INTERNATIONAL MONETARY FUND

Understanding and Predicting Systemic Corporate Distress: A Machine-Learning Approach

Burcu Hacibedel and Ritong Qu

WP/22/153

2022
JUL

WORKING PAPER

 WP/22/153

IMF Working Paper
Strategy, Policy and Review Department

Understanding and Predicting Systemic Corporate Distress: A Machine-Learning Approach
Prepared by Burcu Hacibedel and Ritong Qu*

Authorized for distribution by Martin Cihak and Daria Zakharova
July 2022

ABSTRACT:
In this paper, we study systemic non-financial corporate sector distress using firm-level probabilities of default (PD), covering 55 economies, and spanning the last three decades. Systemic corporate distress is identified by elevated PDs across a large portion of the firms in an economy. A machine-learning based early warning system is constructed to predict the onset of distress in one year's time. Our results show that credit expansion, monetary policy tightening, overvalued stock prices, and debt-linked balance-sheet weaknesses predict corporate distress. We also find that systemic corporate distress events are associated with contractions in GDP and credit growth in advanced and emerging markets at different degrees and milder than financial crises.

JEL Classification Numbers:	C40, E44, G01, G17, G21, G33
Keywords:	Nonfinancial sector; Probability of default; Early warning systems; Macroprudential policy
Author's E-Mail Address:	bhacibedel@imf.org, rqu@imf.org

* The views expressed in this paper are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management. We would want to thank Bruno Albuquerque, Jorge Antonio Chan-Lau, Sophia Chen, Salih Fendoglu, Marco Gross, Weining Xing and seminar participants at the IMF for helpful comments; Chuqiao Bi, Ruofei Hu, Roshan Iyer and Jose Marzluf for excellent research assistance. Special thanks to Bruno Albuquerque for helping us understand Compustat data. All the errors are our own.

Contents

1 Introduction

The Covid crisis and resulting vulnerabilities brought the corporate sector to the forefront of policy debate. Many governments supported corporations through monetary, fiscal and financial policies such as low interest rates, grants, and debt moratoria. Corporations are coming out of the pandemic with higher debt, lower profitability, weaker balance sheets and lower cash buffers (IMF, 2021). In the post-Covid period, the withdrawal of support measures could increase the risk of corporate distress, potentially leading to systemic crises. In this context, the ability to predict corporate distress and understand its macroeconomic consequences is key. Timely detection of the sources of corporate vulnerability beyond general indebtedness enables a more targeted policy choice.

This paper aims to provide an early warning system to forecast corporate distress events to inform timely policy making. We first identify these events using a novel measure and definition based on firm-level probabilities of default (PD) covering the last three decades 1995:2021. Our new measure allows us to take a deeper look at how the corporate sector and economic indicators behave before corporate distress. We build an ensemble of machine-learning models drawing on a set of macroeconomic and balance-sheet variables to predict the onset of systemic corporate distress over a four-quarter horizon. Rather than selecting the best model in a horse race, we take an agnostic view that none of each individual model is the true model, and try to approximate the true model using the combination of individual models (Geweke and Amisano, 2011). Our model not only predicts the approaching distress, but also offers hints about the sources of corporate vulnerability. Our results show that weak balance sheets and global financial conditions play a large role in predicting corporate distress. These allude to the importance of timely use of monetary and macroprudential policies.

There are several challenges to empirical studies of corporate distress. Existing literature lacks a consistent definition of corporate crises or distress. While these two terms are interchangeable, we will use the term "corporate distress" in the remainder of this paper. The lack of cross-country and longitudinal data on corporate distress also makes it hard to analyze cause and aftermath of systemic nonfinancial corporate sector distress. Existing studies are mostly single-country or single-crisis episode focused such as Japan in the 1990s (Caballero et al., 2008), Europe in the 2010s

(Schivardi et al., 2022; Acharya et al., 2020), US (Giesecke et al., 2014), and the issue of corporate crises has not been studied in a long-run cross-country setting so far. With our new definition, we are able to overcome this roadblock. Another major challenge is the intertwined nature of corporate distress with other economic crises. It is hard to find a corporate distress example that is not preceded or does not overlap with crises of other nature such as banking, sovereign debt, and currency crises. In our paper, we also suggest ways to overcome this difficulty, and conduct robustness tests accordingly. One caveat of our data is that it covers listed firms only.

Corporate distress has been mostly linked to credit booms and increasing leverage (Jordà et al., 2020; Müller and Verner, 2021; Lian and Ma, 2021)with significant macroeconomic and microeconomic implications. While there is some consensus on the relatively benign impact of credit distress on the economy compared to financial crises (Giesecke et al., 2014), the type of credit matters. Credit booms driven by corporate and by households affect the economy differently. Household credit driven booms have been shown to have more significant and long lasting impact on GDP growth (Jordà et al., 2013; Mian et al., 2017). The underlying credit and credit dynamics matter for the macro impact. Studies focusing on the microeconomic dynamics and impact focus on the debt overhang and zombification that follow credit booms leading to lower investment by firms (Andrews and Petroulakis, 2019; Gourinchas et al., 2020; Albuquerque, 2021).

To measure corporate distress, existing studies use definitions based on corporate credit growth (Jordà et al., 2020), actual defaults (Giesecke et al., 2014) and distance-to-insolvency (Atkeson et al., 2017)with most in single-country settings. To our knowledge, we are the first paper to use a default probability based measure. The model-based PD uses balance-sheet variables, macro factors and distance-to-default (Merton, 1974) as predictors. Hence, the PD is a comprehensive and timely measure of a firm's difficulty in operation, liquidity, and investors' perception of its underlying risk. Our PD dataset coming from Corporate Research Initiative of National University of Singapore covers more than 60,000 publicly listed firms from 88 economies.

We construct economy-level PD indices by capital-weighted-average of firm-level PDs. Periods of systemic corporate sector distress are characterized by persistently elevated PD indices and identified by a Markov regime-switching model (Hamilton, 1989). We find that economies experienced corporate distress 18 percent of the time on average over 1995:2021. Many, but not all corporate distress events, overlap with

documented financial, sovereign debt and currency crises. Notably, we observe a wave of corporate distress in many economies during the early 2000s that coincides with the Dot-com bubble, which cannot be attributed to other types of crises.

Our paper contributes to the existing literature in several areas. First, we propose a new definition for corporate distress and provide a novel database of corporate distress events covering the last three decades and 55 advanced and emerging markets. Economies experienced corporate distress 18 percent of the time on average over 1995:2021. However, not all corporate distress events lead to systemic financial crises. Secondly, we find that systemic corporate distress has implications for GDP and credit growth. Our results show that around corporate distress, GDP growth slows down in both AEs and EMs while credit growth slows down significantly only in EMs. Thirdly, we construct a machine-learning based model to forecast systemic corporate distress with a forecast horizon of 4 quarters. This allows us to identify indicators that signal increasing risk of crisis. In addition to over-heating in credit markets, the model attribution analysis shows that funding costs, balance sheet vulnerabilities and market overvaluation are also informative about coming corporate crises.

The rest of the paper is structured as follows: Section 2 presents the data and descriptive statistics. Section 3 defines and identifies corporate distress events. Section 4 explains the machine learning based early warning model and examines the precursors of systemic corporate distress. Section 5 analyzes the contemporary macroeconomic impact of corporate distress. Section 6 focuses on the policy implications and concludes.

2 Data

We construct our data focusing on two groups: economy-level PD indices and a set of variables that can predict systemic corporate distress. PD data helps us construct a novel crisis series to overcome the lack of a broad-based definition of systemic corporate distress. Then, we select predictor variables to proxy for four group of indicators closely linked to corporate distress: firm-level balance sheet variables, international macro variables, domestic macro variables and financial market valuation variables. while we start with a larger set of variables, we eliminate a number of these while building up our model if not significant. Once we have the crisis/distress dates, we are able to check for the signaling power of the predictor variables.

2.1 Constructing Economy-level PD Indices

The quarterly PD data is obtained from the Credit Research Initiative database of the National University of Singapore (NUS-CRI PD, henceforth)[1]. The probability of a firm defaulting on its debt encapsulates the firm's difficulty in operation, its liquidity and investors' perception of its underlying risk. Hence, the probability of default (PD) is an apt proxy for corporate distress. PDs are derived from a reduced form model (Duan et al., 2012) that draws on both market-based, such as distance-to-default (Merton, 1974), and accounting-based firm-specific attributes as well as macro-financial factors including the stock returns, cash-to-asset ratio, current assets-to-current liabilities ratio, net-income-to-total-asset ratio, relative market cap and relative market-to-book ratio.The model performs well especially in shorter horizons, achieving an accuracy ratio of 80% in 12 month forecasts. The data set covers the daily probability of default of firm-level corporate bonds with maturities up to 5 years. We focus on the default probability of 12-month corporate bonds (excluding the financial sector),

We construct quarterly economy-level PD indices using capital-weighted averages of firm-level PD at the end of each quarter. At the beginning of each year, the firm-level capital is calculated as the product of stock prices and common shares outstanding from Compustat Global. The NUS-CRI PD is matched with Compustat Global using ISIN. Among the 2,749,611 firm-quarter observations, 2,367,880 are matched with the firm capital data. Missing capital values are imputed with the median of other firms' capital values in the same sector, quarter and from the same economy. After the imputation, the missing capital values are further imputed with the median of other firms' capital values in the same quarter and from the same economy. To make sure the indices are not biased by any limited coverage of firms, we impose a minimum of 20 firms for each economy-quarter observation. Table 1 shows the summary statistics of PD indices for each economy. After discarding economy-quarter observations with insufficient firm coverage, the resulting economy-level average PD indices covers 61,960 firms from 88 economies. The earliest data starts in March 1990. For the purposes of this paper, our sample ends in the last quarter of 2021.

[1]See The Credit Research Initiative of the National University of Singapore (2019) for a detailed description of the methodology.

2.2 Predictors of Systemic Corporate Distress

We draw on the literature to select a total of 43 predictors covering domestic and international macro economic variables, firm balance sheet variables, lagged PDs and variables derived from stock prices. Table 2 shows the full list of variables employed.

Balance sheet Variables

A vast literature, e.g., Altman (1968), Ohlson (1980), Shumway (2001) and Campbell et al. (2008), has shown that firm-level balance-sheet variables can predict corporate defaults. We include the capital expenditure-to-asset ratio, cash-to-asset ratio, debt-to-asset ratio, interest-coverage ratio, net debt-to-asset ratio, return on assets and short-term investment-to-liability ratio. We also include the 12-month and 36-month probabilities of default from NUS-CRI. To allow for economy-specific steady state of PDs when identifying distress periods, we scale the PD series of each economy to unit variance before feeding into the early warning system. In addition to the levels of balance-sheet variables, we also include their quarterly changes as predictors. The balance-sheet variables are obtained from Compustat Global. The details about how the ratios are computed are provided in Appendix B.

Macroeconomic Variables

Besides the accounting ratios, Carling et al. (2007), Duffie et al. (2007) and Koopman et al. (2012) show that domestic macroeconomic variables can predict corporate defaults. Pesaran et al. (2006) find global macroeconomic variables also affect default probabilities. We have a total of 11 macroeconomic variables covering various aspects of the business cycle, credit and external sector. Among these, Fed Funds Shadow Rate[2] and USD appreciation are used as common predictors for corporate distress events across all economies, reflecting the global financial cycle (Rey, 2015) and conditions. The list of macroeconomic variables can be found in Panel B of Table 2.

[2]The Fed Funds Shadow Rate is from Federal Reserve Bank of Atlanta using the method of Wu and Xia (2016).

Market Valuation

Variables related to stock prices are often used to predict defaults, especially when excessive risk taking and asset bubbles. For example, Campbell et al. (2008) and Duffie et al. (2007) use returns of the market index as predictors. To reflect this, we include return volatility, dividend yields and price-earnings ratios because overvalued stock prices could reflect overheating, mispricing and asset bubble risks in financial markets

Missing Predictors

Given that we cover a large set of predictors, the issue of missing predictors is inevitable, especially for emerging markets. Figure 6 shows a heat-diagram of predictor availability over time.The darker the shading is, the larger is the number of economies for which a specific variable is missing in a given year. The color scale is shown to the right of each figure. The number is defined as the ratio of the number of economies for which each predictor is available over the total number of economies. Our sample starts from 1995Q1. Most of balance sheet variables become broadly available from 1996 onwards.

In the early-warning system for corporate distress introduced in Section 4, missing predictors are imputed by the sample median when forecasting. To avoid any biased output, we only include economy-quarter observations where more than two-thirds of the predictors are available. After imposing the cutoff, we end up with 55 economies in our early-warning system.

