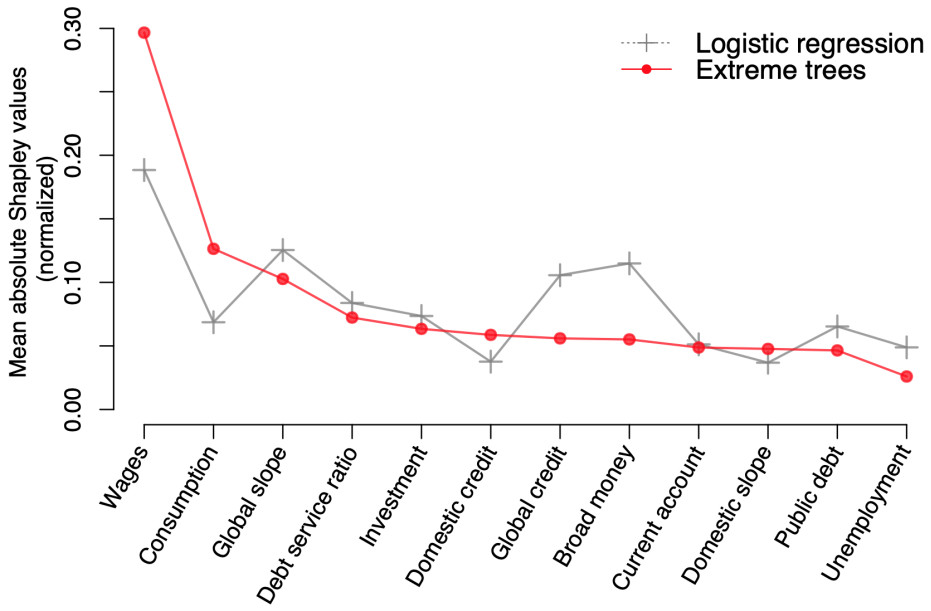


Figure 4.10: Mean absolute Shapley values of individual features for logistic regression and extremely randomized trees using crisisInflation as the outcome variable

| Variable | Direction | Share | Coefficients | p |
|---|---|---|---|---|
| Wages | + | 0.297 | 0.2993 | 0.105 |
| Consumption | + | 0.126 | 0.4308 | 0.016 |
| Global slope | - | 0.103 | 0.5300 | 0.015 |
| Debt service ratio | + | 0.072 | -0.3762 | 1.00 |
| Investment | + | 0.063 | -2.2734 | 1.00 |
| Domestic credit | + | 0.059 | -0.2976 | 1.00 |
| Global credit | - | 0.056 | -0.0415 | 1.00 |
| Broad money | + | 0.055 | -0.1713 | 1.00 |
| Current account | + | 0.049 | -0.2626 | 1.00 |
| Domestic slope | + | 0.048 | 0.2047 | 0.035 |
| Public debt | - | 0.046 | -1.2192 | 1.00 |
| Unemployment | - | 0.026 | -0.3897 | 1.00 |

Table 4.5: Shapley regression. Shows the direction of the relationship between predictor and crisis, coefficients, p-value against the null hypothesis, and the predictive share. crisisInflation used as the outcome variable.