3 Identifying Corporate Sector Distress

The economy-level PD indices constructed in Section 2 serve as proxies for systemic corporate distress in an economy. They align well with corporate distress periods documented in the literature. The top panel in Figure 1 shows that advanced economies experienced high level of PDs during the burst of the dot-com bubble in the early 2000s and during the Global Financial Crisis (GFC) in 2008. In addition to these two episodes of high corporate sector stress, EMs also experienced high PDs during the Asian Financial crisis in late 1990s. As documented in Das et al. (2007), corporate defaults happen in time clusters which implies cyclical waves of economy-level PD

indices. This is exactly what we find in Figure 2 of PD indices for selected economies.

3.1 A Markov-Switching Model for PD Indices

To identify the corporate distress events, we construct a Markov Switching model as characterized by persistently high PDs. Average PD indices across economy blocks (the top panel of Figure 1), and individual economy PD (Figure 2) indices show that the corporate sector is subject to infrequent regimes of high and volatile default probabilities. The Markov Regime-switching model (Hamilton, 1989) is apt for identifying states of high risk, characterized by persistently high PD indices. One challenge is that corporate sector distress periods have few observations for each individual economy, which can lead to large estimation errors. To address the issue, we pool model parameters across different economies to take advantage of the cross-sectional dimension of the data: we can borrow from other economies' experience to estimate each economy's parameters. Pooling the parameters also gives a consistent definition of corporate sector distress across economies. We assume the ratio of mean PD in high risk regime to mean PD in low risk regime to be the same across economies. We make the similar assumption on volatilities. On the other hand, the median level of PD indices are quite dispersed among economies as shown in Table 1, possibly due to different industry compositions and legal restrictions pertinent to defaults. For example, the calm period in Argentina's corporate sector can have a higher expected PD than the high-risk regime of Switzerland. To account for differences in steady states of PDs across economies, we set economy-specific mean and volatility of PDs in low-risk regimes. and set the ratio of the high-risk regime's mean and volatility of PD over those of low-risk regime's to common parameter across economies.

Before specifying the model, we introduce some notations. Let i denote the economy i, $(i \in \{1, 2, ..., N\})$ and t denote period t $(t \in \{1, 2, ..., T\})$. Let $S_{it} \in \{0, 1\}$ index two regimes. S_{it} is driven by a Markov Regime-Switching model:

$$\text{Prob}(S_{it} = m | S_{it-1} = n) = p_{mn}. \tag{1}$$

Under each regime, the dynamics PD_{it} is different in terms of mean and volatility:

$$PD_{it} - (1 + \delta S_{it})\,\mu_{iL} = \rho\,(PD_{it-1} - (1 + \delta S_{it-1})\,\mu_{iL}) + (1 + \gamma S_{it})\,\sigma_{iL}\varepsilon_{it}. \tag{2}$$

μ_{iL} is the unconditional mean of PD_{it} in the low default probability regime. $(1+\delta)$ is the ratio of high-risk regimes' unconditional mean over those of low-risk regimes. We impose the constraint that δ is positive. Similarly, σ_{iL} is the volatility of PD_{it}'s innovations in the low default probability regime. $(1+\delta)$ is the ratio of high-risk regimes' volatility to those of low-risk regimes. We impose the constraint of $1+\gamma > 0$ to make sure the volatility is positive. Parameters δ, γ and ρ are the same across different time series. By using same parameters, δ and γ, we insure crises are identified based on same criteria across economies. μ_{iL} and σ_{iL} are economy-specific to account for economy-specific factors that affect its mean and volatility, as demonstrated in Table 1. Given the proliferation of parameters in a model of many economies, maximum likelihood estimates are computationally difficult, so we estimate the model using Bayesian approach. The MCMC algorithm is elaborated in Appendix A.

3.2 Model Estimates and Periods of High Corporate Sector Distress

Both the mean and the volatility of PDs in the high-PD regime are considerably larger than the ones in the low-PD regime. Table 3 reports the estimates of key parameters in the Markov Regime-switching model. The mean of PDs during the high-PD regime is around 3.7 times the ones in low-PD regime. The volatility of innovation in PDs is also higher in the high-PD regime (around 5.9 times) than in the low-PD regime.

The posterior probability of the corporate distress regime is presented as a heat map in Panel (a) of Figure 3. We use a threshold of 50% on the posterior probability of a high-PD regime to identify the distress periods, as shown in Panel (b) of Figure 3. Darker color indicates high posterior probability. We only include economies with more than 10 years of observations when estimating the model. After imposing the minimum sample length cutoff, we end up with 66 economies. The full list of corporate distress periods is in Table A1 in the Appendix, and distress periods with higher posterior probabilities are also shown here using a different color shading.

With our proposed definition of corporate distress, we can capture several major corporate distress events documented in the literature. These include the 1995 Mexico crisis, 1997 Asian crisis, 2000-01 dot-com bubble/burst, 2012 European Debt crisis, and 2007-08 Global Financial Crisis. Covid-19 also caused high corporate distress in

many economies starting in 2020. Starting from the late 1990s, both the advanced economies and emerging markets went through prolonged periods of high corporate sector distress that ended in the early 2000s. In total, we are able to identify 193 episodes of distress events, and economies experience corporate distress in 18% of the time on average.

Considering the attention banking crises received in the literature, one natural question is how corporate distress events overlap or differ from banking crises (Giesecke et al., 2014). Our analysis shows that while many high corporate distress periods overlap with banking crises, we also identify several episodes of high corporate distress with no overlapping banking crisis. Figure 4 shows corporate distress periods vs banking crises periods. Corporate distress periods are marked in green; banking crises are in blue; the overlapping years are in red. Banking crisis identification is borrowed from Laeven and Valencia (2020) with the sample ending in 2017. Notably, we capture several high corporate distress periods in the early 2000s (the dot-com bubble) and in the early 2010s (the European debt crisis) with no simultaneous banking crises. Figure 5 also shows corporate distress periods vs external crises periods. Most external crises are covered by corporate distress periods.

4 An Early-warning System for Corporate Distress

For our early warning system, we build a machine-learning model to predict the onset of corporate distress periods within a one-year horizon, using the corporate distress definition in Section 3 and the predictors in Section 2.

Essentially, any sample observation can be classified into three categories: pre-distress periods, distress periods and calm periods. Because the early-warning system aims to predict corporate distress in advance, we confine ourselves to differentiating pre-distress periods from calm periods, and discard the distress periods when estimating and evaluating the model. Let C_{it} be an indicator function that equals to 1 when a corporate sector distress event in economy i starts at time t. Let $C_{it+1,t+h}$ be an indicator function that equals to 1 if a distress event starts between $t+1$ and $t+h$:

$$C_{it+1,t+h} = \mathbb{1}_{\{\forall C_{is}=1, s\in[t+1,t+h]\}}. \tag{3}$$

The output of the early-warning system is a probability of a corporate distress event that happens in economy i and starts in the window between $t+1$ and $t+h$, using predictors at time t,

$$\text{Prob}\left(C_{it+1,t+h}=1|x_{it}\right)=f\left(x_{it}\right). \tag{4}$$

The function f in Eq. (4) can be approximated by many functional forms, from linear logit models to more flexible machine learning methods. We estimate five models of distinctive functional form including Logit Lasso, linear discriminant analysis (LDA), support vector machine (SVM), random forest (RF) and extreme gradient boost classifier (XGBoost). These methods are decsribed in detail in the Appendix C.

4.1 Model Combination

Rather than selecting one single model, we combine the five models to proximate the true data generating process using the optimal pooling approach in Geweke and Amisano (2011). The method was shown to outperform Bayes model averaging in out-of-sample forecasting by making a more general and realistic assumption that none of the candidate models is the true form of f[3]. Let $\hat{f}_1$, $\hat{f}_2$, ..., $\hat{f}_M$ be estimation of f from M models. Define the final model as weighted average of individual models:

$$\hat{f}\left(x\right)=\sum_{m=1}^{M}w_m\hat{f}_m\left(x\right), \tag{5}$$

$$\sum_{m=1}^{M}w_m=1,\ 0\leq w_m\leq 1, \tag{6}$$

where w_m is the weight attached to the mth model. The optimal weights would be the ones that minimize the cross-entropy of the combined model to the true data generating process. Empirically, we estimate weights by maximizing the likelihood function

$$w^*=\arg\max_{w}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\{\log\sum_{m=1}^{M}w_m\left[C_{it+1,t+h}\hat{f}_m\left(x_{it}\right)+\left(1-C_{it+1,t+h}\right)\left(1-\hat{f}_m\left(x_{it}\right)\right)\right]\right\}. \tag{7}$$

[3]For an introduction to Bayes model averaging , see, e.g., Raftery (1995) and Hoeting et al. (1999). See Gross and Población (2019) for an application of Bayes model averaging to bank stress testing.

Given that we aim to minimize the expected cross-entropy out of sample, a caveat is that x_{it} should not be included in the sample used to estimate $\hat{f}_m$ when computing $\hat{f}_m(x_{it})$ in equation Eq. (7). We address this issue using cross-validation that is adapted to the case of panel data at the presence of time series and cross-section dependence. The next subsection elaborates on the cross-validation method we used.

4.2 H-block Cross-validation

The panel data for pre-distress periods exhibits cross-sectional dependence as is evident in the clustering of economies experiencing corporate distress shown in Panel (b) of Figure 1 . By the definition of pre-distress periods, the target variable $C_{it+1,t+h}$ is also auto-correlated in the time dimension, because the forecast horizon is longer than one quarter. This violates the independence assumption underlying traditional leave-k-out cross-validation. We adapt the block cross-validation method based on Burman et al. (1994) and Racine (2000) to address the issue. The idea is to break linkages between training and validation sets by putting blocks of sample observed at consecutive time periods into the same validation set and leaving time gaps between training and validation sets.

For a K fold cross-validation, let $(B_{k,train}, B_{k,test})$ be the kth set of training and validation set, $k = 1, 2, ..., K$. Let $\mathcal{T} = \{1, 2, ...T\}$ be the set of time in the sample. $\mathcal{T}$ is divided into K blocks of consecutive time blocks:

$$\mathcal{T}_k = \{t | (k-1) \lfloor T/K \rfloor + 1 \leq t \leq k \lfloor T/K \rfloor\} \text{ for } 1 \leq k < K, \tag{8}$$

$$\mathcal{T}_K = \{t | (K-1) \lfloor T/K \rfloor + 1 \leq t \leq T\}. \tag{9}$$

$B_{k,test}$ is defined as

$$B_{k,test} = \{(i,t) | i = 1, ..., N;\ t \in \mathcal{T}_k\}. \tag{10}$$

$B_{k,train}$ is defined as

$$B_{k,train} = \{(i,t) | i = 1, ..., N;\ t < \min \mathcal{T}_k - h \text{ or } t > \max \mathcal{T}_k + 2h\}, \tag{11}$$

where we leave a gap of h before the oldest observation in $B_{k,test}$ to account for the fact that $C_{it+1,t+h}$ is not known until $t+h$; and a gap of $2h$ after the latest observation in $B_{k,test}$ to further decrease the time-series dependence between the training and the

validation set.

Our proposed cross-validation method is used for the selection for model weights and hyper-parameters of machine-learning models. We focus on a forecast horizon of 4 quarters. The hyper-parameters involved in each model is selected to maximize cross-validation AUROC, using data up to 2005Q1. After selection of hyper-parameters, the model weights are estimated with formula adapted from Eq. (7):

$$w^* = \arg\max_{w} \sum_{k=1}^{5} \sum_{(i,t)\in B_{k,test}} \left\{ \log \sum_{m=1}^{M} w_m \left[C_{it+1,t+h} \hat{f}_m^{(k)}(x_{it}) + (1 - C_{it+1,t+h}) \left(1 - \hat{f}_m^{(k)}(x_{it})\right) \right] \right\}, \tag{12}$$

where we use a 5 fold cross-validation on predictors before 2004Q1, and $\hat{f}_m^{(k)}$ is model m estimated with data in $B_{k,train}$. The combined model is the sum of 76% of XGBoost, 5% of Logit regression and 19% of LDA. The SVM and Random Forest have zero weights. We kept the hyper-parameters and model weights fixed after selecting them based on predictors before 2004Q1.We discuss how we estimate and evaluate the model in the next subsection.

4.3 Model Performance

We adopt two approaches to simulate the model's performance. The first is "back-testing": we make predictions with the data available in each time period, and re-cursively re-estimate the model as new data arrives. The out-of-sample approach addresses the question about how the model would perform, if it were to be deployed in the past. A drawback of the out-of-sample approach is that it fails to account for the fact that the model estimation error decreasing with the sample size, especially for flexible machine learning models. To evaluate the expected performance of the model estimated from the full sample, we use the h-block cross-validation such that the training data covers most of the full sample. We measure the model performance with log-likelihood and the Area Under the Receiver Operating Characteristic[4] Curve (AUROC).

The out-of-sample exercise starts in 2005Q1. To avoid any forward-looking bias, we make sure the training data only uses observations available before the testing

[4]An receiver operating characteristic curve is a curve showing false positive rates and true positive rates of a classification model at all classification thresholds. AUC higher than 0.5 indicates the model is more informative than white noise.

data. Given the definition of $C_{it+1,t+h}$, it means the predictors in the training data are at least 4 quarters behind the corresponding testing set, and the first training-set uses the predictors before 2004Q1. Each model is re-estimated in Q1 of each year. Missing predictors are imputed using sample median, when estimating the model and making forecasts[5]. We restrict to observations with no more than 10 missing predictors, such that imputations will not significantly bias model estimates and predictions. Forecasts are generated conditional on predictors in each quarter of the year, using the model estimated at Q1 of the year.

The top panel of Table 4 shows the out-of-sample AUROC of each sub model and the combined model. Each row shows the AUROC based on groups of all economies, the advanced economies and emerging markets. The combined model generates an out-of-sample AUROC of 0.67 which is highest across all sub models. The AUROC computed for AEs and EMs yields similar results, suggesting the combined model is robust across different groups. The bottom panel of Table 4 shows the out-of-sample log-likelihood of each sub model and the combined model. The combined model has the highest likelihood, which validates the Geweke and Amisano (2011)'s assumption that none of the sub-models is the true data-generating process. By combining sub models, we ends up with a better model.

To evaluate the expected performance of models estimated from the whole sample, we compute AUROC and log-likelihood using a 10-fold cross-validation on the whole sample. The result presented in Table 5 is qualitatively similar to Table 4. The combined model yields the highest AUROC and log-likelihood relative the sub-models. The cross-validation AUROC of the combined model is 0.71 which is higher than the out-of-sample AUROC. The improvement may be attributed to reduced estimation errors, as larger sample is used to estimate models.

We further check the robustness of the model by looking at the AUROC computed with forecasts in different time blocks. Figure 7 plots the AUROC of models in time blocks: 1995-1999, 2000-2004, 2005-2009, 2010-2014 and 2015-2020. With the exception of the period 1995-1999, the combined model has an AUROC above 0.6 in all time blocks. The under performance in the period 1995-1999 is possibly due to the many missing predictors in the earlier sample period, which makes reduce AUROC of four out of five sub models below 0.55.

[5]Exceptions are XGBoost and LDA: the XGBoost package handles missing data internally, while LDA generates probability conditional on the predictors available.

4.4 Interpreting Model Predictions with Shapley Values

One caveat and noteworthy cost of flexible machine learning models is the loss of interpretability relative to linear models. Interpretability is important, because the model output should, by no means, be taken as the conclusion and applied to all circumstances in the future. Our model is simply a statistical extrapolation of past corporate distress events with a set of predictors. Any future corporate sector distress may have a different nature from past ones, and policy makers may need a broader set of predictors. By interpreting how the model arrives at its prediction, policy makers can evaluate the model output and use discretion.

In this section, we use Shapley values to attribute the output of the model to each predictor. The Shapley values use classical equations from cooperative game theory to compute explanations of model predictions (Shapley, 1953). It is widely used in model explanation because of the additive feature that the sum of each predictor's Shapley value plus a constant is the model prediction. We use the SHAP Python package based on the algorithm in Lundberg and Lee (2017) to compute Shapley values.

The top-five contributors are the Fed Funds shadow rate, 12-month default probability, policy rate, dollar annual appreciation and market index return of the past year, as shown in the left panel of Figure 8 which presents the mean absolute Shapley values of the top 20 contributors. Among these, the dividend yield, change in cash-to-asset ratio, capital expenditure to asset ratio, change in return on assets, and interest coverage ratio contribute negatively to the crisis probability. In other words, when these values are negative, they contribute positively to the probability of corporate distress. The remaining predictor variables increase the corporate distress risk as their values increase.

Individual Shapley values of predictors are consistent with our priors. The right panel of Figure 8 plots the distribution of Shapley values. Each dot represents a Shapley value from one observation. The color represents the level of the corresponding predictors. The dots are jittered to reflect the distribution of Shapley values. Hence, a distribution of Shapley values with red dots on the right and blue dots on the left suggests higher predictor values have positive impact on the outcome. Tight financial conditions increase the risk of corporate distress. The Fed Funds shadow rate proxies for global financial conditions and global financial cycle a la Rey (2015) and comes out as the most powerful predictor variable in our model.The policy rate variable proxies

for local financial conditions and their tightening when increasing. Our results also show that the capital expenditure and its change among the top predictors implying that when firms' capital expenditure is high and increasing rapidly, it increase the likelihood of corporate distress. Additionally, over-valued stock prices can be a harbinger of subsequent drastic corrections and corporate distress as reflected by the distributions of previous market index growth, the price-earnings ratio, market-to-book value and dividend yield. Traditional measures of balance-sheet vulnerability have the right sign in predicting corporate distress: high net debt to asset ratio, high capital expenditure to asset ratio, low cash to asset ratio and low interest coverage ratio increase risks of corporate distress.

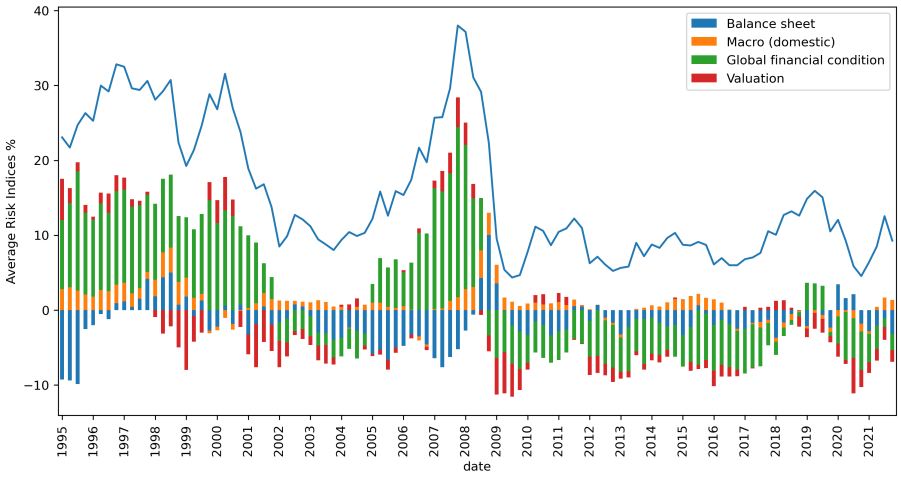
It is also informative to examine how variables in each category contribute to the predictions, because many individual variables each with a small contribution can have large combined effects. We divide the predictors into four classes: firm-level balance-sheet variables, market valuation, domestic macroeconomic variables and global financial conditions as listed in Table 2. The global financial condition category includes the USD appreciation and Fed Funds shadow rate. The Shapley value of each category is the summation of Shapley values of individual predictors in the category. Figure 9 presents the mean absolute values of predictors from each category. The top two categories are the global financial condition variables and balance-sheet variables which have similar contributions. These imply that firms' financial health and indebtedness have the highest signaling value for a looming systemic corporate distress. (Unexpected) Changes in global financial conditions, i.e., interest rates and other push factors could trigger systemic bankruptcies in the corporate sector considering firms' exposure to international economic developments. The third and fourth are valuation variables and domestic macroeconomic variables whose respective contributions are about half of the top two. The deviation from the fundamentals and asset price bubbles also collectively signal and precede corporate distress albeit to a lesser degree. Lastly, domestic macro variables such as inflation, GDP slowdowns, domestic financial cycles and banking sector exposure to firms are included in the fourth group of variables, indicating their relatively lower power collectively as a group.

4.4.1 An Application: Corporate Distress Risk Index for Emerging Markets

We use the average model results for emerging markets to illustrate an application and how our model works in terms of measuring corporate distress probability over time. The results are forward looking with a four-quarter horizon, i.e., the risk index at end-2021 shows the probability of corporate distress in 2022. Starting in 2005, the evolution of the risk index, as illustrated in the solid blue line in Figure 10, captures the key past corporate crises in EMs, namely, the 2007-08 Global Financial Crisis and 2020 Covid shock.

The contribution of each of the four variable categories can also be seen in Figure 10. The gap between the total risk index and the aggregate Shapley values of variables is due to the constant in the model that is the same for all economies in each quarter. The breakdown of the risk index shows that ahead of the GFC, the risk index spiked due loose global financial conditions and market valuation anomalies. At the onset of the GFC, which stemmed from AEs, EM corporates suffered from increasing balance sheet vulnerabilities and higher default probabilities. The most recent surge in the risk index 2020Q1 shows increasing corporate distress on the back of corporates' balance sheet vulnerabilities and tight global financial conditions. the risk subsides in the following quarters with loose global financial conditions. In 2022Q2, we observe an increase in systemic corporate distress probability with the tightening of global financial conditions and the presence of balance sheet vulnerabilities.

Figure 10: **Shapley Value Decomposition of Average EM Indices**



5 Macroeconomic Implications of Systemic Corporate Distress: Initial Findings

As the final part of our empirical analysis, we study the macroeconomic implications of systemic corporate distress with the aim to shed light on our future research. We examine how key macroeconomic variables behave around corporate distress. To this end, we focus on GDP growth, bank credit to nonfinancial corporations (NFCs), foreign direct investment (FDI) and exchange rates and compare around the corporate distress episodes identified in our model. Utilizing a panel of advanced economies and emerging markets data, we report the preliminary findings on the real effect of high corporate distress in different economy groups.

To see the general effects of high corporate distress, we regress the posterior probability of high stress regimes on annualized GDP growth and credit growth with economy-level fixed effects.

$$Y_{it} = c_{it} + \beta_{unconditional}\text{Prob}_{it} + \varepsilon_{it}. \tag{13}$$

The model is at quarterly frequency. The first row of Table 4 shows the estimates: on average, GDP growth during high corporate distress periods is lower by 3.0% relative to low stress periods. The unconditional effect on credit growth is -2.1% and marginally significant at the 10% level. The second and third rows illustrate the effects on AEs and EMs respectively by estimating the model with fixed effects:

$$Y_{it} = c_{it} + \left[\beta_{AE} I_{AE}(i) + \beta_{EM}\left(1 - I_{AE}(i)\right)\right]\text{Prob}_{it} + \varepsilon_{it}, \tag{14}$$

where $I_{AE}(i)$ is 1 for AEs and 0 for EMs. The estimates shows high corporate distress regime significantly reduces GDP growth by 2.4% and 3.9% for AEs and EMs respectively. In terms of credit growth, high corporate distress regimes don't have significant effect on AEs, but the effect is strong on EMs: credit growth is reduced by 6.5%. Hence, corporate distress has significant negative effect on GDP growth of both AEs and EMs. But the effect on credit is limited to EMs.

In comparison to financial/banking crises, the evidence on the macroeconomic consequences of corporate crises is scarce in the literature. Empirically, financial crises induce more severe disruptions to economic growth than typical recessions as

demonstrated by Schularick and Taylor (2012) and Claessens et al. (2012). Related to the corporate sector distress, Giesecke et al. (2014) study the macroeconomic effect of corporate defaults and find that corporate default crises have far fewer impacts than banking crises. Considering that our high corporate distress periods overlap with many banking crises (Figure 4), our results might be biased. To overcome this, we estimate the conditional growth of GDP and credit during only high corporate distress but non-banking-crisis periods, and also estimate the effect of banking crises jointly using the following model:

$$Y_{it} = c_{it} + \beta_{Nonfinancial}\text{Dummy}_{Nonfinancial,it} + \beta_{Financial}\text{Dummy}_{Financial,it} + \varepsilon_{it}, \quad (15)$$

where $\text{Dummy}_{Nonfinancial,it}$ is an indicator function for high-corporate-stress periods, but set to 0 when the is a concurrent banking crisis. The fourth and fifth rows of Table 6 present the estimates. Both our results and Giesecke et al. (2014) suggest that corporate distress has a milder effect on GDP growth than banking crises. We find that GDP growth decreases by 1.6% during corporate distress versus by 3.8% during banking crises. This can be attributed to the credit channel. During banking crises, credit growth decreases by around 5.6% while the effect of corporate distress on credit is insignificant. Contrary to Giesecke et al. (2014)'s finding that corporate defaults do not affect growth in the US, we do find that corporate sector distress has a significant negative impact on growth. The difference can be explained by the cross-economy nature of our analysis, where the results come from a panel of 55 economies. It can also be attributed to our forward-looking measure: we use default probabilities derived from Merton's distance to default, firm-level solvency and liquidity measures as a proxy for corporate sector stress, which is more forward-looking than the actual default data.

We further examine the impact of high PDs during financial crises and high PDs without financial crises (i.e. pure corporate distress) by conducting the above analysis for AEs and EMs blocks separately. Rows 6-10 in Table 6 present the results. Consistent with the previous estimates, pure corporate distress has a milder impact than banking crises in terms of GDP growth: high PD regimes reduce GDP growth by 1.6% for AEs and 2.2% for EMs. The difference in impact is more prominent for credit growth during pure corporate distress periods: high PD regimes increase credit growth by 0.8% for AEs, though the impact is not statistically significant.