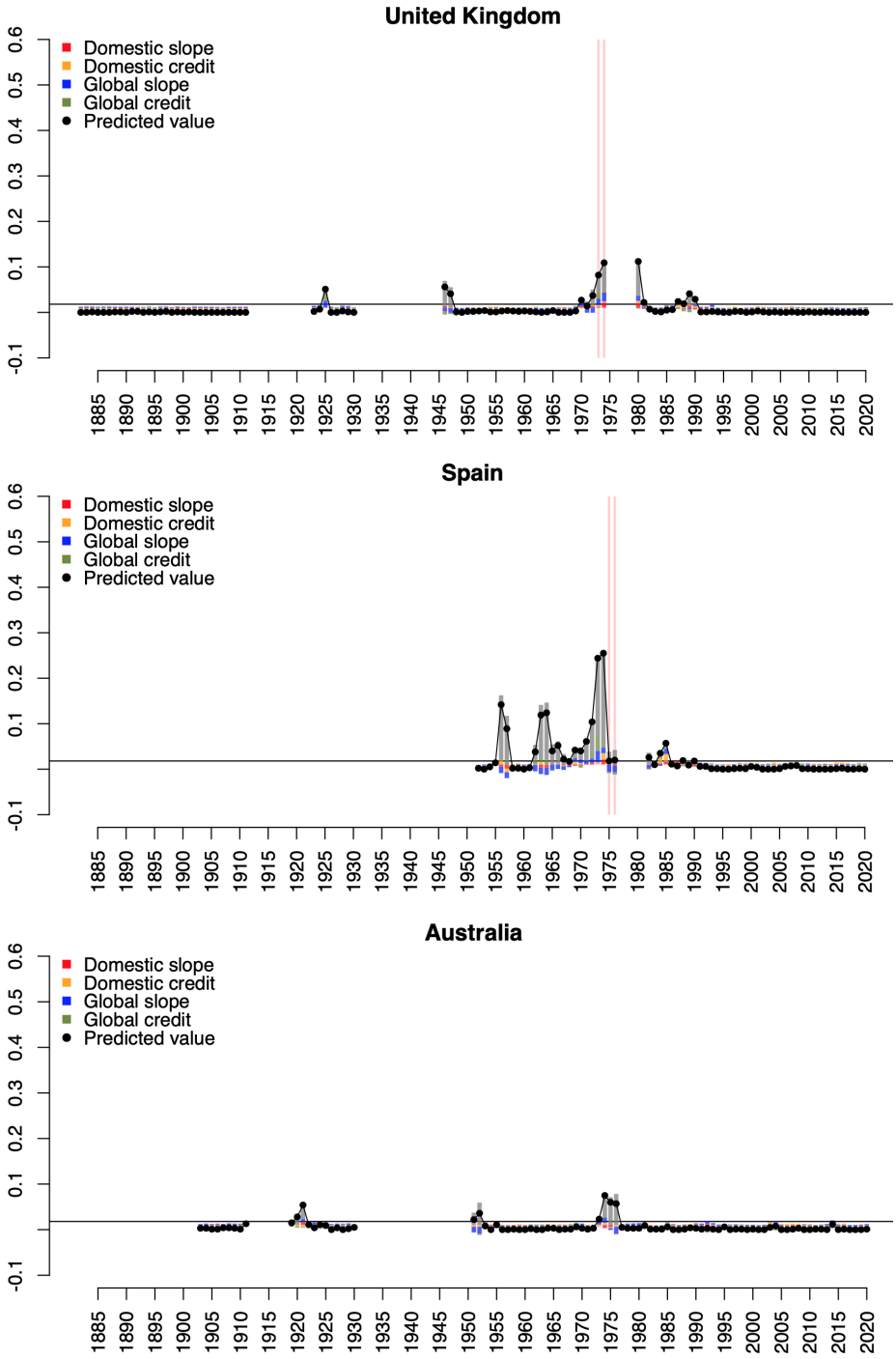


Figure 4.11: Shapley values as a function of time using crisisInflation as the outcome variable

able to differentiate between noise and signal for many of the predictors. Indeed, all but 4 predictors have negative coefficients and p-values of 1. This indicates that the model is making little use of the majority of the predictors. Table 4.5 shows that consumption, the global slope of the yield curve, and the domestic slope of the yield curve are the most significant predictors, yielding the highest coefficients and lowest p-values. The wages predictor, which dominates the predictions regarding Shapley share, is not considered statistically significant when regressed on the Shapley values. However, it obtains a larger coefficient than the consumption predictor. This means that it is still contributing meaningfully to the correct predictions. The model has a marginally more difficult time distinguishing between its signals and the noise.

### 4.2.7 Nonlinearities

Figure 4.12 shows the relationship between Shapley values and the predictor values for the 4 top predictors of inflation crises. The interpretations of these results are less obvious than in the other 3 cases, mostly due to the small number of crisis observations in the inflation crisis case. Figure 4.12 shows that cubic polynomials better fit the data than a straight line for all 4 predictors. Looking at the consumption plot, we can see there is definitely a spike in Shapley values when the consumption rises more than 10% in a year, but this does not necessarily mean that correlates to a crisis prediction. The global slope of the yield curve ranks third in variable importance. Again, Figure 4.12 seems to show a nonlinear relationship between the predictor values and the Shapley values, but the crisis observations are spread over a range of both predictor and Shapley values. The debt servicing ratio seems to have most of the crisis observations occurring when the model is attributing negative Shapley values to the predictor. This interpretation is in line with Table 4.5, where we see that the debt servicing ratio is not statistically significant and that the model has a hard