However, high PD regimes reduce credit by 3.8% for EMs. Hence, the credit channel is significant for EMs even without concurrent bank crises.

One possible explanation for the sharp decrease in credit growth in EMs could the decrease in capital flows during corporate distress periods. While the direction of causality is not clear, capital inflows to EMs increase the supply of credit in the economy leading to credit booms. However, if either due to corporate sector distress or due to exogenous global factors, any sudden stops in capital inflows would have a significant impact on credit. Otherwise, corporate distress might increase international investors' risk aversion. To shed some light on this, we re-estimate Eq. (13), (14) and (15) with exchange rate growth and foreign direct investment growth as dependent variables. Table 7 shows the results. The second and third rows in Table 7 show that high corporate distress regimes are linked to currency depreciation in EMs of about 8.8%, and decrease in FDI by 20%, while the impact on AE exchange rates and foreign direct investment is insignificant. Rows 6 and 8 show the impact of high PDs during corporate distress periods on AEs and EMs, respectively. Without concurrent financial crises, high corporate sector distress is linked to drops in EM exchange rates and foreign direct investment by 6.2% and 11.5%, respectively, while the impact on AEs is insignificant.

While contemporaneous results indicate a lower GDP growth during corporate crises in AEs and EMs, it is lower than those reported during banking crises. We also find evidence on a significant decrease in credit, FDI and exchange rates (depreciation) during corporate crises in EMs. These findings are robust to the impact of concurrent banking crises. However, these results do not clearly identify a causality or Granger causality that corporate distress is likely to follow or coincide with other types of macroeconomic crises in economies.

6 Conclusion

In this paper, we study corporate crises and distress by proposing a new cross-economy measure, constructing an early warning model, and analyzing the macroeconomic consequences. Our early-warning system of corporate sector distress combines several state-of-the-art machine learning methods with robust out-of-sample performance. Our findings illustrate the importance of corporate balance sheet variables and global financial conditions in predicting corporate crises. Furthermore, the results show the

significant impact of corporate distress on GDP and credit growth.

Our results have important policy implications for macroprudential and monetary policies.Our analysis shows that corporate distress has a milder impact than financial crises and its effect on GDP and credit growth is larger in EMs than AEs. From a macrofinancial point of view, the spillover and spillback channels are significant warranting preemptive policies to mitigate the macroeconomic cost of systemic bankruptcies and corporate crises. Surveillance of corporate sector stability and its linkages to the rest of the economy could provide early enough signals of accumulating risks highlighting the importance of integrated policy frameworks. Understanding the risks created in different segments of an economy by monetary, fiscal and financial policies could help policymakers avoid systemic crises and their long lasting cost. This paper underscores the importance of close and timely monitoring of corporate vulnerabilities and implemetation of contingency plans to address these risks in case of materialization.

There are a number of caveats in our work. First, our analysis is limited to publicly listed firms due to data availability. This unfortunately forces us to leave SMEs and private firms out, although in some economies these players constitute a key part of the economy. Secondly, we focus only on advanced and emerging/frontier economies leaving out low-income economies and fragile states. Since our analysis requires availability of high-frequency longitudinal data, we are limited to economies with good data availability. As more data becomes available, we plan to extend our model and work beyond the current economy coverage.

Using our new dataset and measure of corporate distress, our plans for future research include a more thorough analysis of its macroeconomic impact, linking policy effectiveness around corporate crises and the role of macroprudential policy in avoiding systemic financial crises. Another research avenue is looking into sectoral differences as well as the public vs private ownership. The macroeconomic-impact results we document in this paper for EMs warrants a thorough analysis of the underlying reasons for differences between AEs and EMs.

References

Acharya, V. V., Crosignani, M., Eisert, T., and Eufinger, C. (2020). Zombie credit and (dis-) inflation: evidence from europe. Working paper, National Bureau of Economic Research.

Albuquerque, B. (2021). Corporate debt booms, financial constraints and the investment nexus. Working paper, Bank of England Working Paper.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

Andrews, D. and Petroulakis, F. (2019). Breaking the shackles: Zombie firms, weak banks and depressed restructuring in Europe. Working Paper Series 2240, European Central Bank.

Atkeson, A. G., Eisfeldt, A. L., and Weill, P. O. (2017). Measuring the financial soundness of us firms, 1926–2012. *Research in Economics*, 71(3):613–635.

Burman, P., Chow, E., and Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358.

Caballero, R. J., Hoshi, T., and Kashyap, A. K. (2008). Zombie lending and depressed restructuring in Japan. *American Eonomic Review*, 98(5):1943–77.

Campbell, J. Y., Hilscher, J., and Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, 63(6):2899–2939.

Carling, K., Jacobson, T., Lindé, J., and Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of Banking & Finance*, 31(3):845–868.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Claessens, S., Kose, M. A., and Terrones, M. E. (2012). How do business and financial cycles interact? *Journal of International Economics*, 87(1):178–190.

Das, S. R., Duffie, D., Kapadia, N., and Saita, L. (2007). Common failings: How corporate defaults are correlated. *The Journal of Finance*, 62(1):93–117.

Duan, J.-C., Sun, J., and Wang, T. (2012). Multiperiod corporate default prediction-a forward intensity approach. *Journal of Econometrics*, 170(1):191–209.

Duffie, D., Saita, L., and Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financia lEconomics*, 83(3):635–665.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164(1):130–141.

Giesecke, K., Longstaff, F. A., Schaefer, S., and Strebulaev, I. A. (2014). Macroeconomic effects of corporate default crisis: A long-term perspective. *Journal of Financial Economics*, 111(2):297–310.

Gourinchas, P. O., Kalemli-Özcan, e., Penciakova, V., and Sander, N. (2020). Covid-19 and sme failures. Working paper, National Bureau of Economic Research.

Gross, M. and Población, J. (2019). Implications of model uncertainty for bank stress testing.

Journal of Financial Services Research, 55(1):31–58.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, pages 357–384.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and EI george, and a rejoinder by the authors. *Statistical science*, 14(4):382–417.

IMF (2021). Global financial stability report, Apri 2021: Covid-19, crypto, and climate: Navigating challenging transitions.

Jordà, Ò., Kornejew, M., Schularick, M., and Taylor, A. M. (2020). Zombies at large? corporate debt overhang and the macroeconomy. Working paper, National Bureau of Economic Research.

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

Koopman, S. J., Lucas, A., and Schwaab, B. (2012). Dynamic factor models with macro, frailty, and industry effects for us default counts: the credit crisis of 2008. *Journal of Business & Economic Statistics*, 30(4):521–532.

Laeven, L. and Valencia, F. (2020). Systemic banking crises database II. *IMF Economic Review*, 68(2):307–361.

Lian, C. and Ma, Y. (2021). Anatomy of corporate borrowing constraints. *The Quarterly Journal of Economics*, 136(1):229–291.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, 29(2):449–470.

Mian, A., Sufi, A., and Verner, E. (2017). Household debt and business cycles worldwide. *The Quarterly Journal of Economics*, 132(4):1755–1817.

Müller, K. and Verner, E. (2021). Credit allocation and macroeconomic fluctuations. *Available at SSRN 3781981*.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, pages 109–131.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pesaran, M. H., Schuermann, T., Treutler, B.-J., and Weiner, S. M. (2006). Macroeconomic dynamics and credit risk: A global perspective. *Journal of Money, Credit and Banking*, 38(5):1211–1261.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39–61.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, pages 111–163.

Rey, H. (2015). Dilemma not trilemma: the global financial cycle and monetary policy independence. Working paper, National Bureau of Economic Research.

Schivardi, F., Sette, E., and Tabellini, G. (2022). Credit misallocation during the european financial crisis. *The Economic Journal*, 132(641):391–423.

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

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games, Edited by Harold W. Kuhn and Albert W. Tucker*, 2.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1):101–124.

The Credit Research Initiative of the National University of Singapore (2019). Probability of default (PD) white paper.

Wu, J. C. and Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *Journal of Money, Credit and Banking*, 48(2-3):253–291.

Table 1: **Summary Statistics for the Economy-level PD Indices**

Economy	Starting period	N observations	N firms covered	Average	Standard deviation	25 quartile	Median	75 quartile
Argentina	1995Q3	105	56	0.62	0.53	0.27	0.50	0.76
Australia	1992Q1	119	1,044	0.14	0.11	0.08	0.10	0.16
Austria	1992Q2	118	62	0.25	0.24	0.09	0.16	0.29
Bangladesh	2011Q4	40	167	0.13	0.06	0.08	0.11	0.17
Belgium	1991Q2	122	83	0.23	0.41	0.08	0.12	0.24
Brazil	1995Q2	106	203	0.65	0.39	0.35	0.52	0.82
Bulgaria	2006Q1	63	49	0.48	0.28	0.27	0.46	0.59
Canada	1994Q1	111	708	0.52	0.74	0.15	0.25	0.50
Chile	1995Q2	106	104	0.18	0.16	0.08	0.11	0.27
China	1995Q2	106	1,936	1.13	1.00	0.33	0.56	2.06
Colombia	2006Q2	34	21	0.13	0.12	0.05	0.07	0.20
Croatia	2006Q2	62	97	0.20	0.16	0.12	0.16	0.22
Cyprus	1999Q2	90	56	0.57	0.87	0.19	0.32	0.56
Czech Republic	1996Q2	39	48	0.23	0.20	0.11	0.16	0.33
Denmark	1991Q2	122	104	0.16	0.11	0.09	0.13	0.20
Egypt	2007Q1	59	134	0.46	0.22	0.32	0.43	0.61
Finland	1992Q3	117	108	0.22	0.31	0.08	0.14	0.25
France	1990Q2	126	470	0.22	0.26	0.09	0.14	0.26
Germany	1990Q4	124	501	0.23	0.23	0.09	0.17	0.29
Greece	1993Q2	114	202	0.46	0.32	0.21	0.37	0.58
Hong Kong SAR	1992Q1	119	636	0.15	0.12	0.08	0.12	0.16
Hungary	1996Q3	100	27	0.21	0.28	0.08	0.12	0.20
Iceland	2000Q4	18	31	0.33	0.20	0.22	0.28	0.42
India	1994Q2	110	2,006	0.87	0.38	0.55	0.86	1.15
Indonesia	1992Q4	116	238	0.73	0.86	0.27	0.36	0.69
Ireland	1993Q1	115	53	0.16	0.19	0.06	0.10	0.19
Israel	1995Q2	103	311	0.24	0.16	0.12	0.20	0.28
Italy	1990Q1	127	179	0.66	1.04	0.10	0.18	0.69
Jamaica	2011Q3	37	32	0.36	0.23	0.22	0.33	0.41
Japan	1995Q3	105	3,215	0.19	0.20	0.06	0.11	0.21
Jordan	2001Q3	81	83	0.14	0.07	0.10	0.13	0.18
Kenya	2009Q3	49	33	0.16	0.09	0.11	0.14	0.18
Korea	1993Q2	114	1,325	0.66	1.11	0.15	0.32	0.70
Kuwait	2001Q1	83	52	0.20	0.15	0.12	0.16	0.23
Latvia	2007Q1	8	23	0.10	0.05	0.07	0.08	0.10
Lithuania	2005Q2	66	26	0.17	0.23	0.07	0.09	0.20
Luxembourg	2004Q4	64	26	0.23	0.24	0.11	0.15	0.25
Malaysia	1991Q3	121	617	0.23	0.17	0.11	0.16	0.29
Mauritius	2011Q1	36	22	0.44	0.38	0.20	0.30	0.60
Mexico	1995Q2	106	79	0.25	0.21	0.09	0.18	0.36
Morocco	2003Q4	72	46	0.07	0.03	0.05	0.07	0.09
Netherlands	1990Q2	126	115	0.21	0.25	0.08	0.13	0.21
New Zealand	1993Q3	113	71	0.06	0.06	0.02	0.03	0.06
Nigeria	2004Q2	70	83	0.37	0.21	0.23	0.32	0.45
Norway	1991Q3	121	129	0.21	0.15	0.12	0.17	0.24
Oman	2011Q1	43	46	0.21	0.22	0.05	0.07	0.41
Pakistan	2004Q2	70	155	0.27	0.18	0.16	0.22	0.33
Peru	1997Q2	98	43	0.31	0.41	0.06	0.11	0.50
Philippines	1992Q4	116	116	0.40	0.54	0.12	0.20	0.36
Poland	1995Q4	104	258	0.28	0.17	0.18	0.24	0.32
Portugal	1994Q2	110	44	0.27	0.27	0.09	0.13	0.40
Qatar	2012Q4	36	22	0.05	0.02	0.04	0.05	0.06
Romania	1999Q2	85	57	0.36	0.81	0.11	0.16	0.36
Russia	1999Q2	88	130	0.37	0.67	0.16	0.25	0.34
Saudi Arabia	2001Q2	82	87	0.16	0.20	0.06	0.09	0.16
Serbia	2009Q1	43	64	0.39	0.34	0.18	0.20	0.54
Singapore	1992Q1	119	335	0.12	0.12	0.05	0.08	0.14
Slovenia	2004Q2	55	30	0.13	0.12	0.06	0.10	0.14
South Africa	1993Q4	112	251	0.23	0.13	0.13	0.19	0.27
Spain	1992Q1	119	106	0.21	0.20	0.09	0.14	0.28
Sri Lanka	2006Q3	61	152	0.25	0.13	0.15	0.22	0.29
Sweden	1991Q3	121	332	0.18	0.38	0.06	0.09	0.18
Switzerland	1990Q2	126	156	0.10	0.08	0.05	0.07	0.12
Taiwan Province of China	1992Q2	118	572	0.10	0.12	0.03	0.05	0.12
Thailand	1994Q1	111	377	0.43	0.68	0.13	0.19	0.37
Tunisia	2004Q1	71	33	0.15	0.07	0.11	0.14	0.18
Turkey	2002Q4	76	259	0.37	0.23	0.22	0.29	0.45
Ukraine	2007Q1	39	37	0.61	0.37	0.38	0.57	0.67
United Arab Emirates	2007Q2	58	39	0.18	0.11	0.11	0.15	0.21
United Kingdom	1990Q1	127	1,036	0.17	0.14	0.08	0.13	0.20
United States	1990Q4	124	3,672	0.42	0.44	0.11	0.19	0.68
Vietnam	2007Q1	59	448	0.24	0.10	0.17	0.23	0.29

Notes: The third column, labeled 'N observations', shows the number of quarters when the economy level index is available. The fourth column, labeled 'N firms covered', shows the average number of firms each quarter.

Table 2: **Predictor Definition and Data Sources**

Panel A: Balance-sheet Variables and PDs		
Lable	Definition	Source
Investment Rate	Median of capital expenditure to lagged capital stock of firms from each economy	Compustat Global
Investment Rate (change)	Annual changes in investment rates	Compustat Global
Net Current Assets to Total Assets	Median of net current assets to total asset ratio across firms from each economy	Compustat Global
Net Current Assets to Total Asset Ratio (change)	Annual changes in net current assets to total asset ratio	Compustat Global
Debt to Asset Ratio	Median of debt to asset ratio across firms from each economy	Compustat Global
Debt to Asset Ratio Change	Annual changes in debt to asset ratio from each economy	Compustat Global
Default probability 12 month	Capital-weighted averages of 12-month default probability across non-financial firms from each economy	NUS Credit Research Initiative
Default probability 12 month (change)	Quarterly changes in default probability 12 month	NUS Credit Research Initiative
Default probability 36 month	Equal weighted averages of 36-month default probability across non-financial frims from each economy	NUS Credit Research Initiative
Default probability 36 month (change)	Quarterly changes in default probability 36 month	NUS Credit Research Initiative
Interest Coverage Ratio (moving average)	Annual moving averages of median of interest coverage ratio across firms from each economy	Compustat Global
Interest Coverage Ratio (change)	Annual changes in median of interest coverage ratio	Compustat Global
Net Debt (exl. liquid assets) to Asset Ratio	Median of net debt (exl. liquid assets) to asset ratio	Compustat Global
Net Debt (exl. liquid assets) to Asset Ratio (change)	Annual changes in net debt (exl. liquid assets) to asset ratio	Compustat Global
ROA	Return on assets computed as EBIT divided by total assets	Compustat Global
ROA (change)	Annual changes in return of asset	Compustat Global
Short-term investment to liability ratio	Median of cash and short-term invest to liability ratio	Compustat Global
Short-term investment to liability ratio (change)	Annual changes in short-term investment to liability ratio	Compustat Global
Panel B: Macroeconomic Variables		
Lable	Definition	Source
GDP gap	One-sided GDP gap, computed by HP filter with Lambda = 1,600	World Economic Outlook
Inflation	YoY% inflation rate	World Economic Outlook
Foreign Reserve gap	One-sided foreign reserve gap, computed by HP filter with Lambda = 1,600	International Financial Statistics - IMF Data
Gov Bond Yield 10y	Gov Bond Yield 10y	Global Financial Data
Policy Rate	Policy Rate	Global Financial Data
Credit GDP Ratio	Credit GDP Ratio	Bank of International Settlement, and IMF staff calculates
Credit GDP Gap	Credit GDP Gap	Bank of International Settlement, and IMF staff calculates
Quarterly real GDP growth	Real GDP growth	World Economic Outlook
BOP to corporate sector to GDP ratio	BOP other inv. (net) to non-official, non-bank sector-to-GDP ratio	Bank of International Settlement
Fed Funds Shadow Rate	Wu-Xia Shadow Federal Funds Rate	Federal Reserve Bank of Atlanta
Dollar Appreciation	YoY% Dollar Appreciation	Information Notice System - IMF Data
Panel C: Stock Price Valuation		
Lable	Definition	Source
Dividend Yield	Dividend Yield	Datastream
Market Index Growth	YoY% return of market index	Datastream
Price Earning Ratio	Price Earning Ratio	Datastream
Market to Book Value	Maket value over book value of market index	Datastream
Volatility of Market Index	Volatility of Market Index	Datastream

Table 3: **Posteiors of Key Parameters in the Markov Regime-switching Model**

Posteior	ρ	$\delta+1$	$\gamma+1$	p_{11}	p_{22}
Mean	0.91	3.67	5.9	0.95	0.81
Standard deviation	0.01	0.85	0.15	0.00	0.01
Median	0.91	3.68	5.9	0.95	0.81
5% Quantile	0.90	2.33	5.63	0.94	0.78
95% Quantile	0.92	5.09	6.17	0.96	0.83

Table 4: **Out-of-sample Model Performance**

AUROC	XGBoost	Logit	SVM	Random Forest	LDA	Combination
All Economies	0.65	0.64	0.55	0.59	0.64	0.67
Advanced Economy	0.68	0.63	0.55	0.61	0.63	0.68
Emerging Market	0.62	0.65	0.54	0.57	0.65	0.65
Log-likelihood	**XGBoost**	**Logit**	**SVM**	**Random Forest**	**LDA**	**Combination**
All Economies	−1083.42	−1163.00	−1476.90	−1146.92	−1274.32	−1071.11
Advanced Economy	−583.98	−666.75	−866.67	−646.96	−758.36	−590.81
Emerging Market	−499.45	−496.25	−610.23	−499.96	−515.96	−480.31

Table 5: **Cross-validation Model Performance**

AUROC	XGBoost	Logit	SVM	Random Forest	LDA	Combination
All Economies	0.68	0.66	0.56	0.67	0.70	0.71
Advanced Economy	0.70	0.68	0.57	0.69	0.71	0.72
Emerging Market	0.66	0.63	0.54	0.63	0.67	0.68
Log-likelihood	**XGBoost**	**Logit**	**SVM**	**Random Forest**	**LDA**	**Combination**
All Economies	−1550.17	−1572.41	−2215.94	−1539.27	−1736.38	−1471.33
Advanced Economy	−919.30	−940.87	−1380.72	−924.51	−1027.45	−883.95
Emerging Market	−630.86	−631.54	−835.22	−614.76	−708.94	−587.38

Table 6: **Conditional Growth Rates of GDP and Credit to Private Non-financial Sector during High Corporate Distress Periods, Relative to Low Corporate Distress Periods**

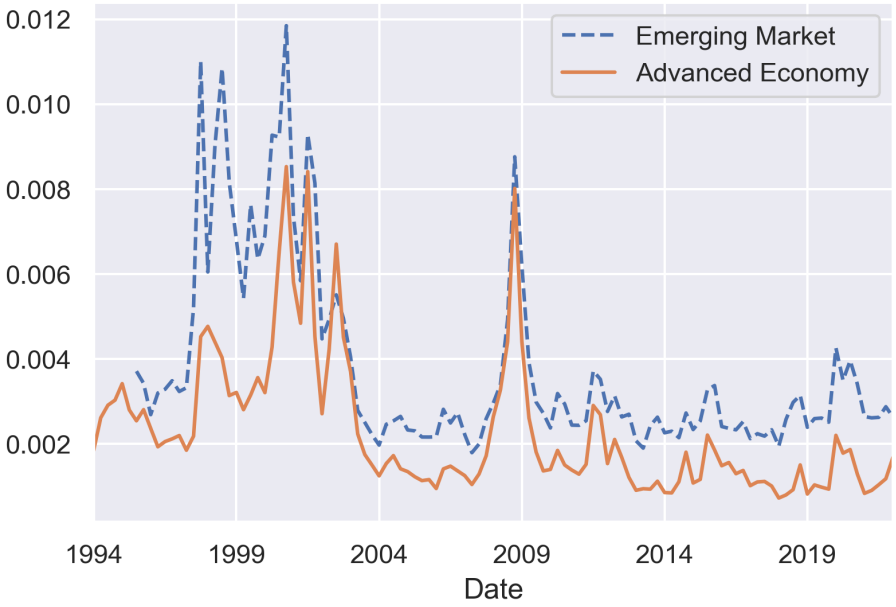
	GDP				Credit			
Unconditional	−2.96***				−2.06*			
	(0.43)				(1.13)			
AE		−2.40***				0.28		
		(0.53)				(1.00)		
EM		−3.93***				−6.49***		
		(0.66)				(2.06)		
Nonfinancial			−1.82***				−0.78	
			(0.34)				(0.82)	
Financial			−4.03***				−5.62***	
			(0.50)				(1.08)	
AE Nonfinancial				−1.60***				0.84
				(0.47)				(0.76)
AE Financial				−4.15***				−3.82***
				(0.69)				(0.76)
EM Nonfinancial				−2.23***				−3.80**
				(0.43)				(1.57)
EM Financial				−3.75***				−9.03***
				(0.63)				(2.58)

Table 7: **Conditional Growth Rates of Exchange Rates and Foreign Direct Investment during High Corporate Distress Periods, Relative to Low Corporate Distress Periods**

	Exchange Rate				Foreign Direct Investment			
Unconditional	-3.38^{***}				-8.38^{***}			
	(1.09)				(2.75)			
AE		-0.53				-3.85		
		(0.94)				(2.84)		
EM		-8.86^{***}				-19.87^{***}		
		(1.80)				(4.62)		
Nonfinancial			-2.40^{**}				-5.85^{**}	
			(0.98)				(2.31)	
Financial			-2.51^{***}				-9.60^{***}	
			(0.96)				(3.68)	
AE Nonfinancial				-0.25				-2.79
				(0.87)				(2.59)
AE Financial				-0.44				-4.30^{**}
				(0.71)				(2.04)
EM Nonfinancial				-6.23^{***}				-11.52^{***}
				(1.83)				(3.87)
EM Financial				-7.80^{***}				-41.69^{***}
				(2.24)				(7.58)

Figure 1: **Average PD Indices and Number of Economies in Corporate Distress across Advanced Economies and Emerging Markets**

(a) Average PD Indices



(b) Number of Economies in Corporate Distress

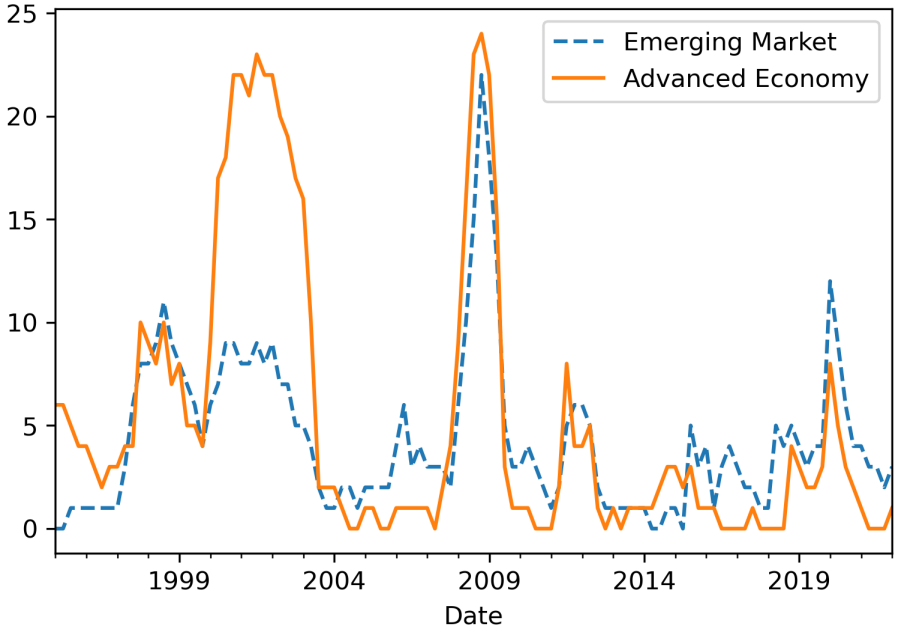
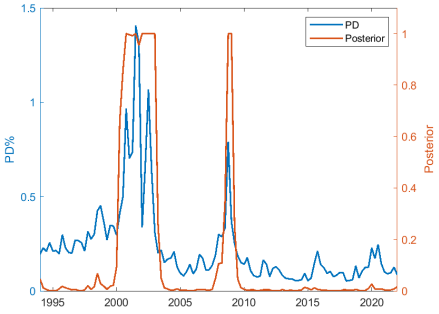
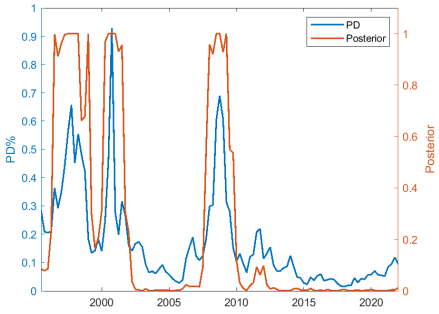