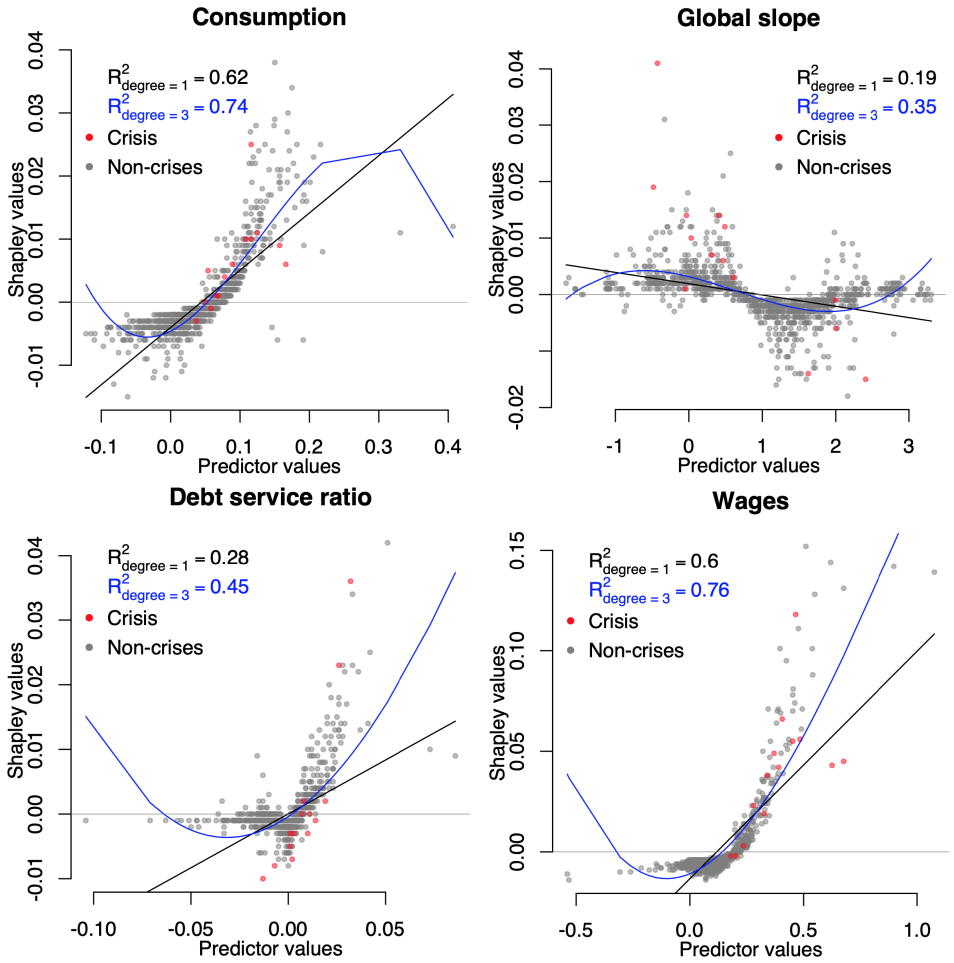


Figure 4.12: Shapley values of key predictors as functions of the predictor values. Inflation crisis observations shown in red.

distinguishing it from the noise. The top predictor of inflation crises by a long shot is the wage variable. The cubic fits the data slightly better than the straight line, and we get the general trend that as the wages rise from year to year, the model attributes a higher likelihood of being in a pre-inflation crisis period.

Table 4.6 shows the results of the univariate regressions comparing the predictor values and their Shapley values. 8 of the 12 predictors achieved a higher AUROC score when regressed on the Shapley values compared to when they were regressed on the predictor values. Because of the imbalanced dataset, the losses and gains when

| Feature | Linear | Shapley |
|---|---|---|
| Domestic slope | 0.52 | 0.56 |
| Broad money | 0.44 | 0.63 |
| Unemployment | 0.50 | 0.61 |
| Wages | 0.95 | 0.94 |
| Consumption | 0.81 | 0.81 |
| Public debt | 0.45 | 0.79 |
| Investment | 0.61 | 0.72 |
| Current account | 0.61 | 0.60 |
| Domestic credit | 0.59 | 0.56 |
| Debt service ratio | 0.72 | 0.60 |
| Global credit | 0.49 | 0.69 |
| Global slope | 0.60 | 0.70 |

Table 4.6: AUROC of univariate linear regressions compared to the crisisInflation variable being regressed on the individual predictor Shapley values.

regressing on the Shapley values are greater. The biggest loss is incurred by the debt servicing ratio, which loses 13 percentage points when regressed on its Shapley values. In contrast, the biggest gain is achieved by the public debt predictor, which has an increase of 33 percentage points when regressed on its Shapley values. For both consumption and wages, the AUROC change is less than 1 percentage point.

Combining the results from Table 4.6 and Figure 4.12 shows the importance of accounting for the nonlinearities in the variables for the machine learning models.

# Chapter 5

# Concluding Remarks and Future Research

This thesis aims to show the potential benefits that can be gained by using machine learning in financial crisis prediction. This was done by examining a variety of machine learning algorithms and evaluating their performance on $\langle country, year \rangle$ observations in both cross-validation and forecasting experiments.

Chapter 1.2 gives an overview of the framework of the procedure of the experiments, including descriptions of the features, their transformations, the definitions of crises, the cross-validation method, the forecasting method, and the AUROC scores for model evaluation. Chapter 2.5 gives a brief description of all the machine learning algorithms used in this work. A description of the Shapley value framework is then presented, which is the method of choice for attempting to understand the drivers of the predictions in these machine learning models.

Chapter 3 then shows the results of the experiments for the 2 different definitions of banking crises used. The first set of banking crises used is from the Macrohistory database (Jordà et al., 2017). The other set of banking crises is taken from Reinhart

and Rogoff, 2009. The results of the experiments show that machine learning models can indeed achieve better AUROC scores than logistic regression, particularly tree-based algorithms like random forests and extremely randomized trees. Through the Shapley framework, we find that the credit growth and slope of the yield curve, both domestically and globally, are the top 4 predictors for both definitions of a banking crisis.

We then turn our attention to different types of financial crises. Chapter 4 shows the results of the experiments for both currency crises and inflation crises. The binary variables for these crises were constructed from the thresholds described in Reinhart and Rogoff, 2009. The results hold for these different financial crises, with the tree-based machine learning algorithms outperforming logistic regression in cross-validation and forecasting. The Shapley values show that the global credit growth, the global slope of the yield curve, the CPI, and the public debt are the most important predictors of a currency crisis in the extremely randomized trees model. For an inflation crisis, the most important predictors are wages, consumption, the global slope of the yield curve, and the debt servicing ratio.

At the end of the analysis for each financial crisis definition, we investigated the nonlinear relationships between the predictor values and their Shapley values. Using a univariate regression experiment, we find that using these nonlinear relationships can improve model performance. This experiment confirms the presence of these nonlinear relationships, which have been pointed out in the literature (Bluwstein et al., 2021).

The models presented in this work serve as a baseline. The aim of this work is to show the potential use of machine learning in financial crisis prediction, not to build the most accurate model. This work lays the foundation of machine learning for predicting different types of crises ahead of time. However, there are a few potential directions for future research. Firstly, reinforcement learning is another branch of machine learn-

ing where the model will learn from past mistakes or successes. However, the use of reinforcement learning for financial crisis prediction is less explored in the literature. Reinforcement learning models require lots of training data for the model to learn from, which may be one barrier in applying it to financial crisis prediction. Alkhafaji et al., 2023 looks at using reinforcement learning for financial crisis prediction, but they do not use a common dataset. Thus, there is room for research in training a reinforcement learning model on one of the main financial crisis databases mentioned in Chapter 1.2. Namaki et al., 2023 confirms that reinforcement learning was not even on the radar of the financial crisis prediction literature up to 2022.

Second, this work focuses in a sense on crisis prediction by country. The economic conditions in each country are different, though. We can see, for example, in the JST crisis case that Canada experienced no crises throughout the entire sample. If one wanted to train a model specifically for Canada, it could not be done under this approach, as there would be no positive cases to learn from. There is a topic in machine learning called *transfer learning*, where one applies a pre-trained model to a second task. A similar phenomenon occurs quite frequently in our everyday lives. Take a person who knows how to play the violin, for example. This person would be able to learn how to play the piano much quicker than someone who has no musical training at all (Zhuang et al., 2021). In the context of financial crisis prediction in Canada, the goal would be to apply a model that was trained on data from a country with a similar economy to that of Canada. Then, even though Canada had no positive cases in the training data, we could still use the model to try and predict what would happen in the future. Zhuang et al., 2021 shows that this realm of study is up and coming in the machine learning world. We hope that this work serves as the groundwork for future work done in financial crisis prediction.