Figure 2: **Probability of Default Indices and Posterior Probability**

(a) US



(b) Japan



(c) China

(d) Brazil

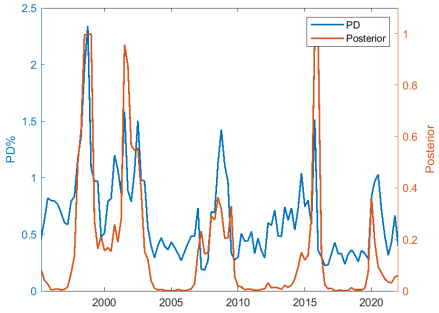
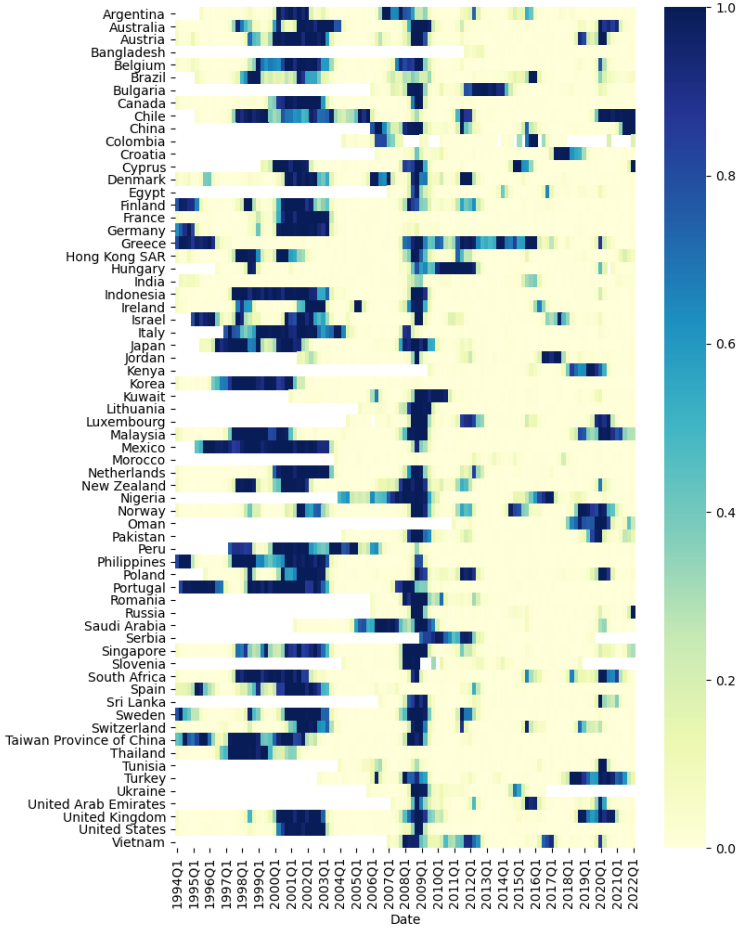


Figure 3: **Posterior Probability of the High Corporate Distress Regime**

(a) Posterior



(b) Crises Periods

Figure 4: **Corporate Distress Periods vs. Banking Crises Periods**

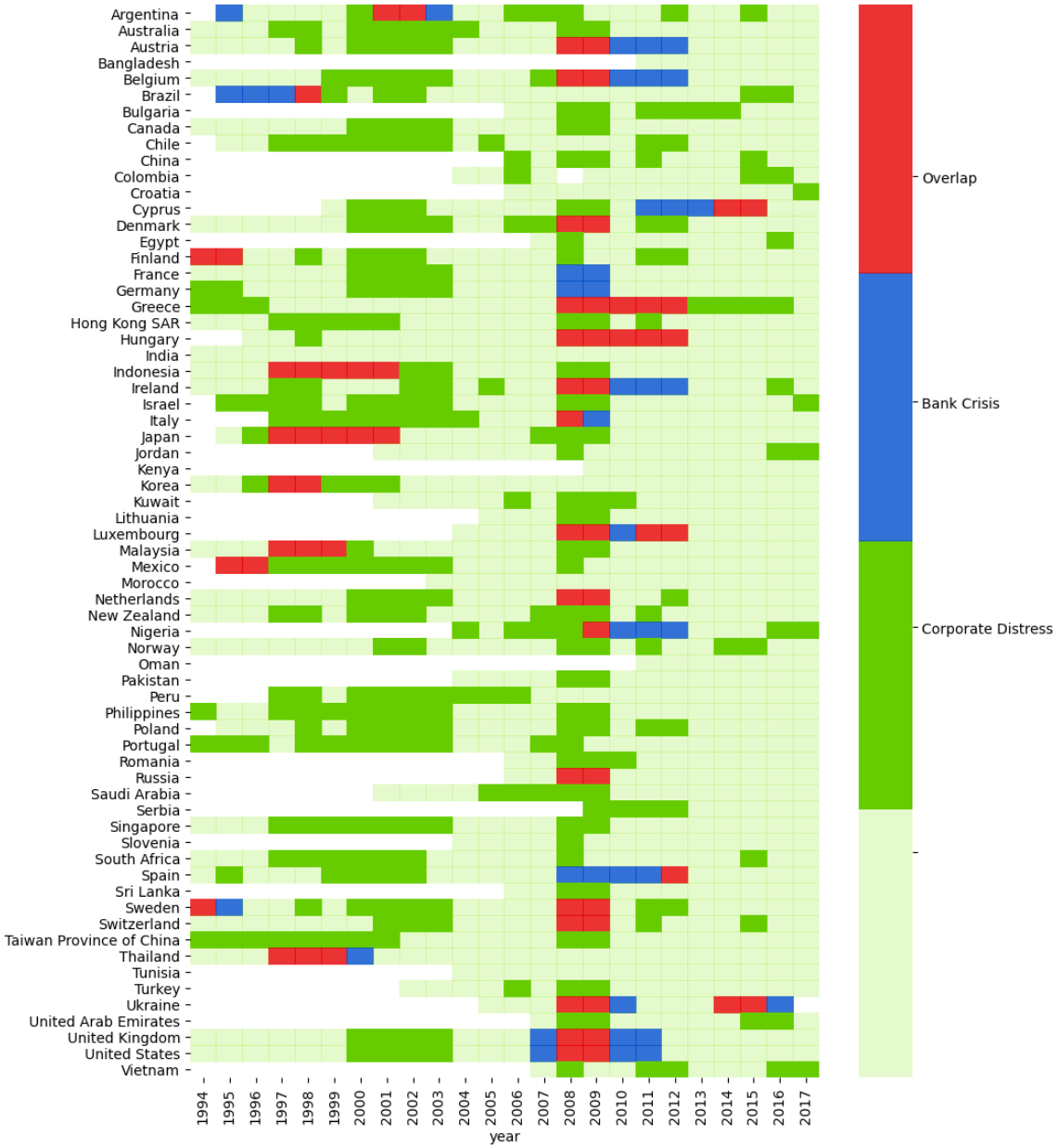


Figure 5: **Corporate Distress Periods vs. Currency and Sovereign Debt Crises Periods**

Figure 6: **Predictor Availability**

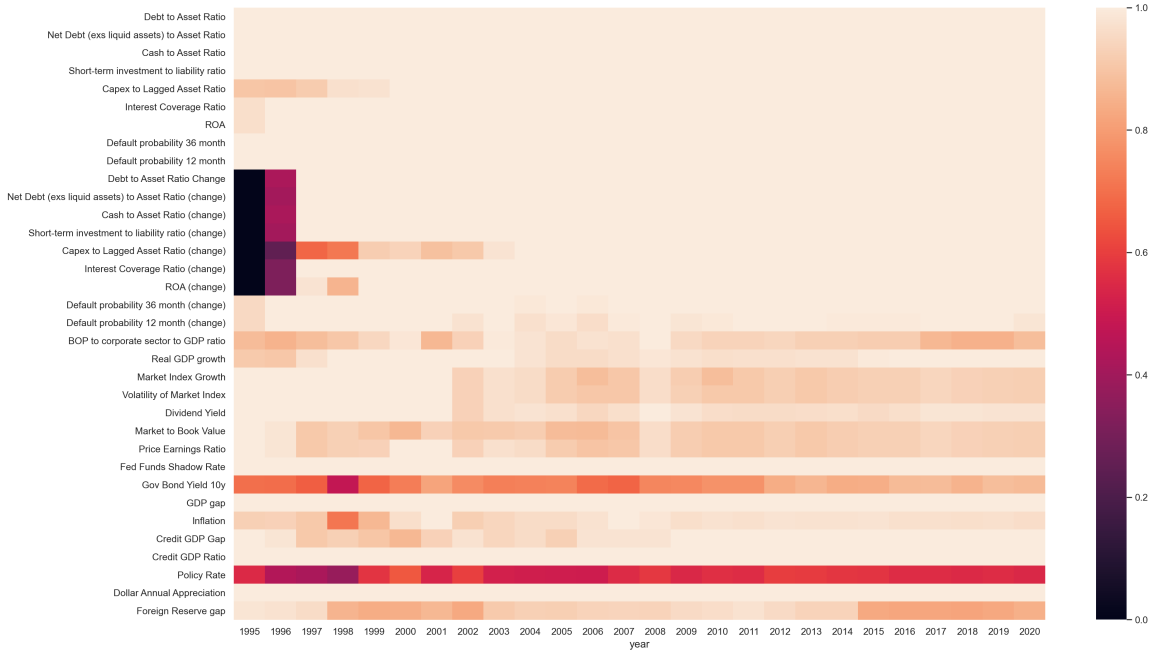


Figure 7: **Cross-validation AUROC**

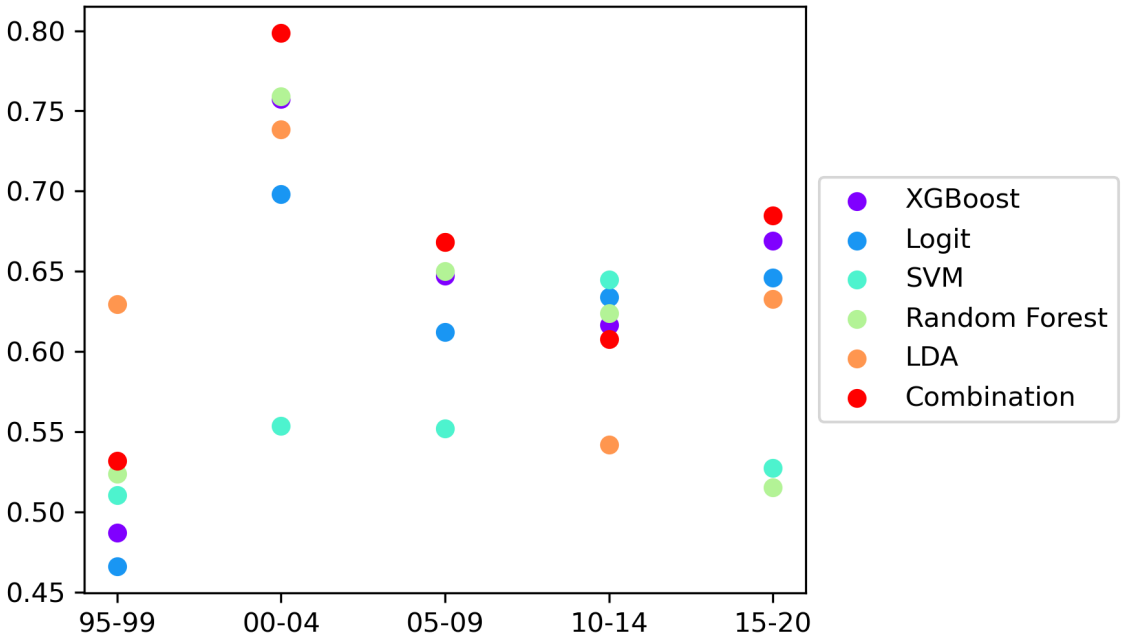
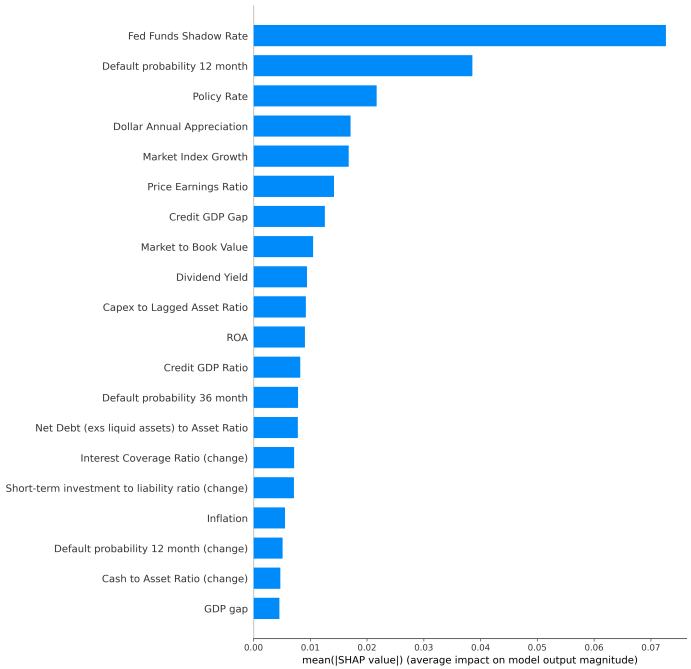