# Bibliography


Aliber, R. Z., & Kindleberger, C. P. (2015). *Manias, panics, and crashes: A history of financial crises, seventh edition* (7th). Palgrave Macmillan.

Alkhafaji, M. A., Ameer, S. A., Alawadi, A. H., & Sharif, H. (2023). Design of hyperparameter tuned deep reinforcement learning based prediction model for financial crisis. *2023 6th International Conference on Engineering Technology and Its Applications (IICETA)*, 287–293. https://doi.org/10.1109/IICETA57613.2023.10351362

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Awan, A. A. (2023). An introduction to shap values and machine learning interpretability.

Bank failures case study: - johnson matthey. (n.d.).

Baron, M., Verner, E., & Xiong, W. (2020). Banking crises without panics. https://doi.org/10.2139/ssrn.3116148.

Bernanke, B. S., & Blinder, A. S. (1992). The federal funds rate and the channels of monetary transmission. *American Economic Review*, *82*(4), 901–921.

Betran, C., & Pons, M. A. (2013). Understanding spanish financial crises, 1850-2000: What determined their severity? *European Historical Society Conference*, *6*.

Beutel, J., List, S., & von Schweinitz, G. (2019). Does machine learning help us predict banking crises? [https://ideas.repec.org//a/eee/finsta/v45y2019ics1572308918305801.html]. *Journal of Financial Stability*, *45*.

Bhamani, F., Grasselli, M., Haussamer, N., Jiang, Y., & Mathonsi, T. (2018). An early warning system for financial crises and long-term asset management [unpublished].

Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Şimşek, Ö. (2021). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics*, *145*, 103773. https://doi.org/10.1016/j.jinteco.2023.103773

Borio, C., & Drehman, M. (2008). Assessing the risk of banking crises- revised. *Bank for International Settlements*, 29–46.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Bussiere, M., & Fratzscher, M. (2002). Towards a new early warning system of financial crises.

Butkiewicz, J. L., & Ohlmacher, S. (2021). Ending bretton woods: Evidence from the nixon tapes. *The Economic History Review*, *74*(4), 922–945. https://doi.org/10.1111/ehr.13052

Caprio, G., & Klingebiel, D. (2003). Episodes of systemic and borderline financial crises.

Casabianca, E. J., Catalano, M., Forni, L., Giarda, E., & Passeri, S. (2022). A machine learning approach to rank the determinants of banking crises over time and across countries. *Journal of International Money and Finance*, *129*, 102739. https://doi.org/10.1016/j.jimonfin.2022.102739

Cesa-Bianchi, A., Eguren Martin, F., & Thwaites, G. (2019). Foreign booms, domestic busts: The global dimension of banking crises. *Journal of Financial Intermediation*, *37*, 58–74.

Chen, M., DeHaven, M., Kitschelt, I., Lee, S. J., & Sicilian, M. J. (2023). Identifying financial crises using machine learning on textual data. *International Finance Discussion Paper*, *1374*, 1–40. https://doi.org/10.17016/ifdp.2023.1374

Davis, E., & Karim, D. (2008). Comparing early warning systems for banking crises. *Journal of Financial Stability*, *4*, 89–120.

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Frankel, J. (2015). The plaza accord, 30 years later. National Bureau of Economic Research. https://doi.org/10.3386/w21813.

Frankel, J., & Saravelos, G. (2012). Can leading indicators assess country vulnerability? evidence from the 2008-09 global financial crisis. *Journal of International Economics*, *87*, 216–231. https://doi.org/10.1016/j.jinteco.2011.12.009.

Fundação Francisco Manuel Dos Santos. (2023). 1973-1978: Three crises, one long recession.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*, 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Greenwood, R., Hanson, S. G., Shleifer, A., & Sørensen, J. A. (2020). Predictable financial crises.

Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *3*(5).

Irwin, D. (2010). Did france cause the great depression? National Bureau of Economic Research. https://doi.org/10.3386/w16350.

Jonung, L., Kiander, J., & Vartia, P. (2009). The great financial crisis in finland and sweden—the dynamics of boom, bust and recovery, 1985-2000. *European Economy Economic Papers*, *350*.

Jorda, O., Schularick, M., & Taylor, A. M. (2013). When credit bites back. *Journal of Money, Credit and Banking*, *45*(s2), 3–28.

Jordà, Ò., Schularick, M., & Taylor, A. M. (2023). Jst financial crisis chronology.

Jordà, Ò., Schularick, M., & Taylor, A. M. (2017). Macrofinancial history and the new business cycle facts. *NBER Macroeconomics Annual*, *31*.

Joseph, A. (2019). Shapley regressions: A framework for statistical inference on machine learning models. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3351091.

Kaminsky, G., & Reinhart, C. M. (1997). Leading indicators of currency crises. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2882556.

Kawai, M., & Takagi, S. (2011). Why was japan hit so hard by the global financial crisis? In *The impact of the economic crisis on east asia*. Edward Elgar Publishing. https://doi.org/10.4337/9780857931702.00016

Laeven, L., & Valencia, F. (2018). Systemic banking crises revisited. IMF Working Papers 18/206, International Monetary Fund.

Li, J. Z. (2018). Principled approaches to robust machine learning and beyond. Massachusetts Institute of Technology.

Liu, L., Chen, C., & Wang, B. (2021). Predicting financial crises with machine learning methods. *Journal of Forecasting*, *41*(5), 871–910. https://doi.org/10.1002/for.2840

Lundberg, S. M. (2018). Shap (shapley additive explanations).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, pp. 4765-4774.

Martin-Acena, P. (2014). The savings banks crises in spain: When and how? Documentos de Trabjo (DT-AEHE) 1404, Asociacion Espanola de Historia Economica.

Mitchener, K. J., & Weidenmier, M. D. (2008). The baring crisis and the great latin american meltdown of the 1890s. *The Journal of Economic History*, *68*(2), 462–500.

Namaki, A., Eyvazloo, R., & Ramtinnia, S. (2023). A systematic review of early warning systems in finance. arXiv:2310.00490. http://arxiv.org/abs/2310.00490.

Reid, M. (1982). *The secondary banking crisis, 1973-1975: Its causes and course*. Springer.

Reimann, C. (2024). Predicting financial crises: An evaluation of machine learning algorithms and model explainability for early warning systems. *Review of Evolutionary Political Economy*. https://doi.org/10.1007/s43253-024-00114-4

Reinhart, C. M., & Rogoff, K. S. (2009). *This time is different: Eight centuries of financial folly*. Princeton University Press.

Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2023). The roc-auc accurately assesses imbalanced datasets. https://doi.org/10.2139/ssrn.4655233.

Saab, G., Jamhour, T., El-Hayek, M.-M., & Yaacoub, H. K. (2024). The logit model: A prediction of future economic events. *Journal of Mathematical Finance*. https://doi.org/10.4236/jı

Santhanam, R., Uzir, N., Raman, S., & Banerjee, S. (2017). Experimenting xgboost algorithm for prediction and classification of different datasets.

Schularick, M., & Taylor, A. M. (2012). Credit booms gone bust: Monetary policy, leverage cycles, and financial crises, 1870-2008. *American Economic Review*, *102*(2), 1029–61. https://doi.org/10.1257/aer.102.2.1029

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317.

Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, *44*(2), 293–335. https://doi.org/10.1016/S0304-3932(99)00027-6

Tölö, E. (2019). Predicting systemic financial crises with recurrent neural networks. *Journal of Financial Stability*, *49*, 100746. https://doi.org/10.1016/j.jfs.2020.100746

Tomczak, K. (2023). The impact of covid-19 on the banking sector. are we heading for the next banking crisis? *Qualitative Research in Financial Markets*, ahead-of–print. https://doi.org/10.1108/QRFM-09-2021-0157

Tryon, R. W. (1979). The french franc in the 1920q's. Massachusetts Institute of Technology.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

# Appendix A

# Appendix

## A.1 Replication Material

The following link redirects to a github page with all the necessary scripts and data for running this code and reproducing the results.

## A.2 Machine Learning model implementation

The following section of the appendix describes the packages used to train the machine learning models and any tuned hyperparameters. Italicized words refer to parameters in the model.

### A.2.1 Logistic Regression

From the python package `sklearn`, the `SGDCClassifier` was used with *penalty* = None and *loss* = log.