Figure 8: **Summary of Shapley Values**

(a) Mean absolute Shapley values



(b) Shapley values distribution

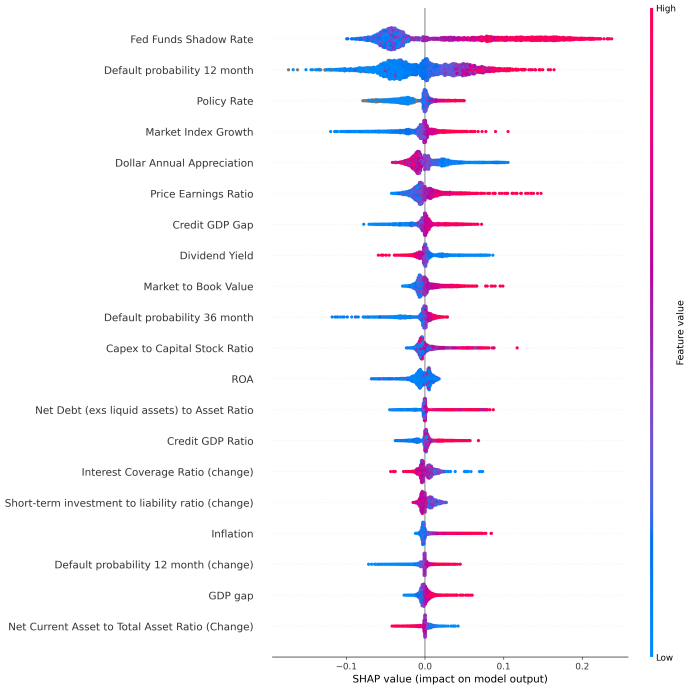
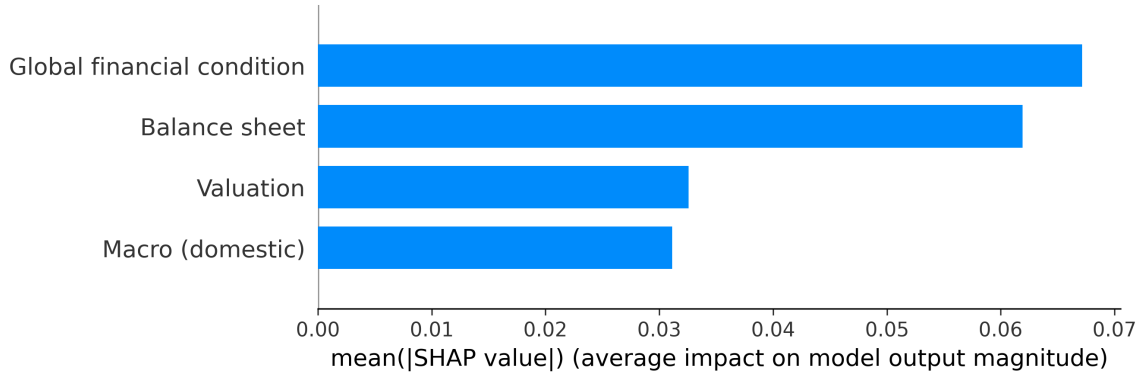


Figure 9: **Summary of Shapley Values of Predictors from Different Categories**



A MCMC Algorithm to Identify Corporate Distress

Commonly used conjugate priors are adopted to increase the speed of estimation. We assume p_{11} and p_{22} have independent priors Beta(1,1). We assume δ has flat Gaussian prior, $\delta \sim N(0, \infty)$ with positive constraints. For identification purposes, we assume ρ has flat Gaussian prior, $\rho \sim N(0, \infty)$, with constraints $\rho \in (0, 1)$ to ensure stationarity. The inverse of variance ratio of high-PD regime over low-PD regime, $1/(1+\gamma)^2$ has prior of Gamma(2,2). Conditional mean of PDs in low-risk regime, μ_{iL} is assumed to have flat Gaussian prior, $\mu_{iL} \sim N(0, \infty)$, and $1/\sigma_{iL}^2$ has prior of Gamma(2,2).

Before elaborating on the MCMC algorithm, we define some notations. Let $\tilde{S}_{it}$ denotes vector of states $\tilde{S}_{iT} = [S_{i1}, S_{i2}, ..., S_{iT}]$ of economy i, and let $\tilde{y}_{iT}$ denote the vector of observations, $\tilde{y}_{iT} = [PD_{i1}, PD_{i2}, ..., PD_{iT}]$. MCMC sampling are implemented using the steps below:

1. Initial values of $p_{11}, p_{22}, \delta, \gamma, \rho, \mu_{iL}, \sigma_{iL}^2$ $(i \in \{1, 2, ..., N\})$ are proposed.
2. For each $i = 1, 2, ..., N$, sample vector of states $\tilde{S}_{iT} = [S_{i1}, S_{i2}, ..., S_{iT}]$ is sampled from $f\left(\tilde{S}_{iT} | p_{11}, p_{22}, \delta, \gamma, \rho, \mu_{iL}, \sigma_{iL}^2, \tilde{y}_{iT}\right)$ using multi-move Gibbs-sampling as proposed in Carter and Kohn (1994).
3. For each $i = 1, 2, ..., N$, sample μ_{iL} from $f\left(\mu_{iL} | \tilde{S}_{iT}, p_{11}, p_{22}, \delta, \gamma, \rho, \sigma_{iL}^2, \tilde{y}_{iT}\right)$.
4. For each $i = 1, 2, ..., N$, sample σ_{iL}^2 from $f\left(\sigma_{iL}^2 | \tilde{S}_{iT}, p_{11}, p_{22}, \delta, \gamma, \rho, \mu_{iL}, \tilde{y}_{iT}\right)$.
5. Sample p_{11}, p_{22} from $f\left(p_{11}, p_{22} | \tilde{S}_{1T}, \tilde{S}_{2T}, ..., \tilde{S}_{NT}\right)$.
6. Sample ρ from $f\left(\rho | \tilde{S}_{iT}, p_{11}, p_{22}, \delta, \gamma, \mu_{iL}, \sigma_{iL}^2, \tilde{y}_{iT}, \text{for all } i \in \{1, 2, ..., N\}\right)$.
7. Sample δ from $f\left(\delta | \tilde{S}_{iT}, p_{11}, p_{22}, \rho, \gamma, \mu_{iL}, \sigma_{iL}^2, \tilde{y}_{iT}, \text{for all } i \in \{1, 2, ..., N\}\right)$.
8. Sample γ from $f\left(\gamma | \tilde{S}_{iT}, p_{11}, p_{22}, \rho, \delta, \mu_{iL}, \sigma_{iL}^2, \tilde{y}_{iT}, \text{for all } i \in \{1, 2, ..., N\}\right)$.

Finally, we repeat steps 2-8 until the Markov chain is properly mixed, and we accumulate enough samples to represent the posterior distribution.

B Constructing Predictors from Compustat Global

We use quarterly data on listed non-financial corporations for 55 economies from S&P Compustat North America and Compustat Global. We exclude financial firms, namely banks, diversified financial, and insurance firms from our analysis. Our final sample comprises an unbalanced panel of 56,758 non-financial firms over 1995q1-2021q3. We first compute the balance-sheet ratio of individual firms, and then take cross-sectional median at each quarter. Below is the definition of the financial ratios:

- Investment Rate = Capital Expenditure / Previous-Quarter Value of Property, Plant and Equipment
- Net Current Asset to Asset Ratio = (Current Asset-Current Liabilities) / Asset
- Debt to Asset Ratio = (Debt in Current Liabilities+Long-term Debt) / Asset
- Interest Coverage Ratio = Earning before interest and taxes / Interest expense
- Net Debt to Asset Ratio = (Debt in Current Liabilities+Long-term Debt-(Current Liability-Debt in Current Liabilities)) / Asset
- ROA = ($0.625 \cdot$ Earning before interest and taxes) / (Asset + Lagged Asset/2)
- Short-term Investment to Liability Ratio = Cash and Equivalent / Liabilities

We make the following adjustments:

- Drop observations for missing assets and liabilities.
- Drop observations when acquisitions are larger than 5% of total assets.
- Winsorize variables at the 1%/99% percentiles at the economy level.
- To ensure representativeness, we drop economies with fewer than 5 firms at each point in time.
- To ensure further representativeness, we only compute medians for each indicator at each point in time and for each economy when the coverage is at least 30% of all firms reporting data.

- We use annual moving average of quarterly interest coverage ratio, investment rate and ROA to filter out the seasonality in the quarterly earnings and investments.

C Machine Learning Models and Hyperparameter Selection

In this section, we elaborate on machine-learning models that we used.

C.1 Logistic Regression with Regularization

Logistic regression belongs to the class of generalized linear model. The outcome variable y_i takes values 0 and 1. Let X_i be a column vector of predictors. The probability that $y_i = 1$ takes the form

$$\text{Prob}\left(y_i = 1\right) = \frac{\exp\left(X_i'\beta\right)}{1 + \exp\left(X_i'\beta\right)},$$

where β is a column vector. The parameter β is estimated by minimizing the summation of the log-likelihood function and a regularization function:

$$\hat{\beta} = \arg\min_{\beta} -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \ln\left(\text{Prob}(y_i = 1)\right) + \left(1 - y_i\right)\ln\left(1 - \text{Prob}(y_i = 1)\right)\right] + C\left\|\beta\right\|^2.$$

Parameter C is a hyperparameter that determines the amount of regularization: the larger the value of C, the greater the amount of shrinkage and thus the coefficients are less prone to overfit the sample.

C.2 Random Forest

Random forest is an algorithm that combines the predictions from many individual randomized decision trees. The averaging method diversify the forecast errors of individual predictors, and address the issue of overfitting. We first introduce decision trees.

A decision tree is a classification algorithm that recursively makes decisions based on one predictor and one threshold at each node: Once determined the predictor is

above or below the threshold, we proceed to the sub tree and repeat the process, until there is no sub trees, or a leaf has been reached. The resulting predicted probability of each class is the proportion of each sample class in the leaf. The binary classification is estimated recursively: At each node, variable j and split point s are selected to split the sample into two groups, $L(j,s)$ and $S(j,s)$ with N_L and N_R samples respectively. We seek the splitting variable j and split point s that solve

$$\min_{j,s} \left[N_L \bar{y}_{L(j,s)} \left(1 - \bar{y}_{L(j,s)}\right) + N_R \bar{y}_{R(j,s)} \left(1 - \bar{y}_{R(j,s)}\right) \right].$$

We continue to split each nodes until the height of the tree reaches a specified maximum length.

The random forest algorithm averages predictions from randomized decision trees. Each tree in the ensemble is built from a sample drawn with replacement. Furthermore, when splitting each node during the construction of a tree, the best split is selected from a random subset of predictors, with specified number of predictors in the subset. The hyperparameters to be set with cross-validation are the maximum height of individual trees, the size of subset of predictors from which the best split is selected, and the number of individual trees to combine. To reduce computation burden, we left the other hyperparameters to default values in the Scikit-learn (Pedregosa et al., 2011) RanfomForestClassifier function.

C.3 Support Vector Machine

Support vector machine (SVM) is a classification algorithm that produces nonlinear boundary in the feature space. SVM first implement nonlinear function to map the feature space into a new feature space. Then, it separates the new feature space with a hyperplane. We start by introducing SVM when the boundary is linear.