### A.2.2 Decision Tree

The `DecisionTreeClassifier` from the `sklearn` package was used to train the decision tree. *max_depth* and *max_features* were tuned during cross-validation. *max_depth* was tested on values ranging from 2 to 20, whereas *max_features* ranged from 1 to 10.

### A.2.3 Random Forest

From the `sklearn` python package, the `RandomForestClassifier` was used with *n_estimators* = 1000. The other parameters were left as the default values since random forests have been shown in the literature to be insensitive to the choice of parameters Bluwstein et al., 2021.

### A.2.4 Extremely Randomized Trees

The `ExtraTreesClassifier` from the `sklearn` package was used to train the extremely randomized trees model. Similarly to random forests, we set *n_estimators* = 1000 and left the other parameters at their default value.

### A.2.5 Artificial Neural Networks

The `MLPClassifier` from the `sklearn` package was used to train the neural networks. We set *solver* = lbfgs and performed a hyperparameter search over *alpha*, *activation*, and *hidden_layer_sizes*. 25 neural networks were trained on bootstrapped samples after the training set was constructed. The prediction given by the model is the mean of all 25 neural networks.

### A.2.6 Support Vector Machines

The `SVC` from the `sklearn` package was used with hyperparameter tuning being applied to $C$ and *gamma*. Observations are upsampled in the training set of each of the 25 support vector machines being trained. The final prediction is the mean of the predictions across the 25 models.

### A.2.7 XGBoost

The `XGBClassifier` from `sklearn` was used and *learning_rate*, *max_depth*, *n_estimators*, *colsample_bytree*, and *subsample* were hypertuned using a grid search over a range of values.

### A.2.8 K-Nearest Neighbours

The `KNeighborsClassifier` from `sklearn` was used with the following parameter tuning. A grid search was done on *n_neighbors* between 3 and 10, the *weights* being either uniform or distance, the *algorithm* being auto, ball_tree, kd_tree, or brute, and the *leaf_size* being 30, 50, or 100.

## A.3 Results not reported in the main body

### A.3.1 Global variables

One limitation of this work is in the construction of global credit growth and global slope of the yield curve features. The construction of these features in this work does not give adequate weight to countries with larger GDPs than others. This method only investigates the differences between the variables and then averages them.

| | JST | Banking | Currency | Inflation |
|---|---|---|---|---|
| ANN | 0.79 | 0.80 | 0.82 | 0.82 |
| SVM | 0.82 | 0.78 | 0.71 | |
| KNN | 0.77 | 0.77 | 0.75 | 0.55 |

Table A.1: AUROC scores for ANN, SVM, KNN for the baseline cross-validation experiment

### A.3.2 Other machine learning models

Table A.1 shows the AUROC scores for the artificial neural network, support vector machine, and K-nearest neighbour models. The neural network and support vector machine required substantial time to train and obtained poorer results than the tree-based methods. Because of the mediocre performance and substantial training time, these three methods were omitted from the forecasting experiment.

After performing the Shapley regressions, we investigated whether we could obtain comparable results by training models with only the statistically significant indicators included. Table A.2 shows the results of this experiment. Indeed, we actually see improved performance for the currency and JST crisis cases, identical performance with the banking crisis, and worse performance for the inflation crisis. This gives further confidence that the Shapley regressions are indeed sorting out the important predictors from those that are not important.

| | Baseline | Important Predictors |
|---|---|---|
| JST | 0.87 | 0.89 |
| Banking | 0.86 | 0.86 |
| Currency | 0.86 | 0.88 |
| Inflation | 0.91 | 0.85 |

Table A.2: AUROC scores for extremely randomized trees models trained with only the statistically significant predictors compared to the baseline model in the cross-validation experiment

| | all | pre-WW2 | post-WW2 |
|---|---|---|---|
| GDP | 0.02 | 0.05 | 0.00 |
| CPI | 0.02 | 0.05 | 0.00 |
| Current account | 0.06 | 0.13 | 0.01 |
| Short-term rate | 0.07 | 0.15 | 0.01 |
| Long-term rate | 0.03 | 0.06 | 0.00 |
| Broad money | 0.07 | 0.12 | 0.03 |
| Credit | 0.09 | 0.20 | 0.01 |
| Public debt | 0.07 | 0.16 | 0.01 |
| Consumption | 0.06 | 0.15 | 0.00 |
| Investment | 0.08 | 0.18 | 0.01 |
| Wage | 0.02 | 0.06 | 0.00 |
| Unemployment | 0.27 | 0.62 | 0.02 |
| Corporate Debt | 0.28 | 0.60 | 0.04 |
| Business loans | 0.46 | 0.90 | 0.14 |
| Household loans | 0.43 | 0.84 | 0.14 |
| House prices | 0.24 | 0.45 | 0.08 |