Define a hyperplane by

$$\{x : f(x) = x'\beta + \beta_0 = 0\}.$$

SVM solves

$$\min_{\beta,\beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} \xi_i$$

subject to

$$\xi_i \geq 0,\ y_i\left(x_i'\beta + \beta_0\right) \geq 1 - \xi_i\ \forall i.$$

ξ_i measures how much the sample is within the margin or violate the separation by hyperplane. The hyperparameter C characterize the cost of misclassification. Given the solutions $\hat{\beta}_0$ and $\hat{\beta}_1$, the decision function can be written as

$$\hat{G}(x) = \text{sign}\left[x'\hat{\beta} + \hat{\beta}_0\right].$$

The linear boundary can be easily extended to the nonlinear boundary by mapping x to $h(x)$ where h is a vector function. It turns out the solution involve $h(x)$ through inner product:

$$K(x_1, x_2) = \langle h(x_1), h(x_2)\rangle.$$

Three popular choices for K in the SVM literature are

$$\begin{aligned} d\text{th degree polynomial:} K(x_1, x_2) &= \left(\gamma\langle x_1, x_2\rangle + \kappa\right)^d, \\ \text{Radial basis:} K(x_1, x_2) &= \exp\left(-\gamma\|x_1 - x_2\|^2\right), \\ \text{Neural network:} K(x_1, x_2) &= \tanh\left(\gamma\langle x_1, x_2\rangle + \kappa\right). \end{aligned}$$

The form of K can be treated as hyperparameters. To reduce the computational burden of tuning hyperparameters, coefficients γ and κ are set as default in Scikit-learn (Pedregosa et al., 2011) SVC function ($\gamma = 1/\text{number of features}$, $\kappa = 0$). Hence, we only need to select K and C through cross-validation.

SVC does not yield probability about each class. To get the probability, we fit a logit model using scores of sample observations. The calibration of logit parameters is through cross-validation conducted internally in the SVC function.

C.4 Linear Discriminant Analysis

Linear discriminant analysis assume features are generated from distinct multivariate Gaussian distributions from each class. The predicted probability is the posterior conditional on the observed feature.

LDA assumes samples of class 0 and 1 are independently generated with probability $1-p$ and p respectively. Conditional on the class c, features are generated from Gaussian distributions $N(\mu_c, \Sigma)$, $c \in \{0, 1\}$. The posterior probability that sample i

is from class 1 is

$$\frac{p\phi\left(x_i|\mu_1,\Sigma\right)}{\left(1-p\right)\phi\left(x_i|\mu_0,\Sigma\right)+p\phi\left(x_i|\mu_1,\Sigma\right)},$$

where ϕ is the probability density function. Model estimation is straightforward: p is the sample frequency of class 1; $\mu_0$and $\mu_1$are sample mean of features in each classes. To reduce estimation error, the covariance matrix Σ is estimated shrinkage method. The resulting estimate $\hat{\Sigma}_s$ is the weighted average of sample covariance matrix $\hat{\Sigma}$ and an identity matrix multiplied the average of diagonal components in$\hat{\Sigma}$.

$$\hat{\Sigma}_s = (1-\alpha)\hat{\Sigma} + \frac{\alpha}{N}\text{tr}(\hat{\Sigma})I,$$

where N is the number of features. α is a hyperparameter that is selected by cross-validation.

One advantage of LDA is that its Bayesian framework allows rigorous treatment of missing predictors: we can just generate posterior from existing predictors without resorting to imputations. We customize our own LDA functions.

C.5 Extreme Gradient Boosting Tree

Extreme gradient boosting tree (Chen and Guestrin, 2016) is a from of gradient boosting tree that combines the outputs of many "weak" classifiers to produce a powerful "committee". Hence the fitted value of sample (x_i, y_i) is

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),$$

where $f_k\left(x_i\right)$ is a base estimator. The model is trained in an additive manner. Let $\hat{y}_i^{(k-1)}$ prediction at the $k-1$ th iteration, we seek f_k to further decrease the objection function

$$\sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)\right) + \Omega\left(f_k\right),$$

where l is the loss function and Ω penalized the complexity of the model:

$$\Omega\left(f\right) = \gamma T + \frac{1}{2}\lambda\left\|w\right\|^2,$$

where T is the number of leaves in the tree and $w \in \mathbb{R}^T$ is f's predicted value at each node.

Besides regularization, extreme gradient boosting also adopts shrinkage and subsampling to reduce overfitting. Rather than adding f_k to $\hat{y}_i^{(k-1)}$, shrinkage only add ηf_k where $0 < \eta < 1$. Each f_k is estimated with a random selected subset of sample to reduce the correlations across basis classifiers. Computationally, extreme gradient boosting tree uses approximate algorithm for split finding of individual trees to speed up the program.

In our exercise, we use an equal average of 50 trees as basis predictor f_k to further reduce overfitting. Extreme boosting involves many hyperparameters. To reduce computational burden, we set some of them to the number commonly used in the literature: maximum depth of tress is 3; number of samples per tree is 8 (2-years of observation); random sample size is 50% of total sample size; λ in regularization function Ω is 1; scale the gradient of samples of $y_i = 1$ (pre-distress periods) to 5 to address unbalanced classes. We use cross-validation to find hyperparameters for total number of iterations K, γ in regularization function Ω and shrinkage parameter η.

Because we scaled up the pre-distress periods class, the resulting output of Xgboost is biased estimate of the probability of $y_{it} = 1$. In order to combine it with output from other models, we need the predictions to be comparable across models. To address this issue, we debiase the probability of Xgboost output by

$$\text{Prob}_{\text{unbiased}} = \frac{\text{Prob}_{\text{biased}}}{5 + \text{Prob}_{\text{biased}}}.$$

Table A1: **Corporate Distress Episodes**

	Periods of corporate distress
Argentina	2000Q2-2002Q1, 2006Q4-2008Q3, 2012Q2-2012Q2, 2015Q3-2015Q3
Australia	1997Q4-1998Q1, 2000Q1-2000Q2, 2001Q3-2004Q1, 2008Q3-2009Q3, 2020Q1-2021Q1
Austria	1998Q3-1998Q3, 2000Q1-2003Q2, 2008Q3-2009Q2, 2018Q4-2019Q1, 2020Q1-2020Q2
Belgium	1999Q1-2003Q2, 2007Q3-2009Q1, 2020Q1-2020Q1
Brazil	1998Q2-1999Q1, 2001Q3-2002Q3, 2015Q4-2016Q1
Bulgaria	2008Q2-2009Q1, 2011Q4-2014Q1
Canada	2000Q2-2003Q1, 2008Q3-2009Q1
Chile	1997Q4-1999Q3, 2000Q3-2001Q3, 2002Q1-2003Q3, 2005Q2-2005Q4, 2011Q3-2012Q1, 2020Q1-2022Q1
China	2006Q1-2006Q4, 2008Q1-2009Q1, 2011Q3-2011Q3, 2015Q3-2015Q3, 2021Q3-2022Q1
Colombia	2006Q2-2006Q4, 2015Q3-2016Q1
Croatia	2017Q2-2018Q4
Cyprus	2000Q1-2002Q1, 2008Q1-2009Q2, 2014Q4-2015Q3, 2022Q1-2022Q1
Denmark	2000Q4-2002Q3, 2003Q2-2003Q2, 2006Q1-2007Q1, 2008Q3-2009Q2, 2011Q3-2012Q1
Egypt	2008Q3-2008Q4, 2016Q4-2016Q4
Finland	1995Q1-1995Q2, 1998Q2-1998Q3, 2000Q3-2002Q2, 2008Q2-2008Q4, 2011Q3-2011Q3, 2012Q2-2012Q2
France	2000Q2-2003Q2
Germany	1995Q1-1995Q1, 2000Q2-2003Q2
Greece	1995Q1-1996Q2, 2008Q1-2009Q2, 2010Q1-2010Q2, 2011Q2-2012Q3, 2013Q1-2013Q1, 2013Q3-2016Q1, 2020Q1-2020Q1
Hong Kong SAR	1997Q4-1999Q1, 2000Q2-2001Q2, 2008Q3-2009Q1, 2011Q2-2011Q2
Hungary	1998Q3-1998Q4, 2008Q4-2009Q2, 2010Q1-2012Q2
Indonesia	1997Q3-2003Q2, 2008Q3-2009Q2
Ireland	1997Q4-1998Q3, 2002Q1-2003Q1, 2005Q1-2005Q2, 2008Q2-2009Q2, 2016Q2-2016Q2
Israel	1995Q1-1996Q2, 1997Q4-1998Q1, 2000Q4-2002Q3, 2003Q1-2003Q1, 2008Q4-2009Q1, 2017Q3-2017Q3
Italy	1997Q1-1997Q4, 1998Q4-2004Q2, 2008Q1-2008Q2
Japan	1996Q3-1999Q1, 2000Q2-2001Q3, 2007Q4-2009Q4
Jordan	2008Q4-2008Q4, 2016Q3-2017Q3
Kenya	2018Q2-2020Q2
Korea	1996Q4-2001Q1
Kuwait	2006Q2-2006Q2, 2008Q4-2010Q3
Lithuania	2008Q2-2009Q3
Luxembourg	2008Q3-2009Q2, 2011Q3-2012Q2, 2019Q4-2020Q3
Malaysia	1997Q3-2000Q4, 2008Q2-2009Q2, 2018Q4-2019Q1, 2020Q1-2021Q2
Mexico	1995Q3-2003Q2, 2008Q4-2008Q4
Netherlands	2000Q1-2003Q2, 2008Q2-2009Q1, 2012Q2-2012Q2
New Zealand	1997Q4-1998Q4, 2000Q3-2002Q1, 2007Q4-2009Q2, 2011Q3-2011Q3, 2020Q1-2020Q1
Nigeria	2004Q2-2004Q3, 2006Q1-2006Q1, 2007Q1-2009Q3, 2016Q1-2017Q1
Norway	2001Q3-2002Q4, 2008Q3-2009Q2, 2011Q3-2011Q3, 2014Q3-2015Q1, 2018Q4-2020Q2
Oman	2018Q2-2020Q3
Pakistan	2008Q2-2009Q3, 2019Q3-2020Q1
Peru	1997Q2-1998Q3, 2000Q1-2002Q3, 2003Q3-2005Q1, 2006Q1-2006Q2
Philippines	1997Q3-1999Q2, 2000Q1-2003Q2, 2008Q4-2009Q1
Poland	1998Q3-1998Q3, 2000Q3-2003Q1, 2008Q4-2009Q2, 2011Q3-2012Q2, 2020Q1-2020Q3
Portugal	1995Q1-1996Q4, 1998Q3-2003Q1, 2007Q3-2008Q3
Romania	2008Q1-2009Q2, 2010Q2-2010Q2
Russia	2008Q3-2009Q1, 2022Q1-2022Q1
Saudi Arabia	2005Q1-2007Q3, 2008Q4-2009Q4
Serbia	2009Q1-2010Q4, 2011Q2-2012Q1
Singapore	1997Q4-1998Q2, 1999Q2-1999Q3, 2000Q4-2003Q1, 2008Q1-2009Q2
Slovenia	2008Q1-2008Q4
South Africa	1997Q4-2002Q1, 2008Q3-2008Q4, 2015Q3-2015Q3, 2018Q2-2018Q2, 2020Q1-2021Q1
Spain	1995Q2-1995Q3, 1999Q1-1999Q1, 2000Q2-2002Q4, 2012Q2-2012Q2
Sri Lanka	2008Q2-2009Q2, 2020Q1-2020Q1
Sweden	1998Q3-1998Q3, 2000Q4-2003Q2, 2008Q2-2009Q2, 2011Q3-2012Q1
Switzerland	2001Q3-2003Q2, 2008Q3-2009Q2, 2011Q3-2011Q3, 2015Q3-2015Q3, 2018Q4-2018Q4
Taiwan Province of China	1995Q1-1996Q1, 1997Q2-1999Q1, 2000Q1-2001Q4, 2008Q2-2009Q1
Thailand	1997Q2-1999Q3
Tunisia	2020Q1-2020Q2
Turkey	2006Q2-2006Q2, 2008Q1-2009Q2, 2018Q2-2021Q3
Ukraine	2008Q3-2009Q2, 2014Q4-2015Q1
United Arab Emirates	2008Q4-2009Q1, 2015Q3-2016Q1, 2020Q1-2020Q2
United Kingdom	2000Q2-2003Q1, 2008Q2-2009Q2, 2018Q4-2020Q4
United States	2000Q2-2003Q1, 2008Q4-2009Q1
Vietnam	2008Q1-2008Q4, 2011Q4-2012Q3, 2016Q3-2017Q1, 2020Q1-2020Q1

PUBLICATIONS

Understanding and Predicting Systemic Corporate Distress: A Machine-learning Approach
Working Paper No. **WP/2022/153**