Table A.3: Proportions of missing values for the predictors for the whole period, and before and after WW2

### A.3.3 Missing values

Table A.3 shows the proportions of the missing values for the variables used in this work. We see most of the missing values occur in the pre-WW2 era, with very few missing after. Most of the features with higher missing value proportions are only included in the robustness checks, with the exception of the unemployment rate.

### A.3.4 Reinhart and Rogoff Narrative crisis results

Looking at Figure 1.1, we can see that there were 2 definitions for currency crises and 2 definitions for inflation crises. Before beginning the analysis of this work, the baseline models were trained on both definitions for each type of crisis. The models achieved very similar results for both types of inflation crises. This was not the case for currency crises, however! Table A.4 shows a comparison in terms of AUROC scores for these baseline experiments. Recall that the manually constructed variables were made using

| | Currency | Currency (RR) | Inflation | Inflation (RR) |
|---|---|---|---|---|
| Extremely Randomized Trees | 0.87 | 0.71 | 0.91 | 0.90 |
| Logistic Regression | 0.65 | 0.58 | 0.84 | 0.87 |

Table A.4: AUROC scores for the baseline models using Reinhart and Rogoff narrative crisis definitions and the crises constructed using thresholds

the thresholds provided by Reinhart and Rogoff, 2009. The results may differ since the thresholds are being applied to different datasets. Since the manually constructed crisis variables obtained better baseline results, they were chosen for the main analysis in this work.

### A.3.5 Corporate debt

Another new variable introduced in the 6th version of the Macrohistory database was corporate debt. We compared a variety of the 11 features mentioned in Bluwstein et al., 2021 and added different combinations of the corporate debt, unemployment rate, and wages features. Table A.5 shows the results of these preliminary models. What we see if that the removal of the wages variable significantly decreased the performance of the models in predicting inflation crises. We also see that adding the corporate debt feature to the baseline model decreased performance across the board. When corporate debt and the unemployment rate are removed, we see increased AUROC scores for currency crises and inflation crises. Combining all of these facts, the final decision was made to only exclude the corporate debt feature from the analyses in this paper. The unemployment rate was used despite its inclusion weakening the performance of currency and inflation crises because it is well-researched in the literature. The models presented in this paper also do not aim to be the best models they can be, but rather they show the potential use of machine learning models in financial crisis prediction.

| | JST | Banking | Currency | Inflation |
|---|---|---|---|---|
| Baseline (unemployment rate and wages) | | | | |
| Extremely Randomized Trees | 0.87 | 0.86 | 0.86 | 0.91 |
| Random Forest | 0.86 | 0.84 | 0.86 | 0.90 |
| Logistic Regression | 0.82 | 0.79 | 0.65 | 0.84 |
| Baseline + Corporate Debt | | | | |
| Extremely Randomized Trees | 0.86 | 0.84 | 0.86 | 0.91 |
| Random Forest | 0.85 | 0.82 | 0.86 | 0.90 |
| Logistic Regression | 0.82 | 0.78 | 0.65 | 0.83 |
| Baseline - Unemployment | | | | |
| Extremely Randomized Trees | 0.87 | 0.83 | 0.88 | 0.92 |
| Random Forest | 0.85 | 0.81 | 0.87 | 0.91 |
| Logistic Regression | 0.81 | 0.79 | 0.64 | 0.89 |
| Baseline - Unemployment - Wages | | | | |
| Extremely Randomized Trees | 0.86 | 0.84 | 0.88 | 0.71 |
| Random Forest | 0.85 | 0.82 | 0.87 | 0.69 |
| Logistic Regression | 0.81 | 0.79 | 0.65 | 0.61 |
| Averages | | | | |
| Extremely Randomized Trees | 0.865 | 0.843 | 0.870 | 0.863 |
| Random Forest | 0.853 | 0.823 | 0.865 | 0.850 |
| Logistic Regression | 0.815 | 0.788 | 0.648 | 0.793 |

Table A.5: AUROC scores for the baseline models using different combinations of the new features in the sixth version of the Macrohistory database